ELSEVIER
ACADEMIC
PRESS

VOLUME THREE

# ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

EDITOR-IN-CHIEF

DANIEL HILLEL

# ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

# ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

Editor-in-Chief
DANIEL HILLEL

Editors
JERRY L HATFIELD
DAVID S POWLSON
CYNTHIA ROSENZWEIG
KATE M SCOW
MICHAEL J SINGER
DONALD L SPARKS

ELSEVIER
ACADEMIC
PRESS

Amsterdam Boston Heidelberg London New York Oxford
Paris San Diego San Francisco Singapore Sydney Tokyo

Elsevier Ltd., The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK

AGGREGATION/Physical Aspects
ARCHAEA
CONDITIONERS
CROP-RESIDUE MANAGEMENT
CRUSTS/Biological
CRUSTS/Structural
EROSION/Wind-Induced
EROSION/Irrigation-Induced
FRACTAL ANALYSIS
FREEZING AND THAWING/Processes
IMMISCIBLE FLUIDS
IRRIGATION/Methods
LYSIMETRY
ORGANIC RESIDUES, DECOMPOSITION
POROSITY AND PORE-SIZE DISTRIBUTION
PRECIPITATION, WATERSHED ANALYSIS
QUALITY OF SOIL
RANGE MANAGEMENT
RHIZOSPHERE
SUBSOILING
SURFACE COMPLEXATION MODELING
WOMEN IN SOIL SCIENCE (USA)

First edition 2005

Library of Congress Control Number: 2004104447

A catalogue record for this book is available from the British Library

ISBN 0-12-348530-4 (set)

This book is printed on acid-free paper
Printed and bound in the United Kingdom

# EDITORIAL ADVISORY BOARD

# FOREWORD

The *Encyclopedia of Soils in the Environment* is a vitally important scientific publication and an equally important contribution to global public policy. The *Encyclopedia* brings together a remarkable range of cutting-edge scientific knowledge on all aspects of soil science, as well as the links of soils and soil science to environmental management, food production, biodiversity, climate change, and many other areas of significant concern. Even more than that, the *Encyclopedia* will immediately become an indispensable resource for policy makers, analysts, and students who are focusing on one of the greatest challenges of the 21st century. With 6.3 billion people, our planet is already straining to feed the world's population, and is failing to do so reliably in many parts of the world. The numbers of chronically poor in the world have been stuck at some 800 million in recent years, despite long-standing international goals and commitments to reduce that number by several hundred million. Yet the challenge of food production will intensify in coming decades, as the human population is projected to rise to around 9 billion by mid-century, with the increased population concentrated in parts of the world already suffering from widespread chronic under-nourishment.

Unless the best science is brought to these problems, the situation is likely to deteriorate sharply. Food production systems are already under stress, for reasons often related directly to soils management. In Africa, crop yields are disastrously low and falling in many places due to the rampant depletion of soil nutrients. This situation needs urgent reversal, through increasing use of agro-forestry techniques (e.g. inter-cropping cereals with leguminous nitrogen-fixing trees) and increasing the efficient applications of chemical fertilizers. In other impoverished, as well as rich, parts of the planet, decades of intensive agriculture under irrigation have led to salinization, water-logging, eutrophication of major water bodies, dangerous declines of biodiversity and other forms of environmental degradation. These enormous strains are coupled with the continuing pressures of tropical deforestation and the lack of new promising regions for expanding crop cultivation to meet the needs of growing populations. Finally, there looms the prospect of anthropogenic climate change. Global warming and associated complex and poorly understood shifts in precipitation extremes and other climate variables all threaten the world's natural ecosystems and food production systems in profound yet still imperfectly understood ways. The risks of gradual or abrupt climate change are coupled with the risks of drastic perturbations to regional and global food supplies.

The *Encyclopedia* offers state-of-the-art contributions on each of these challenges, as well as links to entries on the fundamental biophysical processes that underpin the relevant phenomena. The world-scale and world-class collaboration that stands behind this unique project signifies its importance for the world community. It is an honor and privilege for me to introduce this path-breaking endeavor.

Jeffrey D Sachs
Director
The Earth Institute at Columbia University
Quetelet Professor of Sustainable Development
Columbia University, New York, USA

# PREFACE

The term 'soil' refers to the weathered and fragmented outer layer of our planet's land surfaces. Formed initially through the physical disintegration and chemical alteration of rocks and minerals by physical and biogeochemical processes, soil is influenced by the activity and accumulated residues of a myriad of diverse forms of life. As it occurs in different geologic and climatic domains, soil is an exceedingly variegated body with a wide range of attributes.

Considering the height of the atmosphere, the thickness of the earth's rock mantle, and the depth of the ocean, one observes that soil is an amazingly thin body – typically not much more than one meter thick and often less than that. Yet it is the crucible of terrestrial life, within which biological productivity is generated and sustained. It acts like a composite living entity, a home to a community of innumerable microscopic and macroscopic plants and animals. A mere fistful of soil typically contains billions of microorganisms, which perform vital interactive biochemical functions. Another intrinsic attribute of the soil is its sponge-like porosity and its enormous internal surface area. That same fistful of soil may actually consist of several hectares of active surface, upon which physicochemical processes take place continuously.

Realizing humanity's utter dependence on the soil, ancient peoples, who lived in greater intimacy with nature than many of us today, actually revered the soil. It was not only their source of livelihood, but also the material from which they built their homes and that they learned to shape, heat, and fuse into household vessels and writing tablets (ceramic, made of clayey soil, being the first synthetic material in the history of technology). In the Bible, the name assigned to the first human was Adam, derived from 'adama,' meaning soil. The name given to that first earthling's mate was Hava (Eve, in transliteration), meaning 'living' or 'life-giving.' Together, therefore, Adam and Eve signified quite literally 'Soil and Life.'

The same powerful metaphor is echoed in the Latin name for the human species – Homo, derived from humus, the material of the soil. Hence, the adjective 'human' also implies 'of the soil.' Other ancient cultures evoked equally powerful associations. To the Greeks, the earth was a manifestation of Gaea, the maternal goddess who, impregnated by Uranus (god of the sky), gave birth to all the gods of the Greek pantheon.

Our civilization depends on the soil more crucially than ever, because our numbers have grown while available soil resources have diminished and deteriorated. Paradoxically, however, even as our dependence on the soil has increased, most of us have become physically and emotionally detached from it. Many of the people in the so-called 'developed' countries spend their lives in the artificial environment of a city, insulated from direct exposure to nature, and some children may now assume as a matter of course that food originates in supermarkets.

Detachment has bred ignorance, and out of ignorance has come the delusion that our civilization has risen above nature and has set itself free of its constraints. Agriculture and food security, erosion and salination, degradation of natural ecosystems, depletion and pollution of surface waters and aquifers, and decimation of biodiversity – all of these processes, which involve the soil directly or indirectly – have become abstractions to many people. The very language we use betrays disdain for that common material underfoot, often referred to as 'dirt.' Some fastidious parents prohibit their children from playing in the mud and hurry to wash their 'soiled' hands when the children nonetheless obey an innate instinct to do so. Thus soil is devalued and treated as unclean though it is the terrestrial realm's principal medium of purification, wherein wastes are decomposed and nature's productivity is continually rejuvenated.

Scientists who observe soil closely see it in effect as a seething foundry in which matter and energy are in constant flux. Radiant energy from the sun streams onto the field and cascades through the soil and the plants growing in it. Heat is exchanged, water percolates through the soil's intricate passages, plant roots extract water and transmit it to their leaves, which transpire it back to the atmosphere. Leaves absorb carbon dioxide from the air and synthesize it with soil-derived water to form the primary compounds of life. Oxygen emitted by the leaves makes the air breathable for animals, which consume and in turn fertilize plants.

Soil is thus a self-regulating bio-physio-chemical factory, processing its own materials, water, and solar energy. It also determines the fate of rainfall and snowfall reaching the ground surface – whether the water thus received will flow over the land as runoff, or seep downward to the subterranean reservoir called groundwater, which in turn maintains the steady flow of springs and streams. With its finite capacity to absorb and store moisture, and to release it gradually, the soil regulates all of these phenomena. Without the soil as a buffer, rain falling over the continents would run off entirely, producing violent floods rather than sustained river flow.

Soil naturally acts as a living filter, in which pathogens and toxins that might otherwise accumulate to foul the terrestrial environment are rendered harmless. Since time immemorial, humans and other animals have been dying of all manner of disease and have then been buried in the soil, yet no major disease is transmitted by it. The term *antibiotic* was coined by soil microbiologists who, as a consequence of their studies of soil bacteria and actinomycetes, discovered streptomycin (an important cure for tuberculosis and other infections). Ion exchange, a useful process of water purification, also was discovered by soil scientists studying the passage of solutes through beds of clay.

However unique in form and function, soil is not an isolated body. It is, rather, a central link in the larger chain of interconnected domains and processes comprising the terrestrial environment. The soil interacts both with the overlying atmosphere and the underlying strata, as well as with surface and underground bodies of water. Especially important is the interrelation between the soil and the climate. In addition to its function of regulating the cycle of water, it also regulates energy exchange and surface temperature.

When virgin land is cleared of vegetation and turned into a cultivated field, the native biomass above the ground is often burned and the organic matter within the soil tends to decompose. These processes release carbon dioxide into the atmosphere, thus contributing to the earth's greenhouse effect and to global warming. On the other hand, the opposite act of reforestation and soil enrichment with organic matter, such as can be achieved by means of conservation management, may serve to absorb carbon dioxide from the atmosphere. To an extent, the soil's capacity to store carbon can thus help to mitigate the greenhouse effect.

Thousands of years are required for nature to create life-giving soil out of sterile bedrock. In only a few decades, however, unknowing or uncaring humans can destroy that wondrous work of nature. In various circumstances, mismanaged soils may be subject to erosion (the sediments of which tend to clog streambeds, estuaries, lakes, and coastal waters), to leaching of nutrients with attendant loss of fertility and eutrophication of water bodies, to waterlogging and impaired aeration, or to an excessive accumulation of salts that may cause a once-productive soil to become entirely sterile. Such processes of soil degradation, sometimes called 'desertification,' already affect large areas of land.

We cannot manage effectively and sustainably that which we do not know and thoroughly understand. That is why the tasks of developing and disseminating sound knowledge of the soil and its complex processes have assumed growing urgency and importance. The global environmental crisis has created a compelling need for a concentrated, concise, and definitive source of information – accessible to students, scientists, practitioners, and the general public – about the soil in all its manifestations – in nature and in relation to the life of humans.

Daniel Hillel
Editor-in-Chief
May 2004

# INTRODUCTION

The *Encyclopedia of Soils in the Environment* contains nearly 300 articles, written by the world's leading authorities. Pedologists, biologists, ecologists, earth scientists, hydrologists, climatologists, geographers, and representatives from many other disciplines have contributed to this work. Each of the articles separately, and all of them in sequence and combination, serve to summarize and encapsulate our present knowledge of the world's variegated soils, their natural functions, and their importance to humans.

Concise articles surveying specific aspects of soils (soil genesis, soil chemistry and mineralogy, soil physics and hydrology, and soil biology) are complemented by articles covering transdisciplinary aspects, such as the role of soils in ecology, the history of soil utilization for agricultural and engineering purposes, the development of soil science as a discipline, and the potential or actual contributions of soils to the generation, as well as to the mitigation, of pollution and of global climate change.

This comprehensive reference encompasses both the fundamental and the applied aspects of soil science, interfacing in general with the physical sciences and life sciences and more specifically with the earth sciences and environmental sciences.

The *Encyclopedia of Soils in the Environment* manifests the expanding scope of modern soil science, from its early sectarian focus on the utilitarian attributes of soils in agriculture and engineering, to a wider and much more inclusive view of the soil as a central link in the continuous chain of processes constituting the dynamic environment as a whole. Thus it both details and integrates a set of topics that have always been of vital importance to human societies and that are certain to be even more so in the future.

Daniel Hillel
Editor-in-Chief
May 2004

# GUIDE TO USE OF THE ENCYCLOPEDIA

## Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of entries in alphabetical order. Some entries consist of several articles that deal with various aspects of a topic and are arranged in a logical sequence within an entry. Other entries comprise a single article.

To help you realize the full potential of the material in the Encyclopedia we have provided three features to help you find the topic of your choice: a Contents List, Cross-References and an Index.

## 1. Contents List

Your first point of reference will probably be the contents list. The complete contents list will provide you with both the volume number and the page number of the entry. On the opening page of an entry a contents list is provided so that the full details of the articles within the entry are immediately available.

Alternatively you may choose to browse through a volume using the alphabetical order of the entries as your guide. To assist you in identifying your location within the Encyclopedia a running headline indicates the current entry and the current article within that entry.

You will find 'dummy entries' where obvious synonyms exist for entries or where we have grouped together related topics. Dummy entries appear in both the contents list and the body of the text.

*Example*
If you were attempting to locate material on Water Erosion via the contents list:

WATER EROSION *see* EROSION: Water-Induced

The dummy entry directs you to the Water-Induced article, in the Erosion entry. At the appropriate location in the contents list, the page numbers for articles under Erosion are given.

If you were trying to locate the material by browsing through the text and you looked up Water Erosion then the following information would be provided in the dummy entry:

**Water Erosion** *see* **Erosion:** Water-Induced

Alternatively, if you were looking up Erosion the following information would be provided:

# EROSION

Contents
**Irrigation-Induced**
**Water-Induced**
**Wind-Induced**

## 2. Cross References

All of the articles in the Encyclopedia have been extensively cross referenced. The cross references, which appear at the end of an article, have been provided at three levels:

i. To indicate if a topic is discussed in greater detail elsewhere

> *See also:* **Acid Rain and Soil Acidification**; **Archaeology in Relation to Soils**; **Carbon Emissions and Sequestration**; **Civilization, Role of Soils**; **Classification of Soils**; **Classification Systems:** USA; **Degradation**; **Desertification**; **Erosion:** Water-Induced; **Heavy Metals**; **Morphology**; **Organic Matter:** Genesis and Formation; **Pedology: Principles**; **Quality of Soil**; **Salination Processes**; **Structure**

ii. To draw the reader's attention to parallel discussions in other articles

> *See also:* **Acid Rain and Soil Acidification**; **Archaeology in Relation to Soils**; **Carbon Emissions and Sequestration**; **Civilization, Role of Soils**; **Classification of Soils**; **Classification Systems:** USA; **Degradation**; **Desertification**; **Erosion: Water-Induced**; **Heavy Metals**; **Morphology**; **Organic Matter:** Genesis and Formation; **Pedology:** Principles; **Quality of Soil**; **Salination Processes**; **Structure**

iii. To indicate material that broadens the discussion

> *See also:* **Acid Rain and Soil Acidification**; **Archaeology in Relation to Soils**; **Carbon Emissions and Sequestration**; **Civilization, Role of Soils**; **Classification of Soils**; **Classification Systems: USA**; **Degradation**; **Desertification**; **Erosion:** Water-Induced; **Heavy Metals**; **Morphology**; **Organic Matter: Genesis and Formation**; **Pedology:** Principles; **Quality of Soil**; **Sallination Processes**; **Structure**

## 3. Index

The index will provide you with the volume number and page number of where the material is to be located, and the index entries differentiate between material that is a whole article, is part of an article or is data presented in a table. Detailed notes are provided on the opening page of the index.

## 4. Color Plates

The color figures for each volume have been grouped together in a plate section. The location of this section is cited in the contents list. Color versions of black and white figures are cited in figure captions within individual articles.

## 5. Contributors

A full list of contributors appears at the beginning of each volume.

# CONTRIBUTORS

**C L Acharya**
Indian Institute of Soil Science, Bhopal,
Madhya Pradesh, India

**J F Adams**
Auburn University, AL, USA

**T M Addiscott**
Rothamsted Research, Harpenden, UK

**D C Adriano**
University of Georgia, Aiken, SC, USA

**L Agnew**
Cornell University, Ithaca, NY, USA

**M F Allen**
University of Califonia – Riverside,
Riverside, CA, USA

**R Allen**
University of Idaho, Kimberly, ID, USA

**I Alves**
International Commission on Agricultural
Engineering (CIGR), Lisbon, Portugal

**J A Amador**
University of Rhode Island, Kingston, RI, USA

**R Amundson**
University of California–Berkeley, Berkeley,
CA, USA

**M C Anderson**
University of Wisconsin, Madison, WI, USA

**G L Anderson**
USDA Agricultural Research Service, Sidney, MT, USA

**R W Arnold**
Formerly with USDA Natural Resources Conservation
Service, Washington, DC, USA

**W Böhm**
Georg-August-University, Göttingen, Germany

**J Böhner**
Universität Göttingen, Göttingen, Germany

**O O S Bacchi**
Center for Nuclear Energy in Agriculture,
Piracicaba, Brazil

**Y Bachmat**
Hebrew University, Jerusalem, Israel

**V C Baligar**
USDA Agricultural Research Service, Beltsville, MD, USA

**T C Balser**
University of Wisconsin–Madison, Madison, WI, USA

**K K Bandyopadhyay**
Indian Institute of Soil Science, Bhopal,
Madhya Pradesh, India

**D A Barry**
University of Edinburgh, Edinburgh, UK

**C D Barton**
University of Kentucky, Lexington,
KY, USA

**Y Bashan**
Center for Biological Research of the Northwest (CIB), La Paz, Mexico

**R L Baumhardt**
USDA Agricultural Research Service, Bushland, TX, USA

**P H Bellamy**
Cranfield University, Silsoe, UK

**J Belnap**
USGS Southwest Biological Science Center, Moab, UT, USA

**J L Berc**
USDA Natural Resources Conservation Service, Washington, DC, USA

**T J Beveridge**
University of Guelph, Guelph, ON, Canada

**G D Binford**
University of Delaware, Newark, DE, USA

**D L Bjorneberg**
USDA Agricultural Research Service, Kimberly, ID, USA

**L Blake**
Rothamsted Research, Harpenden, UK

**C W Boast**
University of Illinois, Urbana-Champaign, IL, USA

**J L Boettinger**
Utah State University, Logan, UT, USA

**N S Bolan**
Massey University, Palmerston North, New Zealand

**J V Bonta**
USDA Agricultural Research Service, Coshocton, OH, USA

**J R Boyle**
Oregon State University, Corvallis, OR, USA

**R H Bromilow**
Rothamsted Research, Harpenden, UK

**S M Brouder**
Purdue University, West Lafayette, IN, USA

**D J Brown**
Montana State University, Bozeman, MT, USA

**D D Buhler**
Michigan State University, East Lansing, MI, USA

**P Bullock**
Cranfield University–Silsoe, Silsoe, UK

**S W Buol**
North Carolina State University, Raleigh, NC, USA

**E P Burford**
University of Dundee, Dundee, UK

**J Burney**
Dalhousie University, Halifax, NS, Canada

**C L Burras**
Iowa State University, Ames, IA, USA

**A J Busacca**
Washington State University, Pullman, WA, USA

**M L Cabrera**
University of Georgia, Athens, GA, USA

**C A Cambardella**
USDA Agricultural Research Service, Ames, USA

**C S Campbell**
Decagon Devices, Inc., Pullman, WA, USA

**G S Campbell**
Decagon Devices, Inc., Pullman, WA, USA

**P G Carter**
Purdue University, West Lafayette, IN, USA

**M R Carter**
Agriculture and Agri-Food Canada, Charlottetown, PE, Canada

**D K Cassel**
North Carolina State University, Raleigh, NC, USA

**M A Chappell**
Iowa State University, Ames, IA, USA

**J Chorover**
University of Arizona, Tucson, AZ, USA

**B E Clothier**
HortResearch, Palmerston North, New Zealand

**R Corstanje**
University of Florida, Gainesville, FL, USA

**M S Coyne**
University of Kentucky, Lexington, KY, USA

**J W Crawford**
SIMBIOS, University of Abertay, Dundee, Scotland, UK

**D Curtin**
New Zealand Institute for Crop and Food Research,
Christchurch, New Zealand

**V D'Orazio**
Università di Bari, Bari, Italy

**J H Dane**
Auburn University, Auburn, AL, USA

**S M de Jong**
Utrecht University, Utrecht, The Netherlands

**C de Kimpe**
Agriculture and Agri-Food Canada, Ottawa, ON, Canada

**L Z de Luna**
USDA Agricultural Research Service, Pullman, WA, USA

**P C de Ruiter**
University of Utrecht, Utrecht, The Netherlands

**L E de-Bashan**
Center for Biological Research of the Northwest (CIB),
La Paz, Mexico

**J A Deckers**
Catholic University of Leuven, Leuven, Belgium

**R D DeLaune**
Louisiana State University, Baton Rouge, LA, USA

**D A DiCarlo**
USDA Agricultural Research Service, Oxford, MS, USA

**R P Dick**
Ohio State University, Columbus, OH, USA

**J W Doran**
USDA Agricultural Research Service, Lincoln, NE, USA

**J Dragun**
Dragun Corporation, Farmington Hills, MI, USA

**L Edwards**
Agriculture and Agri-Food Canada, Charlottetown,
PE, Canada

**L M Egerton-Warburton**
Chicago Botanic Garden, Glencoe, IL, USA

**J A Entry**
USDA Agricultural Research Service, Kimberly, ID, USA

**H Eswaran**
USDA Natural Resources Conservation Service,
Washington, DC, USA

**V P Evangelou**[†]
Iowa State University, Ames, IA, USA

**R P Ewing**
Iowa State University, Ames, IA, USA

**N K Fageria**
EMBRAPA, Santo Antônio de Goiás, GO, Brazil

**N R Fausey**
USDA Agricultural Research Service, Columbus,
OH, USA

**M J Fayer**
Pacific Northwest National Laboratory, Richland,
WA, USA

**R A Feddes**
Wageningen University and Research Centre,
Wageningen, The Netherlands

**S Fendorf**
Stanford University, Stanford, CA USA

**T P A Ferré**
University of Arizona, Tucson, AZ, USA

**N Fierer**
University of California–Santa Barbara, Santa Barbara,
CA, USA

**S L Finkelman**
Containerless Research, Inc., Evanston, IL, USA

**R W Fitzpatrick**
CSIRO Land and Water
Glen Osmond, SA, Australia

**G N Flerchinger**
USDA Agricultural Research Service, Boise, ID, USA

**L Florentin**
Laboratoire Sols et Environnement ENSAIA-INPL/INRA,
France

[†] Deceased.

**W Foissner**
Universität Salzburg, Salzburg, Austria

**G R Foster**
Bryan, TX, USA

**C A Francis**
University of Nebraska, Lincoln, NE, USA

**A J Franzluebbers**
USDA Agricultural Research Service, Watkinsville, GA, USA

**S D Frey**
University of New Hampshire, Durham, NC, USA

**W W Frye**
University of Kentucky, Lexington, KY, USA

**M Fuchs**
Agricultural Research Organization,
Bet Dagan, Israel

**M E Fuller**
Shaw Environmental, Inc.,
Lawrenceville, NJ, USA

**P Gérard-Marchant**
Cornell University, Ithaca, NY, USA

**J H Görres**
University of Rhode Island, Kingston, RI, USA

**G M Gadd**
University of Dundee, Dundee, UK

**M Gallardo**
University of Almeria, Almeria, Spain

**A Gamliel**
The Volcani Center, Bet Dagan, Israel

**W H Gardner**
Washington State University, Pullman, WA, USA

**W R Gardner**
Texas A & M University, USA

**G W Gee**
Pacific Northwest National Laboratory, Richland, WA, USA

**M Gerasimova**
Moscow State University, Moscow, Russia

**C P Gerba**
University of Arizona, Tucson, AZ, USA

**P F Germann**
University of Bern, Bern, Switzerland

**J J Germida**
University of Saskatchewan, Saskatoon, SK, Canada

**J E Gilley**
USDA Agricultural Research Service, Lincoln, NE, USA

**C Gimenez**
University of Cordoba, Cordoba, Spain

**S M Glasauer**
University of Guelph, Guelph, ON, Canada

**R J Glass**
Sandia National Laboratory,
Albuquerque, NM, USA

**A Globus**
Agrophysical Research Institute,
St Petersburg, Russia

**A J Gold**
University of Rhode Island, Kingston, RI, USA

**S Goldberg**
USDA Agricultural Research Service, Riverside, CA, USA

**S A Grant**
US Army Engineer Research and Development Center,
Hanover, NH, USA

**S Green**
HortResearch, Palmerston North, New Zealand

**P H Groenevelt**
University of Guelph, Guelph, ON, Canada

**G Guggenberger**
University of Bayreuth, Bayreuth, Germany

**A Hadas**
The Volcani Center, Bet Dagan, Israel

**S M Haefele**
International Rice Research Institute, Metro Manila, Philippines

**K M Haider**
University Halle, Halle, Germany

**L J Halverson**
Iowa State University, Ames, IA, USA

**R E Hamon**
CSIRO Land and Water, Glen Osmond, SA, Australia

**F A Harper**
University of Dundee, Dundee, UK

**J Harsh**
Washington State University, Pullman, WA, USA

**P G Hartel**
University of Georgia, Athens, GA, USA

**E K Hartig**
New York, NY, USA

**J L Hatfield**
National Soil Tilth Laboratory, Ames, IA, USA

**K M Hati**
Indian Institute of Soil Science, Bhopal,
Madhya Pradesh, India

**P B Hatzinger**
Shaw Environmental, Inc., Lawrenceville, NJ, USA

**J L Havlin**
North Carolina State University, Raleigh, NC, USA

**J D Helms**
USDA Natural Resources Conservation Service,
Washington, DC, USA

**D Hillel**
Columbia University, New York, NY, USA

**A Hodge**
University of York, York, UK

**P A Holden**
University of California–Santa Barbara, Santa Barbara,
CA, USA

**J A Homburg**
Statistical Research Inc., Tucson, AZ, USA

**J W Hopmans**
University of California–Davis,
Davis, CA, USA

**R Horton**
Iowa State University, Ames, IA, USA

**T A Howell**
USDA Agricultural Research Service, Bushland, TX, USA

**P M Huang**
University of Saskatchewan, Saskatoon, SK, Canada

**G W Hurt**
University of Florida, Gainesville, FL, USA

**C A Impellitteri**
USEPA, Cincinnati, OH, USA

**P W Inglett**
University of Florida, Gainesville, FL, USA

**L E Jackson**
University of California–Davis, Davis, CA, USA

**T J Jackson**
USDA Agricultural Research Service, Beltsville,
MD, USA

**A H Jahren**
Johns Hopkins University, Baltimore, MD, USA

**B R James**
University of Maryland, College Park, MD, USA

**P M Jardine**
Oak Ridge National Laboratory, Oak Ridge, TN, USA

**A T Jeffries**
National Soil Tilth Laboratory, Ames, IA, USA

**V G Jetten**
Utrecht University, Utrecht, The Netherlands

**C J Johannsen**
Purdue University, West Lafayette, IN, USA

**D Johnson**
Rothamsted Research, Harpenden, UK

**A E Johnston**
Rothamsted Research, Harpenden, UK

**R Köthe**
Scilands, Gesellschaft zur Bearbeitung digitaler
Landschaften, Göttingen, Germany

**U Kafkafi**
The Hebrew University of Jerusalem, Rehovot, Israel

**E J Kamprath**
North Carolina State University, Raleigh, NC, USA

**K Kandaswamy**
Massey University, Palmerston North, New Zealand

**E Kandeler**
University of Hohenheim, Stuttgart, Germany

**S Kant**
The Hebrew University of Jerusalem, Rehovot, Israel

**D L Karlen**
USDA Agricultural Research Service, Ames, IA, USA

**J Katan**
The Hebrew University of Jerusalem, Rehovot, Israel

**E F Kelly**
Colorado State University, Fort Collins, CO, USA

**J W Kelsey**
Muhlenberg College, Allentown, PA, USA

**A C Kennedy**
USDA Agricultural Research Service, Pullman,
WA, USA

**R Keren**
The Volcani Center, Bet Dagan, Israel

**R Khaleel**
Fluor Federal Services, Richland,
WA, USA

**S A Khresat**
Jordan University of Science and Technology,
Irbid, Jordan

**Y-J Kim**
Cornell University, Ithaca, NY, USA

**C Kirda**
University of Cukurova, Adana, Turkey

**E A Kirkby**
University of Leeds, Leeds, UK

**M B Kirkham**
Kansas State University, Manhattan, KS, USA

**D E Kissel**
University of Georgia, Athens, GA, USA

**D Konyushkov**
V.V. Dokuchaev Soil Science Institute, Moscow, Russia

**B-J Koo**
University of Georgia, Aiken, SC, USA

**H Kosegarten**
Justus Liebig University, Giessen, Germany

**P Kostecki**
University of Massachusetts, Amherst, MA, USA

**R Kretzschmar**
Institute of Terrestrial Ecology, Zurich, Switzerland

**J A LaGro Jr**
University of Wisconsin-Madison, Madison, WI, USA

**R Lal**
Ohio State University, Columbus, OH, USA

**R L Lavigne**
University of Massachusetts, Leeds, MA, USA

**P J Lawrence**
USDA Natural Resources Conservation Service,
Washington, DC, USA

**P Leeds-Harrison**
Cranfield University, Silsoe, Bedford, UK

**G A Lehrsch**
USDA Agricultural Research Service, Kimberly, ID, USA

**R J Lenhard**
Idaho National Engineering Environmental Laboratory,
Idaho Falls, ID, USA

**J Letey**
University of California–Riverside, Riverside, CA, USA

**M J Levin**
USDA Natural Resources Conservation Service,
Washington, DC, USA

**G J Levy**
The Volcani Center, Bet Dagan, Israel

**E Loffredo**
Università di Bari, Bari, Italy

**P Loganathan**
Massey University, Palmerston North, New Zealand

**E Lombi**
CSIRO Land and Water, Glen Osmond, SA, Australia

**P J Loveland**
Cranfield University, Silsoe, UK

**J M Lynch**
University of Surrey, Guildford, UK

**R O Maguire**
North Carolina State University, Raleigh, NC, USA

**A P Mallarino**
Iowa State University, Ames, IA, USA

**D A C Manning**
University of Newcastle, Newcastle upon Tyne, UK

**D A Martens**
USDA Agricultural Research Service, Tucson, AZ, USA

**J J Martinez-Rios**
Universidad Juarez del Estado de Durango, Mexico

**J A Mason**
University of Wisconsin–Madison, Madison, WI, USA

**C J Matocha**
University of Kentucky, Lexington, KY, USA

**A B McBratney**
University of Sydney, Sydney, NSW, Australia

**D K McCool**
USDA Agricultural Research Service, Pullman, WA, USA

**L M McDonald**
West Virginia University, Morgantown, WV, USA

**S P McGrath**
Rothamstead Research, Harpenden, UK

**E McKyes**
McGill University, Ste-Anne de Bellevue, PQ, Canada

**J E T McLain**
USDA Agricultural Research Service, Tucson, AZ, USA

**K Mengel**
Justus Liebig University, Giessen, Germany

**P D Meyer**
Pacific Northwest National Laboratory, Richland,
WA, USA

**D L Mokma**
Michigan State University, East Lansing, MI, USA

**H C Monger**
New Mexico State University, Las Cruces,
New Mexico, USA

**J L Monteith**
Edinburgh, UK

**J C Moore**
University of Northern Colorado,
Greeley, CO, USA

**J L Morel**
Laboratoire Sols et Environnement ENSAIA-INPL/INRA,
France

**D K Morris**
Stennis Space Center, MS, USA

**R Morrison**
DPRA, Inc., San Marcos, CA, USA

**M D Mullen**
University of Kentucky, Lexington, KY, USA

**F O Nachtergaele**
FAO, Rome, Italy

**P K R Nair**
University of Florida, Gainesville, FL, USA

**C H Nakatsu**
Purdue University, West Lafayette, IN, USA

**E A Nater**
University of Minnesota, St. Paul, MN, USA

**D A Neher**
University of Toledo, Toledo, OH, USA

**P A Nektarios**
Agricultural University of Athens, Athens, Greece

**P S Nico**
California State University–Stanislaus, Turlock, CA, USA

**D R Nielsen**
University of California, Davis, CA, USA

**J R Nimmo**
US Geological Survey, Menlo Park, CA, USA

**J M Norman**
University of Wisconsin, Madison, WI, USA

**I O A Odeh**
University of Sydney, Sydney, NSW, Australia

**K R Olson**
University of Illinois, Urbana, IL, USA

**M Oostrom**
Pacific Northwest National Laboratory, Richland, WA, USA

**D Or**
University of Connecticut, Storrs, CT, USA

**W J Orts**
USDA Agricultural Research Service, Albany, CA, USA

**P R Owens**
USDA Agricultural Research Service, Mississippi State, MS, USA

**Y Pachepsky**
USDA Agricultural Research Service, Beltsville, MD, USA

**A Palmer**
Massey University, Palmerston North, New Zealand

**N S Panikov**
Stevens Institute of Technology, Hoboken, NJ, USA

**C E Pankhurst**
CSIRO Land and Water, Townsville, QLD, Australia

**D R Parker**
University of California, Riverside, CA, USA

**J-Y Parlange**
Cornell University, Ithaca, NY, USA

**K Paustian**
Colorado State University, Fort Collins, CO, USA

**D Peak**
University of Saskatchewan, Saskatoon, SK, Canada

**K D Pennell**
School of Civil & Environmental Engineering, Atlanta, GA, USA

**L S Pereira**
International Commission on Agricultural Engineering (CIGR), Lisbon, Portugal

**E Perfect**
University of Tennessee, Knoxville, TN, USA

**G A Peterson**
Colorado State University, Fort Collins, CO, USA

**J J Pignatello**
Connecticut Agricultural Experimental Station, New Haven, CT, USA

**C-L Ping**
University of Alaska Fairbanks, Palmer, AK, USA

**T O Powers**
University of Nebraska, Lincoln, NE, USA

**D S Powlson**
Rothamsted Research, Harpenden, UK

**J I Prosser**
University of Aberdeen, Aberdeen, UK

**J H Prueger**
National Soil Tilth Laboratory, Ames, IA, USA

**M Qadir**
International Center for Agricultural Research in the Dry Areas (ICARDA), Aleppo, Syria; and International Water Management Institute (IWMI), Colombo, Sri Lanka

**J I Querejeta**
University of California – Riverside, Riverside, CA, USA

**P A C Raats**
Wageningen University, Wageningen, The Netherlands

**M Radosevich**
University of Tennessee, Knoxville, TN, USA

**R L Raper**
USDA Agricultural Research Service, Auburn, AL, USA

**K R Reddy**
University of Florida, Gainesville, FL, USA

**P F Reich**
USDA Natural Resources Conservation Service, Washington, DC, USA

**K Reichardt**
Center for Nuclear Energy in Agriculture, Piracicaba, Brazil

**D C Reicosky**
USDA Agricultural Research Service, Morris, MN, USA

**C W Rice**
Kansas State University, Manhattan, KS, USA

**T L Richard**
Iowa State University, Ames, IA, USA

**M C Rillig**
University of Montana, Missoula, MT, USA

**C J Ritsema**
Alterra, Wageningen, The Netherlands

**K Ritz**
Cranfield University, Silsoe, MK45 4DT, UK

**W P Robarge**
North Carolina State University, Raleigh, NC, USA

**D E Rolston**
University of California–Davis, Davis, CA, USA

**C Rosenzweig**
NASA Goddard Institute for Space Studies,
New York, NY, USA

**W F Rowe**
George Washington University, Washington, DC, USA

**H Rubin**
Technion, Israel Institute of Technology, Haifa, Israel

**D Russo**
The Volcani Center, Bet Dagan, Israel

**E M Rutledge**
University of Arkansas, Fayetteville, AR, USA

**S Saggar**
Landcare Research, Palmerston North, New Zealand

**J Sandor**
Iowa State University, Ames, IA, USA

**T J Sauer**
National Soil Tilth Laboratory, Ames, IA, USA

**K G Scheckel**
USEPA, Cincinnati, OH, USA

**A C Scheinost**
Swiss Federal Institute of Technology (ETH), Zurich,
Switzerland

**H W Scherer**
University of Bonn, Bonn, Germany

**M Schnitzer**
Agriculture and Agri-Food Canada, Ottawa, ON, Canada

**S Schubert**
Justus Liebig University, Giessen, Germany

**C P Schulthess**
University of Connecticut, Storrs, CT, USA

**D G Schulze**
Purdue University, West Lafayette, IN, USA

**A P Schwab**
Purdue University, West Lafayette, IN, USA

**C Schwartz**
Laboratoire Sols et Environnement ENSAIA-INPL/INRA,
France

**R C Schwartz**
USDA Agricultural Research Service, Bushland,
TX, USA

**T Selige**
Technische Universität München, Munich, Germany

**J S Selker**
Oregon State University, Corvallis, OR, USA

**G S Senesi**
Università di Bari, Bari, Italy

**N Senesi**
Università di Bari, Bari, Italy

**I Shainberg**
The Volcani Center, Bet Dagan, Israel

**B Sharratt**
USDA Agricultural Research Service, Pullman, WA, USA

**L M Shuman**
University of Georgia, Griffin, GA, USA

**S D Siciliano**
University of Saskatchewan, Saskatoon, SK, Canada

**F W Simmons**
University of Illinois, Urbana-Champaign, IL, USA

**J T Sims**
University of Delaware, Newark, DE, USA

**M J Singer**
University of California–Davis, Davis, CA, USA

**D Smiles**
CSIRO Land and Water, Canberra, Australia

**P Smith**
University of Aberdeen, Aberdeen, UK

**K A Smith**
University of Edinburgh, Edinburgh, UK

**T J Smyth**
North Carolina State University, Raleigh, NC, USA

**V A Snyder**
University of Puerto Rico Agricultural
Experiment Station, San Juan,
Puerto Rico

**B Soane**
Formerly with Scottish Agricultural College,
Edinburgh, UK

**R E Sojka**
USDA Agricultural Research Service, Kimberly,
ID, USA

**O C Spaargaren**
ISRIC/WDC for Soils, Wageningen, The Netherlands

**D L Sparks**
University of Delaware, Newark, DE, USA

**G Sposito**
University of California–Berkeley, Berkeley,
CA, USA

**J I Sprent**
University of Dundee, Dundee, UK

**F Stagnitti**
Deakin University, Warrnambool, VIC, Australia

**G Stanhill**
The Volcani Center, Bet Dagan, Israel

**T S Steenhuis**
Cornell University, Ithaca, NY, USA

**D Steffens**
Justus Liebig University, Giessen, Germany

**G C Steinhardt**
Purdue University, West Lafayette, IN, USA

**M C Sukop**
Florida International University, Miami, FL, USA

**D Swartzendruber**
University of Nebraska-Lincoln, Lincoln, NE, USA

**M R Sweeney**
Washington State University, Pullman, WA, USA

**M A Tabatabai**
Iowa State University, Ames, IA, USA

**M Tagliavini**
University of Bologna, Bologna, Italy

**E S Takle**
Iowa State University, Ames, IA, USA

**J P Tandarich**
Hey and Associates, Inc., Chicago, IL, USA

**J C F Tedrow**
Formerly with Rutgers University, New Brunswick,
NJ, USA

**B B Thapa**
North Carolina State University, Raleigh, NC, USA

**M Thompson**
Iowa State University, Ames, IA, USA

**R B Thompson**
University of Almeria, Almeria, Spain

**T K Tokunaga**
E.O. Lawrence Berkeley National Laboratory, Berkeley,
CA, USA

**M D Tomer**
USDA Agricultural Research Service, Ames, IA, USA

**S Topcu**
University of Cukurova, Adana, Turkey

**G C Topp**
Agriculture and Agri-Food Canada, Ottawa,
ON, Canada

**J Torrent**
Universidad de Córdoba, Córdoba, Spain

**M Toselli**
University of Bologna, Bologna, Italy

**H Tsoar**
Ben Gurion University of the Negev,
Beer Sheva, Israel

**M Tuller**
University of Idaho, Moscow, ID, USA

**G Uehara**
University of Hawaii, Honolulu, HI, USA

**F C Ugolini**
Università degli Studi di Firenze, Firenze, Italy

**A L Ulery**
New Mexico State University, Las Cruces,
NM, USA

**S K Upadhyaya**
University of California–Davis, Davis, CA, USA

**M A Vázquez**
University of Puerto Rico Agricultural
Experiment Station, San Juan,
Puerto Rico

**P A Vadas**
USDA Agricultural Research Service,
University Park, PA, USA

**J C van Dam**
Wageningen University and Research Centre,
Wageningen, The Netherlands

**J J H van den Akker**
Alterra Wageningen UR, Wageningen,
The Netherlands

**R R van der Ploeg**
University of Hannover, Hannover, Germany

**M Th van Genuchten**
US Salinity Laboratory, Riverside, CA, USA

**G W J van Lynden**
ISRIC, Wageningen, The Netherlands

**R S Van Pelt**
USDA Agricultural Research Service, Big Spring,
TX, USA

**J Vangronsveld**
Limburgs Universitair Centrum,
Diepenbeek, Belgium

**G E Varvel**
USDA Agricultural Research Service, Lincoln,
NE, USA

**D H Wall**
Colorado State University, Fort Collins, CO, USA

**M T Walter**
Cornell University, Ithaca, NY, USA

**M C Wang**
National Chung Hsing University, Taichung, Taiwan

**M K Wang**
National Taiwan University, Taipei, Taiwan

**A L Ward**
Pacific Northwest National Laboratory, Richland,
WA, USA

**A W Warrick**
University of Arizona, Tucson, AZ, USA

**R Webster**
Rothamsted Research, Harpenden, UK

**W W Wenzel**
Universität für Bodenkultur, Wien, Austria

**B J Wienhold**
USDA Agricultural Research Service, Lincoln, NE, USA

**P Wierenga**
University of Arizona, Tucson, AZ, USA

**H Wijnja**
Connecticut Agricultural Experiment Station,
New Haven, CT, USA

**K E Williamson**
University of Delaware, Newark, DE, USA

**A R Wilts**
USDA Agricultural Research Service, Morris, MN, USA

**T Winsome**
University of California–Davis, Davis, CA, USA

**C Witt**
Potash & Phosphate Institute/Potash & Phosphate
Institute of Canada and International Potash Institute,
Singapore

**K E Wommack**
University of Delaware, Newark, DE, USA

**B H Wood**
Auburn University, AL, USA

**C W Wood**
Auburn University, AL, USA

**H B Woodruff**
Soil Microbiology Associates, Watchung, NJ, USA

**J M Wraith**
Montana State University, Bozeman, MT, USA

**J C Xie**
The Institute of Soil Science, Academia Sinica, Nanjing, People's Republic of China

**B Xing**
University of Massachusetts, Amherst, MA, USA

**W Yang**
University of Connecticut, Storrs, CT, USA

**T-C J Yeh**
University of Arizona, Tucson, AZ, USA

**C M Yonker**
Colorado State University, Fort Collins, CO, USA

**E G Youngs**
Cranfield University, Silsoe, Bedfordshire, UK

**C W Zanner**
University of Nebraska–Lincoln, Lincoln, NE, USA

**J M Zhou**
The Institute of Soil Science, Academia Sinica, Nanjing, People's Republic of China

**T M Zobeck**
USDA Agricultural Research Service, Lubbock, TX, USA

# CONTENTS

## VOLUME 1

### A

### B

## C

## D

## E

## F

## VOLUME 2

## G

## H

## I

## J

## K

## L

## M

## VOLUME 3

### N

### O

### P

## Q

## R

## S

## VOLUME 4

T

U

V

W

**Z**

COLOR PLATE SECTIONS

# N

## NEMATODES

**D A Neher**, University of Toledo, Toledo, OH, USA
**T O Powers**, University of Nebraska, Lincoln, NE, USA

### Introduction – Life in a Soil Pore

Few people are aware of the most numerous of all soil-dwelling animals, the nematodes. These minute, unsegmented roundworms are actually aquatic organisms, inhabiting water films (1–5 $\mu$m thick) that coat and surround soil particles. (This chapter focuses on nematodes that typically complete their entire lifecycle within terrestrial soils or within plant roots. Numerous parasites of vertebrates have soil-dwelling stages, but their development is dependent on entering an appropriate host species. Vertebrate nematode parasites include hookworms, intestinal roundworms, pinworms, whipworms, and other serious pests of humans and domesticated animals. Information about vertebrate parasites is readily found in textbooks of parasitology.) Their size, typically 0.40–1.0 mm in length, is ideal for navigating this porous matrix. Although fully functional within a water film, they are ultimately bounded by soil structure, restricted in movement within soil pores less than 30 $\mu$m diameter.

Nematodes are simple organisms (**Figure 1**). Comprised of approximately 1000 somatic cells in the adult stage, these worm-like organisms are an example of functional and anatomical economy. Lacking eyes, appendages, and true segmentation, nematodes use mechanosensory and chemosensory neurons embedded in the cuticle to orient and respond to a wide range of environmental stimuli. At the nematode's anterior end is a circlet of sensilla arranged around an oral opening. These sensilla, including two larger, laterally placed chemosensory organs called 'amphids,' detect subtle chemical gradients, providing directional information to the nematode (Figure 2). Plant-feeding nematodes respond to slight variations in $CO_2$ and root exudates. A positive chemosensory signal initiates a snake-like sinuous movement characteristic of most nematodes, created by longitudinal body muscles working in apposition to a hydrostatic skeleton. Directional movement more than a few centimeters per day is unlikely in the soil environment. Once they contact the root surface, plant-feeding nematodes will probe with their stylet, a hollow, protrusible, hypodermic needle-like feeding tube (**Figure 3a**). Some plant-feeding nematodes will feed externally to the root surface, using this stylet to puncture cells, withdrawing the cytoplasmic contents. Others will penetrate the root and establish permanent feeding sites within the root cortex, or migrate cell-to-cell, leaving a trail of damaged necrotic tissue.

Free-living nematodes also occupy the interstitial spaces in soil. Unlike the plant-feeders, free-living nematodes are seldom sedentary, continually moving to feed on a diverse array of food, including bacteria, algae, fungi, protozoa, small invertebrates, and other nematodes. They use the same set of sensilla to track their food source; however, their feeding structures are modified to suit their meal. Bacterial-feeders graze using a relatively simple tubular mouthpart,

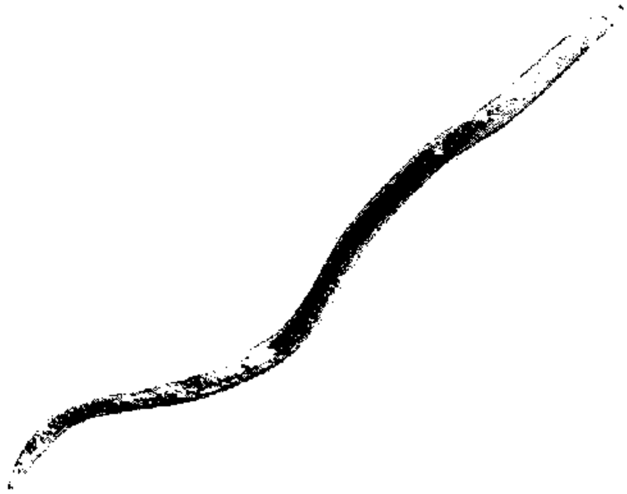


**Figure 1** (see color plate 43) Entire body view of a female *Pratylenchus agilis* (×100 magnification), collected on the Konza Prairie (96° W35′ 39° N05′) beneath Scribner's panicum (*Panicum scribnerianum*) and bluegrass (*Poa pratensis*) near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin (2000).

although the cuticle surrounding the oral opening may be modified elaborately to direct food toward the stoma (**Figure 4**). Predators may be adorned with 'teeth,' used to puncture or shred the integument of various invertebrates (**Figure 5**). Fungal-feeders and omnivores have a stylet similar in appearance to that of plant-feeders, distinguished by the lack of stylet 'knobs,' points of muscle attachment for stylet protrusion (**Figure 3b**).

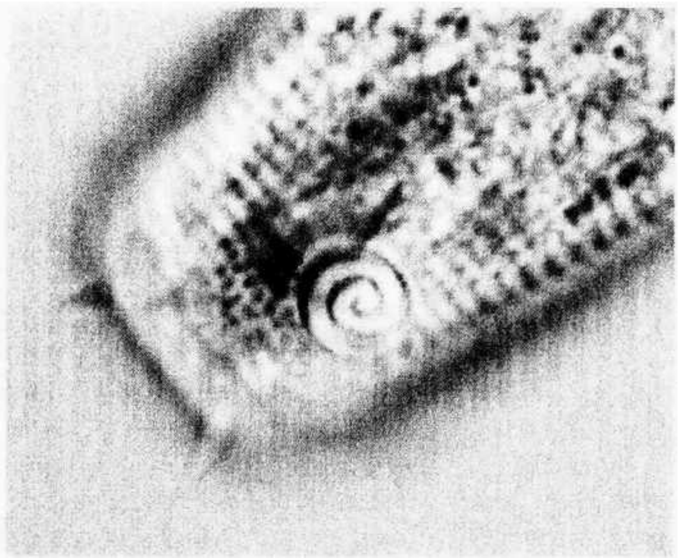


**Figure 2** Spiral-shaped chemosensory organs called 'amphids' in an anterior position of *Achromadora* sp. collected from soil of Jumbo Valley fen in Cherry County, Nebraska.

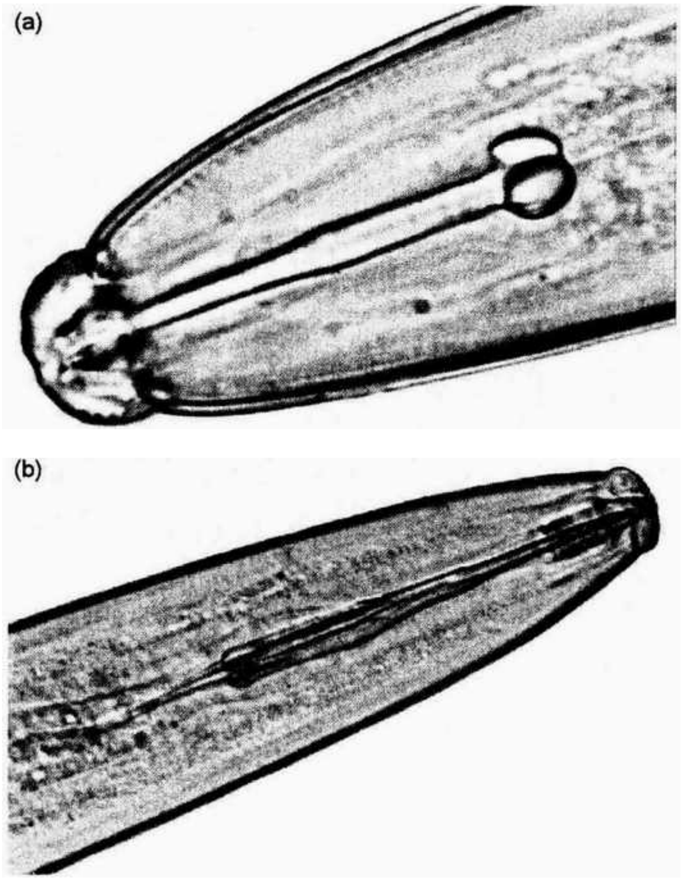


**Figure 3** Variation in morphology of spear-like structure in oral opening: (a) male plant-parasite *Hoplolaimus galeatus* (×1000 magnification), collected from soil with big bluestem (*Andropogon gerardii* Vitman) in the Konza Prairie (96° W35′ 39° N05′), near Manhattan, Kansas; and (b) female fungivore *Enchodelus hopedorus* (×400 magnification), collected from the summit of Long's Peak, Colorado (105° W35′ 40° N16′). Photographs are provided courtesy of Peter Mullin (2000).

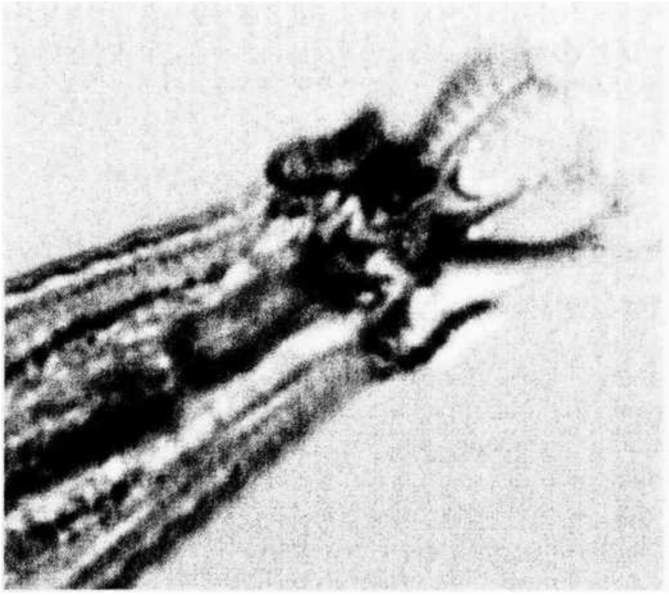


**Figure 4** Cuticle ornamentation of oral opening of *Acrobeles ctenocephalus* (×1000 magnification), collected in soil with little bluestem (*Andropogon scoparius*) in the Konza Prairie (96° W35′ 39° N05′) near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin (2000).

**Figure 5** Teeth of oral opening of predator *Mylonchulus montanus* (×1000 magnification), collected in soil with big bluestem in the Konza Prairie (96° W35′ 39° N05′) near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin (2000).

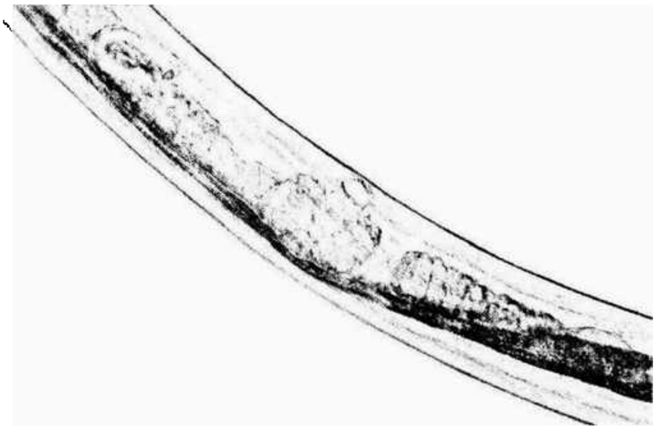


**Figure 6** Ovary in reproductive tract of female *Axonchium micans* (×400 magnification), collected in soil with big bluestem/Scribner's panicum in the Konza Prairie, near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin (2000).

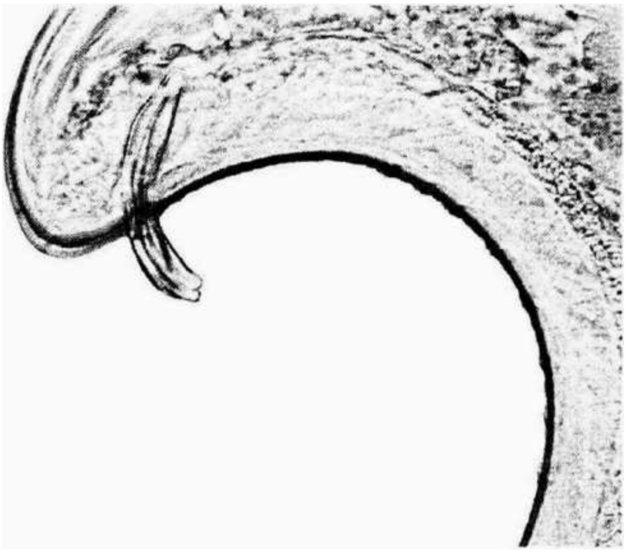


**Figure 7** Protrusible modification of male genitalia of *Aporcelaimellus obscuroides* (×400 magnification), collected in soil with big bluestem/Scribner's panicum in the Konza Prairie, near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin (2000).

Early observations of nematodes by light microscopy led scientists to refer to nematodes as a 'tube within a tube.' Food entering the nematode is channeled quickly into a tubular esophagus and then passed into an intestine that comprises the bulk of the body cavity. Waste products are eliminated through a posterior ventral anus. There is no circulatory or respiratory system within the nematode. In addition to the alimentary system, the reproductive system is a conspicuous feature in the adult nematode (**Figure 6**). Females and hermaphrodites are characterized by a branched genital system, which produces eggs that exit the body via a ventral genital opening in the body wall. Males possess a spicule, a cuticularized, protrusible modification of the genital system that guides sperm into the female for internal fertilization (**Figure 7**). Many nematodes are strictly bisexual, but parthenogenesis and self-fertilizing hermaphroditism are also common among species. Regardless of mode of reproduction, nematodes are prolific animals. Some species complete a life-cycle of 3 or 4 days. Others require several months to go from egg to egg-laying female. *Meloidogyne* spp., an economically important group of plant parasites, will typically begin to produce eggs within 3 weeks of hatching; each female is capable of producing hundreds of eggs during an adult life.

## Nematode Distribution and Abundance

The hallmark of Phylum Nemata is the exceptional abundance and ubiquitous presence of nematodes (**Table 1**). It is a challenging exercise to identify a habitat not occupied by a diverse community of nematodes. Consider the dry valleys of Antarctica, an environment considered so extreme that it is thought to provide the bare minimum for sustaining life on earth. There, in McMurdo Valley, three nematode species occupy the top levels in the soil food chain, two bacterivores and a third species, *Eudorylaimus antarcticus*, that feeds on the two bacterivorous species. Soils of the high Arctic are populated so heavily with nematodes that biologists have searched for explanations for an Arctic nematode biodiversity that exceeds that of the humid tropics. Temperate grasslands appear to sustain the greatest levels of diversity and abundance in the phylum. On Konza Prairie, one of the last remaining large stands of true tallgrass prairie in North America, each handful of soil is likely to contain 50–100 different nematode species. An estimated 1.5 million nematodes can be extracted from a square meter of soil, just within the surface 10 cm. Some researchers have considered the plant-feeding or herbivorous species as 'belowground buffalo,' referring to their impact as grazers of plant roots. Nearly all researchers agree that the approximately 25 000 described species of nematodes represent a small percentage of the species that actually exist in nature.

Finding agreement on those factors responsible for extraordinary nematode abundance and diversity is not so simple. Clearly, the evolutionary age of nematodes has provided considerable time for nematode diversification. Recent molecular studies indicate that the roots of the nematode evolutionary tree are hundreds of millions of years older than the corresponding tree for insects. Marine nematodes, which inhabit even the deepest ocean floors, have a nucleotide diversity that suggests nematodes flourished before life

**Table 1** Nematode communities in selected studies

| *Ecosystem type* | *Abundance ($10^6$ m$^{-2}$)* | *Genera (n)* | *Species (n)* | *Bacterivore (%)* | *Fungivore (%)* | *Herbivore (%)* | *Predator/omnivore (%)* |
|---|---|---|---|---|---|---|---|
| Grassland | 2.3–20 | 30–124 | 71–384 | 24–38 | 5–20 | 16–41 | 26–40 |
| Deciduous forest | 2.3–3.7 | 22–81 | 34–175 | 27–46 | 15–25 | 13–31 | 5–32 |
| Tropical rainforest | NA | 107 | 204 | 35 | 5 | 29 | 30 |
| Tundra | – | 58 | 162 | 55 | 3 | 22 | 19 |
| Heath | 1.2 | – |  | 57 | – | 2 | – |
| Agroecosystems | 3.5–5.0 | 19–44 | 33–100 | 36–46 | 13–19 | 33–41 | 5–7 |
| Temperate desert | – | 18 | 23 | 33 | 22 | 17 | 28 |
| Antarctic coastal | – | 15 | 27 | 47 | 27 | 0 | 27 |
| Antarctic dry valley | – | 3 | 3 | 67 | 0 | 0 | 33 |

NA, not available.

colonized land. The same nematode survival mechanisms that allowed them to be among the first colonizers of land most likely contribute to a competitive advantage in current-day harsh environments.

## Physiology

Although nematodes depend on free water for normal functioning, under harsh conditions such as freezing or drying many nematodes are capable of entering a cryptobiotic state, essentially a reversible state of suspended animation until favorable conditions return. Anhydrobiotic success, the ability to withstand the lack of moisture, depends on a slow rate of dehydration, rapid decline in cuticle permeability, and accumulation of compounds such as trehalose and glycerol to stabilize phospholipids and proteins. Species including *Acrobeloides* spp., *Aphelenchus avenae*, and *Scutellonema brachyurum* start to enter anhydrobiosis by coiling at gravimetric water contents of 3.7% (−300 kPa), 9% (−30 kPa), and 15% (−10 kPa), respectively. It is unknown how long nematodes may persist in this state and still survive hydration. Specimens of the wheat gall nematode *Anguina tritici* have been revived after more than 30 years in anhydrobiosis.

Nematodes are not alone among invertebrates in their ability to enter a cryptobiotic state. Therefore, additional clues to their successful colonization of soil are found in their physiological versatility. Physiological adaptations have allowed nematodes to invade habitats in which few other animals can survive and avoid interspecific competition and many environmental selection pressures. Nematodes are capable of regulating their uptake of oxygen over a wide range of partial pressures from 100 to 5%. When oxygen falls below 5%, nematodes convert their use of energy reserves. For example, *Aphelenchus avenae* uses reserves of neutral lipids under aerobic conditions and glycogen under anaerobic conditions. Not only does a hydrostatic skeleton facilitate movement, but also an ability to expand or contract body size in response to gradual changes in concentration of solutions containing sodium, calcium, magnesium, and/or potassium ions. The adjustment process, osmoregulation, is possible by modifying the permeability of their cuticle. Furthermore, nematodes can tolerate a wide pH range, with some species capable of withstanding strong acids or bases (pH 1.6–11.0). Nematode tolerance for temperature is equally remarkable. In hot springs, nematodes live at temperatures as high as or higher than any other Metazoa. *Aphelenchoides parietinus*, a cosmopolitan species, has been reported from 58°C springs in Chile and 61°C springs in New Zealand.

## Ecology

Although nematodes have adapted mechanisms to survive extremities of climate, their activity is stimulated by the return of more moderate conditions. For example, communities of nematodes are revived after rain in desert soils or after a relatively warm period in soils of polar regions. Species are found interacting with other organisms in a variety of occupations including competitor, opportunist, parasite, host, predator, or prey. As a competitor, nematodes rival protozoa for access to bacterial food sources. However, differences in size and generation time can alleviate competition by specializing on microhabitats fine-tuned to minute spatial and temporal scales. As opportunists, nematode species are phoretically transported by insects, adhering to their bodies, providing an opportunity to reach food sources at a longer distance than they would be capable of reaching alone. For example, Psychodidae flies carry nematodes to fermenting organic matter; dung beetles carry nematodes to fresh dung; and Scolytidae bark beetles transport Tylenchida and Rhabditida to their tunnels, where they feed on bacteria and fungi. Nematodes that are parasites of plants include both specialists and generalists.

Host ranges may extend to hundreds of plant species or be restricted to a single plant variety. Some nematodes specialize in precise feeding sites along the root. Plant-parasitic nematodes affect primary productivity of plants by altering uptake of water and nutrients. These abnormalities may result from changes in root morphology and/or physiology resulting in reduced productivity. Other nematodes infect insects, killing them within 48 h by releasing insect-pathogenic bacteria into the insect. The insect host dies from the bacterial infection, and nematodes feed on the bacteria and develop and reproduce inside the insect cadaver. In this case, the nematode and bacteria have a mutualistic association. Sometimes, a nematode finds itself in a role reversal, becoming the host or target of fungi specialized for trapping nematodes, using constricting rings, sticky knobs, or hyphal nets. Being in the middle of a food chain, nematodes eventually become prey to higher-order consumers. Nematodes provide a portion of the diet for many kinds of small soil arthropods. For example, a symphylan can hold seven large nematodes in its gut at one time. In 1 day, a mite may consume two large nematodes or several smaller nematodes.

Genetic diversity, nematode abundance, and the variety of niches occupied by nematodes has led to the assessment of nematode communities as a means of making predictions about past or present soil conditions. The species composition and relative nematode abundance of species comprising a nematode community allow nematodes to serve as biological indicators for soil disturbance as well as 'soil health.'

## Monitoring

Nematodes contribute directly to biogeochemistry of soils by regulating processes such as decomposition and nutrient cycling. Nematodes do not feed on decaying organic matter directly, but on bacteria, fungi, algae, and actinomycetes that colonize and decompose decaying plant and animal debris. Indirectly, nematodes control availability of nutrients by regulating the abundance or activity of these organisms, releasing nitrogen and phosphorus from microbes they digest, immobilizing nutrients in their live tissues, and excreting excess nitrogen as ammonium. Under field conditions, bacterial-feeding and predatory nematodes contribute 8–19% of nitrogen mineralization in conventional and integrated farming systems, respectively.

Nematode community structure and function are known to change in response to land-management practices such as nutrient enrichment through fertilization by organic or inorganic nitrogen, cultivation, liming, drainage, plant community composition and age, and toxic substances such as heavy metals, pesticides, and polycyclic aromatic hydrocarbons. Disturbance can affect the survival of individuals directly, or indirectly by changing resource levels. For example, a larger ratio of fungal-feeding to bacterial-feeding nematodes suggests less effect by cultivation than a smaller ratio. Any disturbance that results in compaction reduces soil porosity, and the numbers of relatively large-sized nematodes. General biocides such as methyl bromide can nearly eliminate nematodes, returning the ecological succession of soil communities to that assembling a depauperate soil matrix. Although recovery occurs eventually, abundance and diversity of nematode communities may take years to achieve prefumigation levels. Alternatively, some herbicides affect nematodes indirectly by reducing vegetation and smaller additions of organic matter to soil.

The composition of nematode communities may reflect the frequency of disturbance to soil, whether the disturbance is primarily physical and/or chemical in origin. This concept is based on the principle that different species have contrasting levels of sensitivity or tolerance to stress because of unique survival and/or reproductive traits. Smaller values of an index that integrates relative abundance of species within each sensitivity/tolerance class are indicative of a more disturbed environment, and larger values may indicate a less-disturbed environment. Ideally, indicators of soil health would correspond with ecological processes occurring in soil. Because initial experiments have been correlative in nature, it is too soon to claim such an association with great confidence.

The concept of nematodes as indicators of soil condition and their predictive use in soil management represents a tremendous shift in emphasis in the science of nematology. Formerly a science focused on the control of parasitic and harmful species, now nematodes are seen as an integral and potentially useful component of soil systems. In light of this new perspective, it is likely that the study of nematodes will contribute significantly to our understanding of biological processes in the soil.

## Further Reading

Fitter AH (1985) *Ecological Interactions in Soil*. Oxford, UK: Blackwell Press.

Freckman DW (1982) *Nematodes in Soil Ecosystems*. Austin, TX: University of Texas Press.

Freckman DW (1988) Bacterivorous nematodes and organic-matter decomposition. *Agriculture, Ecosystems and Environment* 24: 195–217.

Freckman DW and Baldwin JG (1990) Nematoda. In: Dindal DL (ed.) *Soil Biology Guide*, pp. 155–200. New York: John Wiley.

Ingham RE, Trofymow JA, Ingham ER, and Coleman DC (1985) Interaction of bacteria, fungi, and their nematode grazers: effect on nutrient cycling and plant growth. *Ecological Monographs* 55: 119–140.

Maggenti A (1981) *General Nematology*. New York: Springer-Verlag.

Nicholas WL (1984) *The Biology of Free-living Nematodes*. Oxford, UK: Clarendon Press.

Perry RN and Wright DJ (1998) *The Physiology and Biochemistry of Free-living and Plant-parasitic Nematodes*. New York: CABI Publishing.

Yeates GW (1999) Effects of plants on nematode community structure. *Annual Review of Phytopathology* 37: 127–149.

Zunke U and Perry RN (1997) Nematodes: harmful and beneficial organisms. In: Benckiser G (ed.) *Fauna in Soil Ecosystems: Recycling Processes, Nutrient Fluxes, and Agricultural Production*, pp. 85–133. New York: Marcel Dekker.

---

# NEUTRON SCATTERING

**M J Fayer and G W Gee**, Pacific Northwest National Laboratory, Richland, WA, USA

## Introduction

Soil water makes land-based life possible by satisfying plant water requirements (e.g., for crops and forests), by serving as a medium for nutrient movement to plant roots and nutrient cycling (e.g., fungal and microbial decay, mineralization, fertilization), and by controlling the fate and transport of contaminants in the soil environment. Therefore, understanding the dynamics of plant growth, nutrient cycling, and contaminant behavior in the soil requires knowledge of the soil-moisture status as well as its spatial and temporal variability.

The neutron probe (also known as the neutron moisture meter) is a popular instrument for measuring soil-water content. The probe emits high-energy neutrons and measures the low-energy neutrons that result from interactions with the soil constituents. What makes a neutron probe viable for measuring soil water is that hydrogen nuclei, which are part of the water molecule, play a dominant role in normal soils in reducing the energy level of neutrons.

The neutron probe measures soil-moisture status without the need to collect innumerable soil samples. This feature alone is one of the main reasons the technique is popular, because it allows researchers and farm managers to monitor the water content at multiple soil depths at the same location repeatedly without disturbing the soil. Thus, neutron probes make it possible to conduct long-term studies 'on the same plot of land' without fear of altering the plot by repetitive sampling. Neutron probes are easy to deploy for studies of spatial variability, thus overcoming the cost and destructiveness of sampling or the cost, disturbance, and inflexibility of weighing lysimeters. Another reason for the rise in popularity of the neutron probe is the large sensing volume, which nominally approximates a sphere with a 15- to 30-cm radius. The large sensing volume reduces the influence of small-scale variability in the soil, the number of measurements needed, and the sensitivity to soil disturbance caused by installation.

The idea of using neutrons to measure soil conditions arose in the 1940s, and the first use of a neutron device in a borehole was reported in 1947. By the 1950s, researchers around the world were extending neutron probe technology to agricultural conditions and reporting success in measuring water content in laboratory and field experiments. Since then, the neutron probe has become a standard field method for measuring soil-water content. As shown in **Figure 1**, the agriculture community uses neutron probes to monitor field water status in croplands and identify when irrigation is needed to meet plant water requirements. The environmental community uses neutron probes to monitor water conditions in and around waste disposal sites. The academic research community uses neutron probes to study and understand the hydrologic cycle.

## Theory and Instrumentation

The neutron is an elementary particle with a mass similar to that of a proton. Unlike the proton, the neutron has no charge, which allows it to interact primarily with atomic nuclei. A free neutron eventually decays according to:

$${}^{1}_{0}\mathrm{n} = {}^{1}_{1}\mathrm{H} + \beta \qquad [1]$$

**Figure 1** Neutron probe in use in a field containing soybean stubble. Photo courtesy of Steve Evett, USDA–ARS, Soil and Water Management Research Unit, Bushland, Texas.

with a half-life of 10.4 min and a $\beta$ energy of 0.782 MeV.

Neutrons are produced by the transformation from one radionuclide to another. Three modes of transformation suitable for soil and environmental studies are spontaneous fission, alpha-neutron ($\alpha$-n) interactions, and neutron generation. Of these, the $\alpha$-n mode has proven to be most effective for agricultural soils.

The $\alpha$-n method for generating neutrons involves mixing two materials: one material to provide a source of alpha particles, the other material to provide a source of neutrons when exposed to alpha particles. Potential sources of alpha particles include americium, californium, polonium, plutonium, and radium. The americium decay reaction is:

$$ {}^{241}_{95}\mathrm{Am} - {}^{237}_{93}\mathrm{Np} + {}^{4}_{2}\alpha + \gamma \qquad [2] $$

which generates an $\alpha$ with an energy of 5.48 MeV. The material typically chosen to provide the neutrons is beryllium, although lithium has also been used. For the americium–beryllium (Am-Be) combination, the equation is:

$$ {}^{4}_{2}\alpha + {}^{9}_{4}\mathrm{Be} = {}^{12}_{6}\mathrm{C} + {}^{1}_{0}\mathrm{n} + \gamma \qquad [3] $$

The optimum material combination from the standpoint of commercial application is americium–beryllium, which has the desirable characteristics of a long half-life (458 years), a moderately high neutron-emission rate of 67.6 n $s^{-1}$ $MBq^{-1}$ (2500 n $s^{-1}$ $mCi^{-1}$), and a low gamma dose rate of 2.78 nSv $s^{-1}$ at 1 m (1 mrem $h^{-1}$ at 1 m) under a fast neutron flux of $10^6$ n $s^{-1}$. The long half-life ensures that a sealed source will last many years with negligible decay. The high emission rate ensures that enough neutrons will be present to make detection feasible. The low gamma dose rates and moderate fast neutron fluxes make it relatively safe to use.

The energy of neutrons generated by the Am-Be source ranges from 1 to 14 MeV and averages approximately 4.5 MeV. Neutrons at these energy levels are classified as fast neutrons. As they permeate the surrounding medium, fast neutrons interact with nuclei in scattering (both elastic and inelastic) and absorption events. The probability that a nucleus will interact with a free neutron can be statistically determined and represented using the microscopic cross-section ($\sigma$); the standard unit of measurement is the barn, which is $10^{-28}$ $m^2$ per nucleus. The ability of a soil to interact with neutrons will be a function of the soil's elemental composition, the $\sigma$ values associated with each element, and the element density (atoms per volume).

In elastic scattering events, the neutron strikes the nucleus and rebounds; energy and momentum are conserved. This type of scattering is colloquially referred to as 'billiard ball scattering' and is the dominant process for slowing down neutrons under normal soil conditions. In inelastic scattering events, the neutron strikes the nucleus and rebounds with a loss of energy via gamma-ray emission. In absorption events, neutrons strike and react with the nucleus, resulting in the release of additional radiation (e.g., $\alpha$, $\beta$, $\gamma$, n). Both inelastic scattering and absorption of fast neutrons are minor neutron-slowing processes relative to elastic scattering.

**Table 1** shows that soil constituents differ in their efficiency at slowing down neutrons by elastic scattering. Hydrogen, which is present primarily in soil water, requires the least number of collisions to thermalize a 14 MeV neutron. **Table 1** shows that, although the slowing-down power of the elements is a function of the energy range, hydrogen consistently has the greatest potential.

After repeated scattering events, the energy level of the neutron decreases to less than 1000 eV, at which point the neutron is classified as a slow or thermalized neutron. The petroleum industry uses slow neutrons in the energy range from 1000 to nearly 0.025 eV for geophysical measurements (e.g., petroleum reservoir

**Table 1** Fast and thermal neutron information for the most common soil constituents

| *Element* | *Nominal soil fraction (wt %)* | *Estimated no. collisions to thermalize neutron from 14 MeV to 0.01 eV* | *Estimated fast neutron slowing-down cross section 10–0.183 MeV (barn)* | *Estimated fast neutron slowing-down cross section 0.183 MeV to 1.44 eV (barn)* | *Thermal capture cross section at 0.0253 eV (barn)* |
|---|---|---|---|---|---|
| H | – | 19 | 2.76 | 19.2 | 0.332 |
| O | 49 | 154 | 0.079 | 0.48 | 0.0002 |
| Si | 33 | 297 | 0.164 | 0.16 | 0.16 |
| Al | 7 | 290 | 0.668 | 0.13 | 0.23 |
| Fe | 4 | 539 | 0.132 | 0.32 | 2.62 |
| C | 2 | 112 | 0.234 | 0.77 | 0.00373 |
| Ca | 1 | 380 | 0.172 | 0.12 | 0.44 |
| K | 1 | 418 | 0.151 | 0.11 | 2.07 |
| Na | 0.7 | 190 | 0.204 | 0.28 | 0.505 |
| Mg | 0.6 | 235 | 0.153 | 0.27 | 0.069 |
| Ti | 0.5 | 512 | – | – | 6.09 |
| N | 0.1 | 155 | – | – | 1.88 |
| P | 0.08 | 333 | – | – | 0.166 |
| Mn | 0.08 | 586 | – | – | 13.4 |
| S | 0.05 | 344 | – | – | 0.51 |
| Cl | 0.01 | 343 | – | – | 33.8 |
| B | 0.001 | 89 | – | – | 759 |

Adapted from: IAEA (1970) *Neutron Moisture Gauges*. Technical Reports Series 12. Vienna, Austria: International Atomic Energy Agency; Bohn HL, McNeal BL, and O'Connor GA (1979) *Soil Chemistry*. Chichester: John Wiley & Sons, Inc.; Hearst JR, Nelson PH, and Paillet FL (2000) *Well Logging for Physical Properties*, 2nd edn. New York: John Wiley & Sons, Inc.; Chang J *Table of Nuclides*, KAERI (Korea Atomic Energy Research Institute).

characterization); these slow neutrons are called epithermal neutrons. Slow neutrons near or below the energy level of 0.025 eV are called thermal neutrons because they are in an energy condition similar to thermal equilibrium with their environment.

The density of thermal neutrons that forms around the probe is balanced between thermal neutron creation (via thermalization of fast neutrons) and the processes of thermal neutron capture, decay, and diffusion away from the probe. Once formed, thermal neutrons do not travel far; the diffusion length is highly variable but is on the order of centimeters. If a fast neutron encounters many nuclei close to the source, the neutron will be reduced to thermal conditions while it is still relatively close to the source. If a fast neutron encounters few nuclei close to the source, the neutron will tend to migrate further from the source before it is reduced to a thermal energy level. In other words, the nature of the medium dictates how far from the source the thermal neutrons are created. Because hydrogen (and thus water) has such a high macroscopic cross-section relative to all other normal soil constituents, water content is the primary controller of the density of thermal neutrons near the probe and thus determines how much of the medium is sensed. This phenomenon is called the volume of influence. Figure 2 shows how the volume of soil sensed by the neutron probe decreases as the water content increases. The radius of influence is typically 15 cm for wet soil and as much as 30 cm for dry soil.

Thermal neutrons can be detected using the ionization products that form when the neutron is captured by an element. Early detectors used gas-filled tubes containing boron trifluoride ($BF_3$), but modern detectors tend to use the helium isotope $^3He$, which reacts as follows:

$$^3_2\mathrm{He} + ^1_0\mathrm{n} = ^3_1\mathrm{H} + \mathrm{p} \qquad [4]$$

The $^3He$ isotope has a higher thermal cross-section (5400 barns) than $^{10}B$ (3840 barns) and can be stored in the detector tube to a greater density. For both the B and He detectors, the ionization products within the gas in the tube alter the current through the tube, thus allowing the thermal neutrons to be detected and counted. These detectors are insensitive to fast neutrons.

Radioactive decay is a random process that affects the rate of fast neutrons emanating from the probe. To overcome the randomness of the source, neutron probes are designed to count the number of thermal neutrons for a specific duration. Repeated measurements under the same conditions, such as in a standard medium, produce a sample population of counts that resembles a Poisson distribution, which has the property that the standard deviation of the finite sample population can be approximated by the square root of

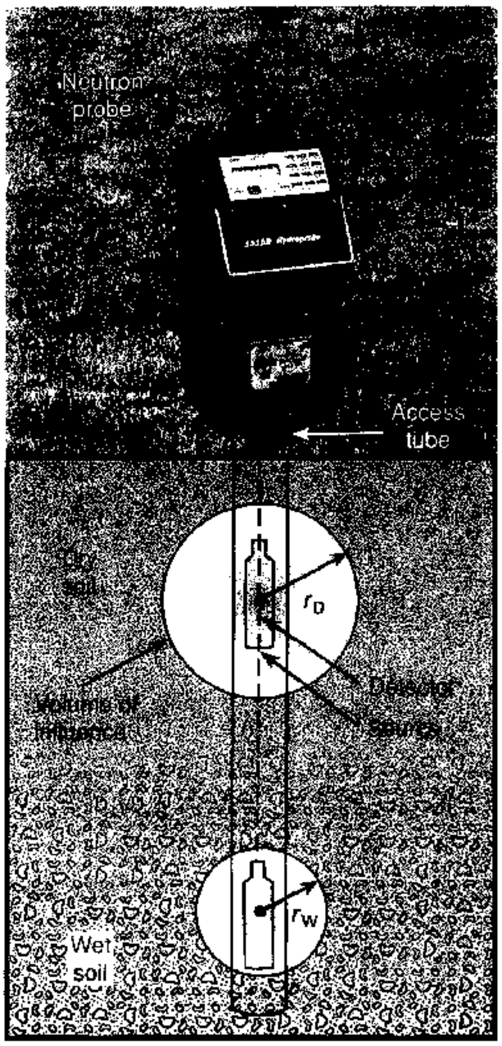


**Figure 2** A neutron probe volume of influence in wet soil and dry soil.

the mean count. Because a Poisson distribution can be approximated with a normal distribution (given sufficient counts), the mean count and standard deviation can be used to describe the sample distribution. Thus, they can be used to control measurement precision by estimating the counting time needed to achieve a specific precision. They can also be used for quality control by comparing the results with a normal distribution to detect abnormal amounts of fluctuation that would indicate a malfunctioning probe.

Neutrons emitted by a neutron probe have the potential to activate (make radioactive) some of the elements (e.g., Na, Mg, Cl, Si, Al, Fe, Mn) in the soil and access tube. For example, the aluminum in an aluminum access tube can become activated as follows:

$$^{27}_{13}\mathrm{Al} + ^{1}_{0}\mathrm{n} = ^{28}_{13}\mathrm{Al} \qquad [5]$$

The resulting aluminum isotope has a short half-life (2.3 min) and yields a beta particle during decay. Other activation product half-lives range from seconds to hours. The probability of significant activation is very low because the concentrations in the soil are generally so small and the neutron flux is relatively low (more than an order of magnitude lower than required to cause measurable activity). Because activation products are belowground and of short duration, the risk of exposure is extremely small.

## Field Methodology

To use a neutron probe to measure soil moisture requires an access hole in the soil into which the probe can be inserted. Ideally, the hole should be just wide enough to permit entry of the probe while minimizing excess air gaps. In reality, drilling a uniform-width hole and keeping it open and unchanging is highly unlikely, even under the best of field conditions. Therefore, the commonly accepted practice is to install a rigid-wall tube (or pipe, conduit, casing) that allows the probe to access the soil repeatedly without degrading the hole geometry.

The access-tube material should be chosen based on ease of installation, durability, availability, and minimal impact on neutron flux. Thin-walled tubing made of aluminum, schedule-10 galvanized steel, or schedule-40 polyvinyl chloride (PVC) is the preferred choice, but thicker iron well casing has been shown to be acceptable for specific applications. Successful installation requires that air gaps and packing material (e.g., grout) be minimized or, even better, eliminated and that the installation procedure be used consistently for all tubes. The key to choosing the correct access-tube material and installation method is to demonstrate, through calibration, that the probe and access-tube combination retains sufficient sensitivity to be an effective soil-moisture measurement system.

The most common orientation for access tubes is vertical so that the operator can monitor the depth distribution of soil water. However, access tubes can also be installed at an angle, or even horizontally, to allow the operator to monitor beneath an object such as a buried storage tank. Access tubes can become damaged during installation; once installed, the tube should be tested with a dummy probe to determine its integrity. Dummy probes are inert objects, usually made of metal, that are shaped like a real probe but slightly larger. If the dummy probe can be passed through the access tube, the operator can be assured that the neutron probe will not become lodged.

In addition to protecting the access tube, the operator must protect the access-tube site from compaction by foot and vehicular traffic. Relative to the tube surroundings, compaction increases the density of neutron-affecting minerals, alters the soil hydraulic properties, and affects plant growth and behavior. The seriousness of compaction can increase if the site is used for multiyear tests involving many site visits.

## Calibration

A neutron probe detects and measures thermal neutrons to indicate soil-moisture status. In addition to soil water, the formation and maintenance of a cloud of thermal neutrons within the sensing volume of the probe depends on the mineral composition of the soil matrix, bulk density, bound water, soil-water constituents, and organic matter. Table 1 shows how the major soil elements vary in their ability to slow down fast neutrons as well as capture thermal neutrons. Clearly, hydrogen has the greatest potential to thermalize neutrons and capture them, and thus control the thermal neutron density around the probe. As noted above, increases in bulk density translate directly into increases in the density of neutron-affecting minerals. Bound (or structural) water is part of the soil matrix (e.g., between clay lamella). Bound water affects neutron thermalization just like free water, but it is not available to plants and does not interact with free-flowing soil water.

Natural soil-water constituents are generally present in concentrations too low to affect a neutron probe and are normally not considered during probe calibration and use. However, operators should be cognizant when soils are subject to significant fertilization, salts (e.g., carbonate and sulfate deposits in semiarid and arid soils), sludge amendments, or waste disposal. Under these conditions, the soil water could contain constituents with significant thermal neutron-capture cross sections (e.g., B, Cl, Fe, Mn) that would render the probe calibration less representative unless modified.

Another soil component normally not considered during probe calibration and use is organic material, which includes living plant roots and associated subsurface plant components, dead and decaying plant parts, and fully developed organic matter. The key feature of organic material that could affect probe sensitivity is the presence of water, either as part of living plant tissue or in response to changes in the water-holding capacity of the soil. These factors should be assessed seasonally and in special zones in the soil where plant root changes are greatest (e.g., soil surface and soil-layer interfaces with contrasting soil texture).

Because they are affected by conditions such as mineralogy, bulk density, and sediment discontinuities, neutron probes should be calibrated for the specific conditions under which they will be used in the field. In a typical field calibration, multiple access tubes are emplaced to allow access to the entire thickness of soil. The neutron probe is lowered to different depths in each access tube and readings performed. Immediately subsequent to collecting neutron readings, the soil is excavated to collect multiple undisturbed soil samples directly adjacent to the tubes and at the same probe-measurement depths. The undisturbed samples are then analyzed in the laboratory to determine the measured volumetric water contents and bulk densities for those depths. These values are then related to the neutron-probe counts taken at the respective depths. The key to a successful calibration is to collect data under wet and dry conditions in each soil layer to cover the complete range of expected water contents.

The thermal neutron count rate is converted to soil moisture using a calibration equation. Separate equations may be required for different soil horizons. The most common calibration equation is a linear function of the form:

$$\theta = m(\mathrm{CR}) + b \qquad [6]$$

where $\theta$ is the volumetric moisture content, CR is the thermal neutron count ratio, and $m$ and $b$ are calibration parameters. The count ratio is the thermal-neutron count in the soil relative to the thermal-neutron count in a standard medium. The rationale for referencing to a standard medium is to minimize the impacts caused by changes in the probe source and electronics due to aging and decay that might not be readily apparent to the probe operator. The standard medium must have neutron-slowing properties and the properties must remain constant. Typical standards are the probe shield or a sealed barrel containing nonchanging material such as water or wax. If the probe shield is used, it ought to be located 0.8 m from the ground surface and up to 3 m from structures, vehicles, and people to minimize interferences. The distance recommendations are less stringent if the larger standards such as water- or wax-filled barrels are used. Figure 3 shows a typical linear calibration equation and the supporting field measurements.

Equation [6] has been shown to be appropriate for a wide variety of soils and moisture conditions, but other forms of calibration equations have been proposed and used for special cases. These other equations include multiple regression equations that use other soil information such as bulk density and quadratic equations that allow for a better fit of data for specific cases (e.g., very low water contents; large-diameter, iron-cased wells).

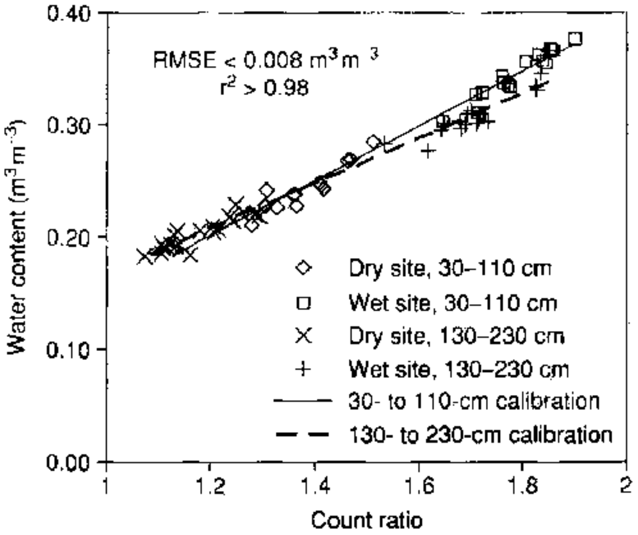


**Figure 3** Calibration of a neutron probe in Pullman soil at Bushland, Texas. RMSE, root mean square error. Courtesy of Dr. Steve Evett, USDA–ARS, Soil and Water Management Research Unit, Bushland, Texas.

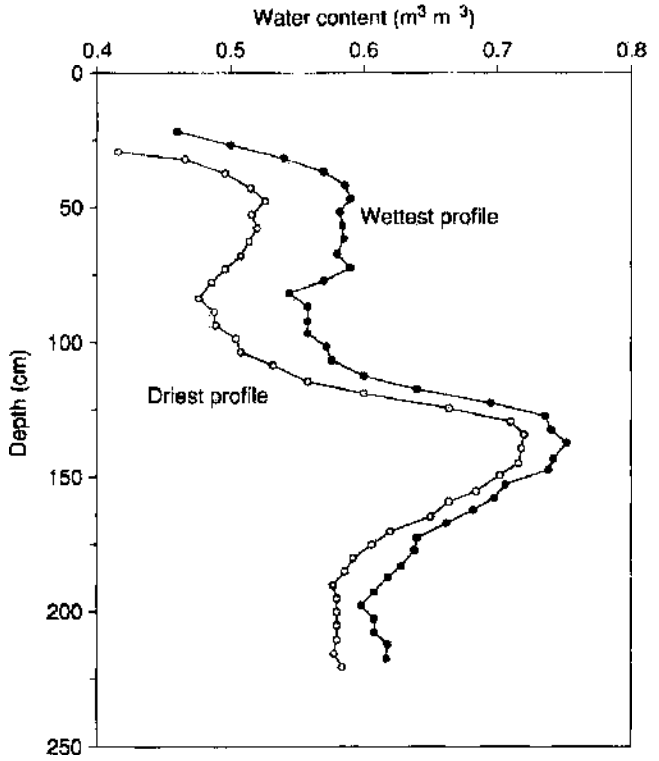


**Figure 4** Water-content profiles in a Tasmanian soil on two different dates. Adapted from Greacen EL (1981) *Soil Water Assessment by the Neutron Method.* Victoria, Australia: CSIRO.

## Applications

The output of the neutron probe is converted to soil moisture data. These data are used to assess the depth distribution of water, the overall soil-water status, and the movement of water in the soil profile. When collected on multiple occasions, the temporal data can reveal long-term changes such as withdrawal of water by plants, loss of water by evaporation, and responses to weather and irrigation conditions. Figure 4 shows a typical variation in water-content profile determined with a neutron probe.

## Safety and Care

Human health is affected by exposure to radiation. The nature and degree of health effects depend on the type of radiation and the length of time one spends in the radiation field. Although the radioactive source in a neutron probe is sealed and unreachable by the operator (barring the use of force to overcome the seal), the probe operator is subject to the radiation emitted through the seal and the probe housing. Emissions from neutron probes consist of neutrons and gamma rays at various energy levels. Measurements adjacent to the probe shield have shown that dose rates are dominated by the fast neutron flux. Thus, most probe monitoring efforts focus on measuring the fast neutron flux. The impact to an operator can be seen in the following example. The radiation field at a distance of 5 cm from the shield of a 3.7-GBq (100 mCi) Am-Be source was measured and found to be 0.12 mSv h$^{-1}$. At that dose rate, a neutron probe operator could be 5 cm from the shield for 8 h every week of the year and not exceed the International Standard worker limit of 50 mSv in any single year.

Radiation from a neutron probe is generally not a health hazard to someone within the radiation field if ALARA (as low as reasonably achievable) principles are followed. For example, the radiation field strength can be reduced by limiting the size of the source and surrounding it with a shield. Typical shields are composed of hydrogenous material to reduce neutrons. The dose from a 3.7-GBq (100 mCi) Am-Be source is reduced by approximately 20% when behind a 7-cm-thick plastic shield; the resulting dose is primarily from fast neutrons (0.1 mSv h$^{-1}$ (10 mrem h$^{-1}$) versus 0.01 mSv h$^{-1}$ (1 mrem h$^{-1}$) from gammas and thermal neutrons). The dose from more modern probes is less because the source activities have been reduced, e.g., 0.37 GBq (10 mCi) and 1.8 GBq (50 mCi). Operators can reduce exposure further by using distance and time to advantage. For example, the radiation filed is inversely proportional to the distance squared. By doubling the distance to the probe, the operator can reduce the radiation field by a factor of four. The operator can also reduce the radiation field to some extent by using intervening material (e.g., shielding between a truck cab and the truck bed when the probe is transported). In addition to increasing distance to the probe, the operator can also reduce exposure by minimizing the time spent within the radiation field. Prior planning, quality-assurance procedures, and efficient work habits can reduce the amount of unnecessary time spent lingering near the probe.

Modern neutron probes are robust but, because of the radioactive source and the electronics, these probes require administrative oversight and care in handling. The International Atomic Energy Agency (IAEA) promulgates safety and regulatory requirements designed to ensure safe use of neutron probes. More than 125 nations belong to the IAEA and they use national organizations to address the IAEA requirements. These requirements include the need for routine periodic leak tests of sealed sources to look for particles of the source (thus indicating a failure of the source seal) and transportation and storage controls for safety, theft protection, and disposal.

Manufacturers provide guidance to operators for probe care, cleaning, and battery replacement. For maintenance and repair work, manufacturer recommendations differ. Some require sending the probe back to the factory; others allow the electronics to be disengaged from the source and sent to the factory while the source remains stored locally. A probe operator can prolong the life of the probe, cable, and electronics by taking care to avoid dropping them, subjecting them to extremes of temperature, pressure, and acceleration, and by minimizing their exposure to solar radiation and water. The operator should also prevent undue stress on the cable.

Access tubes can become damaged during use. Caps or covers on the access tubes help to limit accidental bumps as well as to keep out water, soil, and organic debris. Even so, the tubes should be periodically checked for dents, decay, and debris. Tube cleanliness can be maintained by using a cleaning rag on a pole. The radial uniformity of the tube should be checked occasionally using a dummy probe to ensure that the actual neutron probe will pass easily through the tube.

## Future Use

After more than 50 years, neutron probes remain the most reliable tool available for field monitoring of soil-water content. Neutron probes provide integrated measurements over relatively large volumes of soil and, with proper access, allow for repeated sampling of the subsurface at the same locations. The limitations of neutron probes include costly and time-consuming manual operation, lack of data automation, and costly regulatory requirements. As more nonradioactive systems for soil-water monitoring are developed to provide automated profiling capabilities, neutron-probe usage will probably decline. Until then, neutron probes will continue to be a standard for reliable measurements of field water contents in soils around the globe.

*See also:* **Time-Domain Reflectometry; Water Content and Potential, Measurement; Water Cycle; Water Potential**

## Further Reading

American Society for Testing Materials (2002) *Standard Test Method for Water Content of Soil and Rock In-Place by the Neutron Depth Probe Method,* D 5220-92 (97), vol. 04.08, pp. 1089–1094. West Conshohocken, PA: ASTM International.

Bohn HL, McNeal BL, and O'Connor GA (1979) *Soil Chemistry.* New York: John Wiley & Sons, Inc.

Evett SR (2002) Exploits and endeavours in soil water management and conservation using nuclear techniques. In: Eaglesham A (ed.) *International Symposium on Nuclear Techniques in Integrated Plant Nutrient, Water and Soil Management*, pp. 151–177. Vienna, Austria: International Atomic Energy Agency.

Gardner W and Kirkham D (1952) Determination of soil moisture by neutron scattering. *Soil Science* 73: 391–401.

Gee GW, Stiver JF, and Borchert HR (1976) Radiation hazard from americium-beryllium neutron moisture probes. *Soil Science Society of America Journal* 40: 492–494.

Greacen EL (1981) *Soil Water Assessment by the Neutron Method.* Victoria, Australia: CSIRO.

Hearst JR, Nelson PH, and Paillet FL (2000) *Well Logging for Physical Properties*, 2nd edn. New York: John Wiley & Sons, Inc.

Hignett C and Evett SR (2002) Neutron thermalization. In: Dane JH and Topp GC (eds) *Method of Soils Analysis, part 4, Physical Methods*. Madison, WI: Soil Science Society of America.

IAEA (1970) *Neutron Moisture Gauges*. Technical Reports Series 112. Vienna, Austria: International Atomic Energy Agency.

IAEA (1999) *Nuclear Geophysics and Its Applications*. Technical Reports Series 393. Vienna, Austria: International Atomic Energy Agency.

IAEA (2002) *Neutron and Gamma Probes: Their Use in Agronomy*. Training Course Series No. 16. Vienna, Austria: International Atomic Energy Agency.

Tyler SW (1988) Neutron moisture meter calibration in large diameter boreholes. *Soil Science Society of America Journal* 52: 890–893.

# NITROGEN IN SOILS

Contents

## Cycle

**M S Coyne and W W Frye**, University of Kentucky, Lexington, KY, USA

### Introduction

Nitrogen (N) is the element most often limiting plant nutrition in terrestrial ecosystems. The greatest source of available N is atmospheric dinitrogen ($N_2$), which is relatively inert and can only be used by symbiotic and free-living prokaryotic bacteria with the capacity for $N_2$ fixation (**Table 1**). For other plants and soil organisms, the slow release of N from rocks and minerals and cycling between organic and inorganic forms in soil is crucial to life. Unfortunately, some aspects of N cycling can be problematic. Nitrogen can be readily lost from terrestrial soils, leading to reduced fertility and surface- or groundwater contamination, and several transformations give rise to intermediate or final products that can have negative environmental consequences.

### Properties and Forms of Nitrogen in Soil

There are two overriding chemical properties of the N atom worth noting. First, the dinitrogen molecule, $N_2$, is very stable. Once formed and lost from the soil environment, $N_2$ is relatively inert and requires considerable energy ($946\,\mathrm{kJ\,mol^{-1}}$) to break the triple bond between N molecules and reintroduce N to biological systems. The second significant feature of N, owing to its atomic structure and electron configuration, is its ability to undergo changes in oxidation state from $-5$ to $-3$ depending on its environment. This property is important for soil transformations, because N, in its oxidized forms, can be an electron acceptor in anaerobic respiratory metabolism. N, in its reduced forms, can also be an energy source for various chemolithotrophic bacteria. The N cycle in soil is driven by these transformations between oxidized and reduced states and between organic and inorganic forms (**Figure 1**).

The principal soluble inorganic N forms in soil are $NO_3^-$, $NO_2^-$, and $NH_4^+$. The inorganic gaseous N forms in soil are $N_2$, $N_2O$, NO, and $NH_3$. Organic N takes many forms, such as amino acids (e.g., glycine, glutamine), amino sugars (e.g., glucosamine, galactosamine), nucleosides (e.g., adenine, guanine), peptides, phospholipids (e.g., phosphatidylethanolamine, phosphatidylserine), vitamins (e.g., niacin), and other compounds such as creatine, cyanide, allantoin, various alkyl amines, and urea. Unidentifiable forms of N are typically polymerized with soil organic matter fractions such as humic and fulvic acids.

### N-Cycling Pathways

The major organic and inorganic N transformations that occur in soil can be grouped into five pathways: mineralization (ammonification), assimilation (immobilization), nitrification, nitrate reduction, and $N_2$ fixation.

**Table 1** Global sources of available N

| *Source* | *N content ($\times 10^{16}$ kg)*[a] | *Total (%)* |
|---|---|---|
| *Atmosphere* | 386.000 | 87.0000 |
| *Oceans and lakes* | | |
| Dissolved $N_2$ | 2.190 | 0.5000 |
| Soluble inorganic N | 0.110 | 0.0200 |
| *Sediments* | 35.000–55.000 | 12.0000 |
| *Sea bottom organic N* | 0.054 | 0.0100 |
| *Terrestrial biosphere* | 0.028–0.065 | 0.0100 |
| *Terrestrial soils* | | |
| Organic N | 0.022 | 0.0050 |
| Clay-fixed $NH_4^+$ | 0.002 | 0.0005 |

[a] The data exclude N in the lithosphere such as igneous rocks, sediments, and coal that do not actively participate in global N cycling.

Source: Stevenson FJ and Cole MA (1999) *Cycles of Soil*, 2nd edn. New York: John Wiley and Sons, Inc.

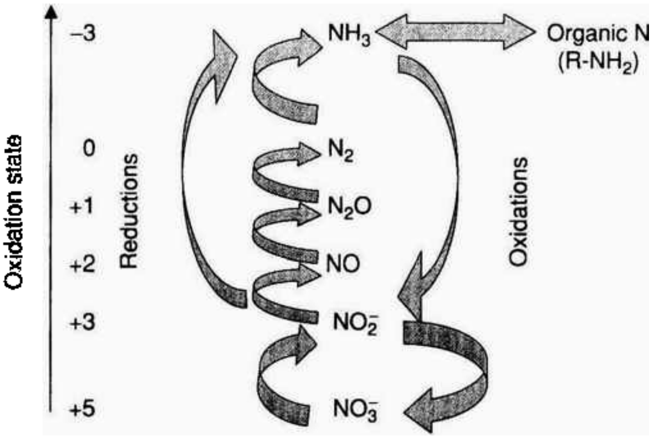


**Figure 1** Oxidation–reduction transformations of nitrogen.

### Mineralization (Ammonification)

During mineralization, organic N is hydrolyzed to release $NH_3$, which is subsequently protonated in the soil solution to become the cation $NH_4^+$ (eqn [1a]). Consequently, mineralization is frequently referred to as ammonification:

$$\text{Organic N (R-NH}_2\text{)} \longrightarrow NH_3 \longrightarrow NH_4^+ \quad [1a]$$

$$\text{Glutamate} + NAD^+ + H_2O \xrightleftharpoons{\text{Glutamate dehydrogenase}} \alpha\text{-Ketoglutarate} + NH_4^+ + NADH + H^+ \quad [1b]$$

Another example of mineralization that has great importance in soil is urea hydrolysis by the seemingly ubiquitous enzyme urease, which is shown in eqn [2]. Urea is one of the dominant organic N forms excreted by animals, and it is also one of the most popular forms of solid fertilizer used on a global basis. Urea hydrolysis in soil is rapid, and the subsequent $NH_3$ formation can cause considerable N loss through volatilization. One of the immediate effects of urea hydrolysis is also a rapid increase in the soil pH:

$$H_2O + NH_2CONH_2 \xrightarrow{\text{Urease}} CO_2 + 2NH_3 \quad [2]$$

### Assimilation (Immobilization)

Assimilation (immobilization) is the incorporation of inorganic N as $NH_4^+$, $NO_3^-$, or $NO_2^-$ into biomass (eqns [3] and [4]). Immobilization can also refer to the binding of $NH_4^+$ to certain clays in soil and to the interaction of inorganic N with soil organic matter:

$$NH_4^+ \longrightarrow \text{Organic N (R-NH}_2\text{)} \quad [3]$$

$$NO_3^- \text{ or } NO_2^- \longrightarrow NH_4^+ \longrightarrow \text{Organic N (R-NH}_2\text{)} \quad [4]$$

### Nitrification

Nitrification is the general term that describes oxidation of reduced inorganic and organic N and formation of $NO_2^-$ or $NO_3^-$. It is an aerobic process that occurs by one of two pathways: autotrophic nitrification, carried out by chemolithotrophic bacteria, and heterotrophic nitrification, carried out by chemoheterotrophic bacteria and fungi.

**Autotrophic nitrification** Autotrophic nitrification is a two-step process that is carried out by chemolithotrophic bacteria, which obtain energy by oxidizing $NH_3$ to $NO_3^-$ (eqn [5]).

$$NH_4^+ \xrightleftharpoons[+H^+]{-H^+} NH_3 \longrightarrow NO_2^- \longrightarrow NO_3^- \quad [5]$$

The first step in the pathway, which is performed by $NH_3$-oxidizing bacteria such as *Nitrosomonas*, involves the sequential oxidation of $NH_3$ to hydroxylamine ($NH_2OH$) and then to $NO_2^-$ (eqns [6a] and [6b]):

$$NH_3 + O_2 + 2H^+ \xrightarrow{\text{Ammonia monooxygenase (AMO)}} NH_2OH + H_2O \quad [6a]$$

$$NH_2OH + H_2O \xrightarrow{\text{Hydroxylamine oxidoreductase (HAO)}} NO_2^- + 5H^+ \quad [6b]$$

The second step in the pathway is performed by $NO_2^-$ oxidizing bacteria typified by bacteria in the genus *Nitrobacter*. Nitrite oxidation is generally so rapid that $NO_2^-$ rarely accumulates to any significant extent in soil:

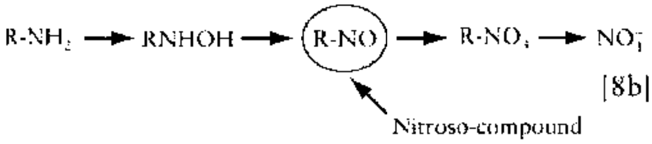
$$H^+ + NO_2^- + H_2O \xrightarrow{\text{Nitrite dehydrogenase}} NO_3^- + 3H^+ \quad [7]$$

Autotrophic nitrification is believed to be the dominant type of nitrification occurring in agricultural soils. Nitrification is an acidifying process and use of reduced organic and inorganic N for plant nutrition will lower the soil pH with time because of nitrification.

**Heterotrophic nitrification** Heterotrophic nitrification is the oxidation of organic N and $NH_3$ by various chemoheterotrophic bacteria and fungi such as *Arthrobacter*, *Streptomyces*, and *Aspergillus*. Heterotrophic nitrification produces $NO_2^-$ and $NO_3^-$ in soils typically too acid to permit autotrophic nitrification (eqn [8a]).

$$\text{Organic N (R-NH}_2) \longrightarrow NO_2^- \text{ or } NO_3^- \quad [8a]$$

Products of heterotrophic nitrification also include nitroso-compounds, which have been identified as potential carcinogens (eqn [8b]):



$$\text{R-NH}_2 \longrightarrow \text{RNHOH} \longrightarrow \underset{\text{Nitroso-compound}}{\text{(R-NO)}} \longrightarrow \text{R-NO}_2 \longrightarrow NO_3^- \quad [8b]$$

## Nitrate Reduction

In anaerobic or waterlogged environments, denitrification and dissimilatory nitrate reduction to ammonium (DNRA) are the two major pathways by which the $NO_2$ and $NO_3^-$ formed by nitrification can be subsequently reduced. Facultatively anaerobic bacteria perform these N transformations, although it has been suggested that some fungi are also capable of denitrification.

**Denitrification** Denitrification is the major reductive fate of $NO_3$ in most soil environments. It is a respiratory process in which nitrogen oxides are used as terminal electron acceptors in place of $O_2$ and produces gas as the terminal reduction product (eqn [9]). Denitrification is a multistep process catalyzed by distinct enzymes for each step:

$$NO_3^- \xrightarrow{\substack{\text{Nitrate}\\ \text{reductase}}} NO_2^- \xrightarrow{\substack{\text{Nitrite}\\ \text{reductase}}} NO \xrightarrow{\substack{\text{Nitric oxide}\\ \text{reductase}}} N_2O \xrightarrow{\substack{\text{Nitrous oxide}\\ \text{reductase}}} N_2 \quad [9]$$

**Dissimilatory nitrate reduction to ammonium** The $NH_4^+$ produced during DNRA is not assimilated but is excreted into the environment (eqn [10]). The bacteria that perform DNRA are distinct from denitrifiers. They predominate in environments that are continuously saturated, anaerobic, and have high concentrations of available C relative to $NO_3$:

$$NO_3^- \longrightarrow NO_2^- \longrightarrow NH_4^+ \quad [10]$$

## $N_2$ Fixation

Nitrogen fixation is an energy-demanding process by which prokaryotic bacteria reduce $N_2$ gas to $NH_3$ either independently or in symbiosis with various leguminous and nonleguminous plant species (eqn [11]). The most significant types of terrestrial nitrogen fixation are the legume/rhizobia symbiosis and the actinorhizal symbioses, which occur between woody shrubs and bacteria of the genus *Frankia*. In aquatic systems, photosynthetic cyanobacteria such as *Anabaena* and *Nostoc* are responsible for $N_2$ fixation.

$$N_2 + 16\text{Mg-ATP} + 8H^+ \xrightarrow{\text{Nitrogenase}} 2NH_3 + H_2 + 16\text{ADP} + 16\text{Pi} + 16Mg^{2+} \quad [11]$$

Nitrogenase is actually an enzyme complex consisting of dinitrogenase reductase and dinitrogenase, which are highly conserved, $O_2$-sensitive enzymes working in tandem to reduce $N_2$. Nitrogenase is somewhat nonspecific with respect to substrate, so it will also reduce $H^+$ to $H_2$, $N_2O$ to $N_2$, and $C_2H_2$ to $C_2H_4$. The latter reduction is the basis for the acetylene reduction assay (ARA), which is used to detect and quantify $N_2$ fixation in different environments.

## Fixed (Nonexchangeable) Ammonium

An important part of the inorganic soil N of some soils is fixed or nonexchangeable $NH_4^+$. Ammonium ions in soil exist in a chemical equilibrium as follows:

$$NH_4^+\text{(solution)} \leftrightarrow NH_4^+\text{(exchangeable)} \leftrightarrow NH_4^+\text{(fixed)}$$

**Figure 2** shows the crystalline structure of a weathered, micaceous silicate clay mineral, such as illite, showing the locations of solution, exchangeable, and fixed $NH_4^+$. Much of the fixed $NH_4^+$ can be slowly available to growing plants and soil

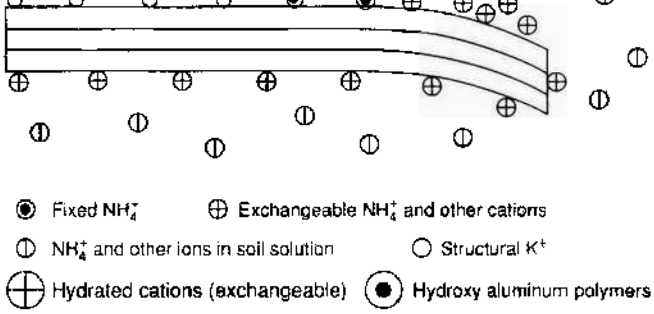


**Figure 2** Schematic of micaceous-type soil clay (vermiculite. hydrous mica) showing wedge-shaped zones and weathered edges where $NH_4^+$ can be fixed. Adapted from Dixon JB and Weed SB (eds) (1989) *Minerals in Soil Environments*, 2nd edn. Soil Science Society of America Book Series 1. Madison, WI: Soil Science Society of America with permission.

microorganisms. As plant uptake or biomass immobilization depletes the readily available $NH_4^+$, i.e., solution $NH_4^+$ and exchangeable $NH_4^+$, fixed $NH_4^+$ is released by cation replacement, especially with $K^+$. Conversely, when $NH_4^+$ is added to the soil from any source, more $NH_4^+$ is fixed to establish a new equilibrium. Therefore, the fixed $NH_4^+$ sites serve as a reserve source of N and as a sink for excess $NH_4^+$.

## Regulation and Environmental Control

Nitrogen transformations between organic and inorganic forms in soil depend on the type and availability of inorganic N and on the relative availability of inorganic N to mineralizable C.

### Cellular and Environmental Regulators of N Transformations

Microorganisms in soil will preferentially assimilate $NH_4^+$ rather than $NO_3^-$ or $NO_2^-$. Even low $NH_4^+$ concentrations in soil solutions will quickly stop $NO_3^-$ uptake and, soon after, inhibit enzymes associated with assimilatory $NO_3^-$ reduction (**Table 2**). If $NH_4^+$ concentrations in soil solutions exceed 1 mmol $l^{-1}$, assimilation by glutamate dehydrogenase will occur in which the first recognizable product is glutamate. If soil N is limiting, microorganisms will utilize an enzyme system referred to as GS/GOGAT (glutamine synthetase/glutamine oxoglutarylamino transferase) in which the first identifiable form of organic N is glutamine. Nitrate, if it is assimilated, will first be reduced to $NO_2^-$ by assimilatory $NO_3^-$ reductase and the $NO_2^-$ subsequently reduced to $NH_4^+$ by assimilatory $NO_2^-$ reductase. Thereafter, the pathways are the same as for $NH_4^+$ assimilation.

Organic N will be mineralized in soil in the absence of available inorganic soil N, although, as a general rule, N-containing compounds such as amino acids are not mineralized only to be subsequently resynthesized. Most N cycling in soil and inorganic N release is actually driven by the mineralization of soil C for biomass and energy; inorganic N accumulates in soil only when the microbial demand for N is met. When the C/N ratio of decomposing material exceeds approximately 30, net immobilization of inorganic N from soil occurs because there is insufficient N in the mineralizing material to sustain microbial biomass formation. In contrast, when the C/N ratio is less than 30, the N needs for microbial growth are met, and inorganic N is released into the soil environment. The decomposition of organic materials with very low C/N ratios can therefore cause inorganic N to accumulate briefly in excess of microbial and plant needs, with the potentially adverse effects of volatilization, runoff loss, or plant toxicity in the case of $NH_4^+$.

The three environmental factors that regulate nitrification in soil are $NH_4^+$ availability, aeration, and pH (**Figure 3**). As long as the oxidation–reduction potential ($E_h$) of soil exceeds approximately 400 mV, $NO_3^-$ will remain oxidized and nitrification of $NH_4^+$ will proceed. In contrast to most other N transformation processes, autotrophic nitrification is extremely pH-sensitive. Below pH 5.5, $NH_4^+$ oxidation will not occur, and above pH 8 $NO_2^-$ oxidation slows, because $NH_3$ inhibits $NO_2^-$-oxidizing bacteria. Ammonification and urease activity can rapidly increase the $NH_3$ concentration in soil sufficiently to prevent $NO_2^-$ oxidation and thereby cause brief $NO_2^-$ accumulation. Plant toxicity can result.

In grassland and prairie environments, autotrophic nitrification is often minimal. Although it has been suggested that this is due to allelopathic effects of the native vegetation, the most likely explanation is

**Table 2** Inhibition of $NO_3^-$ assimilation by $NH_4^+$

| *$NO_3^-$ concentration ($\mu g$ $NO_3^-$-N $ml^{-1}$)* | *N source* | *Inhibition (%)* | *Response time (min)* |
|---|---|---|---|
| 2 | $NH_4^+$ | 55 | 1 |
| | Arginine | 51 | 6 |
| | Asparagine | 50 | 1 |
| | Glutamine | 48 | 4 |
| | Glutamate | 22 | nd |
| | Aspartate | 21 | nd |
| | Alanine | 8 | nd |
| | Glycine | 1 | nd |
| 10 | $NH_4^+$ | 79 | 5 |
| | Glutamine | 72 | 3 |

nd, not detectable.

Source: Rice CW and Tiedje JM (1989) Regulation of nitrate assimilation by ammonium in soils and in isolated soil microorganisms. *Soil Biology and Biochemistry* 21: 597–602.

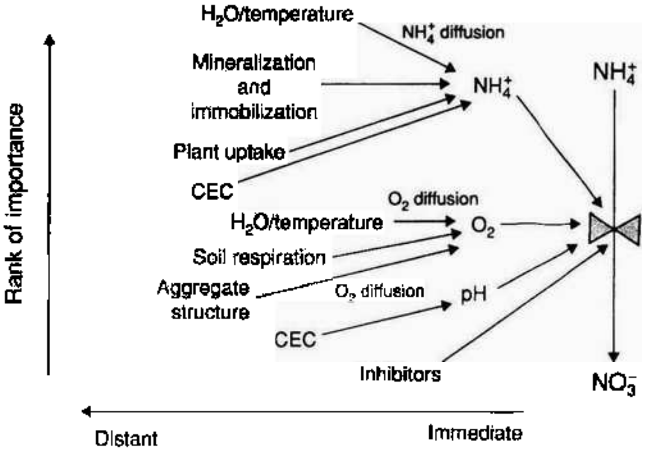


**Figure 3** Regulators of nitrification in the soil environment. CEC, cation exchange capacity.

that intense competition from plant roots and heterotrophic microorganisms deprives nitrifiers of the $NH_4$ that they require for growth.

Nitrification slows and ceases as soils become increasingly saturated. The dominant fate of $NO_3$ in most waterlogged soil environments is denitrification. Oxygen concentration is the overwhelming environmental factor controlling denitrification and DNRA (**Figure 4**). Limited C and $NO_3$ availability will also slow denitrification and affect whether incompletely reduced intermediates such as NO and $N_2O$ accumulate.

Environmental control over whether denitrification or DNRA predominates in an environment appears to be governed by the extent and duration of anoxia and by the relative proportions of soluble C and $NO_3^-$ (**Figure 5**). Denitrifiers are more competitive when C is limiting because they obtain more energy through anaerobic respiration with $NO_3$ than do $NH_4^+$ dissimilators (1121 versus 600 kJ mol$^{-1}$ of $NO_3$ reduced).

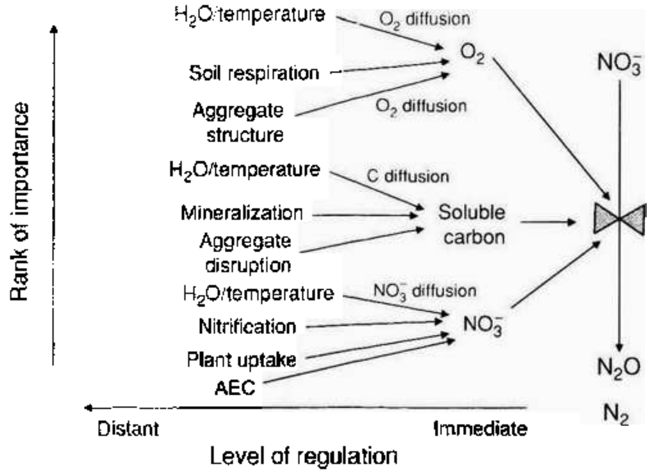


**Figure 4** Regulators of $NO_3$ reduction in the soil environment. AEC, anion exchange capacity.

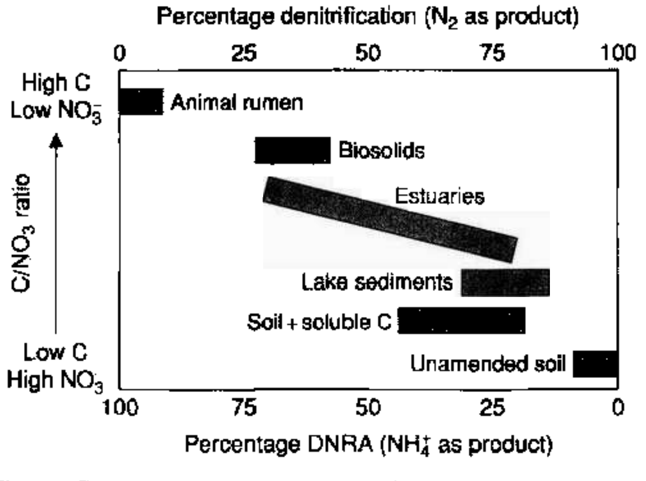


**Figure 5** Environmental control of the partitioning between denitrification and dissimilatory nitrate reduction to ammonium (DNRA). Adapted from Tiedje JM, Sexstone AJ, Myrold DD, and Robinson AJ (1982) Denitrification: ecological niches, competition, and survival. *Antonie van Leeuwenhoek* 48: 569–583.

557

2005

04

Nitrogen fixation is an energy-demanding process. Consequently, when N is available in the form of inorganic N or readily mineralizable organic N, the synthesis and activity of nitrogenase are suppressed. High soil N concentrations will also suppress the establishment of symbiotic $N_2$-fixing relationships between bacteria and plants.

### Chemical Regulators of N Transformations

Chemicals have been developed that can effectively regulate (inhibit) two important N transformations on an agricultural field scale – nitrification and urea hydrolysis. A nitrification inhibitor slows the rate of autotrophic nitrification by selectively suppressing *Nitrosomonas* bacteria in the soil (eqn [5]). A urease inhibitor functions by interfering with the enzymatic activity of urease in catalyzing the hydrolysis of urea (eqn [2]). The use of these chemicals will be discussed later as part of management practices to decrease N losses.

## Routes of N Loss From the Environment

The N transformations in soil inevitably result in intermediate or final products that can be lost from the soil environment through gaseous diffusion, leaching, or surface erosion (**Figure 6**). The route of N loss greatly depends on the type of transformation – whether oxidative or reductive.

### Reductive Losses

The major reductive losses of N occur during denitrification, which produces N oxides such as NO and $N_2O$ as intermediates and $N_2$ as a final product. All three compounds are gases that can be lost to the atmosphere. Nitrate reduction to $NH_4^+$ seemingly preserves N from loss; however, the $NH_4^+$ formed can still be leached in soils with low cation exchange capacities or lost through erosion.

### Oxidative Losses

Autotrophic nitrification frequently results in the production of NO and $N_2O$ as gaseous intermediates. The major concern with nitrification is the production of $NO_3$, which is readily leached from soil environments.

### Volatilization

Ammonium exists in chemical equilibrium with $NH_3$. The pK for this equilibrium is approximately pH 9.5. Consequently, when the soil pH is sufficiently high, or

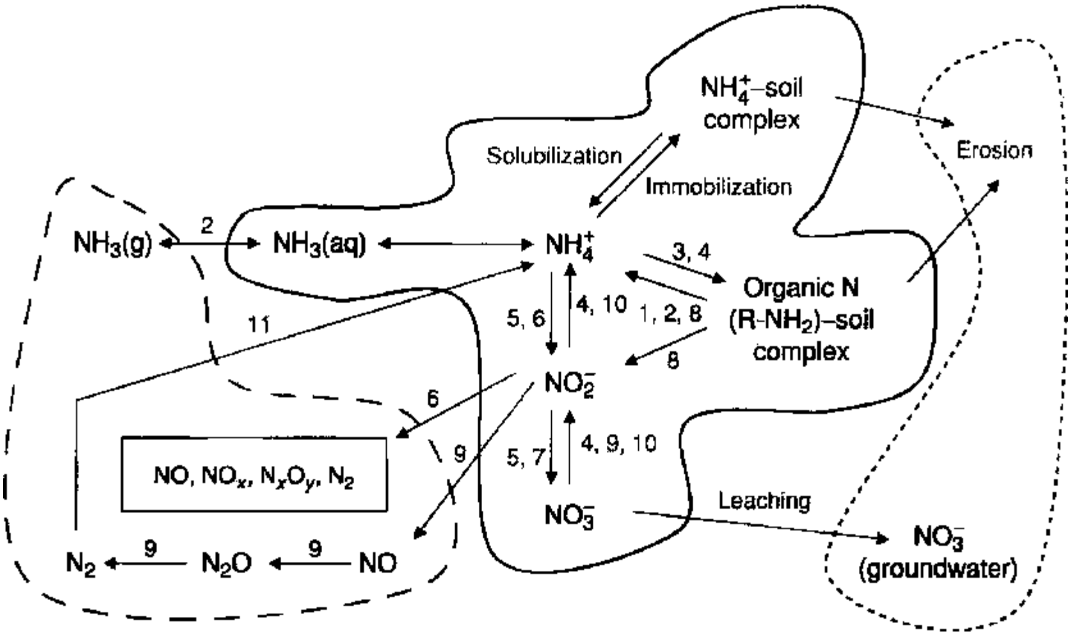


**Figure 6** Paths of nitrogen transformation and loss from soil. Numbered pathways correspond to the chemical equations in the text.

**Table 3** Some patented nitrification inhibitors

| *Chemical* | *Common name* | *Developer* | *Inhibition by day 14 (%)* |
|---|---|---|---|
| 2-Chloro-6-(trichloromethyl)pyridine | N-serve | Dow Chemical | 82 |
| 4-Amino-1,2,4-6-triazole-HCl | ATC | Ishihada Industries | 78 |
| 2,4-Diamino-6-trichloromethyltriazine | CL-1580 | American Cyanamid | 65 |
| Dicyandiamide | DCD | Showa Denko | 53 |
| Thiourea | TU | Nitto Ryuso | 41 |
| 1-Mercapto-1,2,4-triazole | MT | Nippon | 32 |
| 2-Amino-4-chloro-6-methylpyrimidine | AM | Mitsui Toatsu | 31 |

Adapted from Paul EA and Clark FE (1989) *Soil Microbiology and Biochemistry*. San Diego, CA: Academic Press with permission.

if NH$_4^+$ production is sufficiently rapid, aqueous NH$_3$ will be volatilized from soil and lost to the atmosphere as a gas. Animal manures, which contain readily mineralizable forms of organic N such as urea, can produce considerable amounts of volatile NH$_3$ due to this process.

### Interaction of N with Soil Organic Matter

A significant fate of inorganic N is its interaction with soil organic matter. Although this can retard N loss from the environment, it also temporarily renders part of the inorganic N unavailable. Studies in California and Iowa have demonstrated that much of the inorganic NO$_3^-$ and NH$_4^+$ added to soil can almost immediately become immobilized. Nitrification-mediated immobilization involves the reaction of NO$_2^-$, formed during the first step of nitrification, with soil organic matter at acidic and neutral pH to form nitrosophenols. Nitrite can also react with amino acids and other forms of reduced organic N to liberate gases as illustrated in eqns [12] and [13]:

$$NH_3 + HNO_2 \longrightarrow NH_4NO_2 \longrightarrow 2H_2O + N_2 \quad [12]$$

$$RNH_2 + HNO_2 \longrightarrow ROH + H_2O + N_2 \quad [13]$$

### Management Practices to Decrease N Losses

Figure 6 shows the transformation pathways that potentially result in losses from the N cycle. Some of these losses, including soil erosion, NO$_3$ leaching, denitrification, and NH$_3$ volatilization, can be managed to varying extents. Inhibiting nitrification and urease activity affects all of these pathways except erosion.

**Nitrification inhibitors** Several chemicals have been shown to inhibit nitrification (Table 3), but few have been proven agronomically and economically effective. The first nitrification inhibitor to be approved by the US Environmental Protection Agency was nitrapyrin, manufactured by Dow Chemical Company and marketed as N-Serve. Nitrapyrin is 2-chloro-6-(trichloromethyl) pyridine. It may be used with any $NH_4^+$-N fertilizer, including anhydrous ammonia, urea, urea–ammonium nitrate solutions (UAN), ammonium nitrate, ammonium sulfate, and animal manures.

Nitrapyrin has a relatively high vapor pressure ($2.8 \times 10^{-3}$ mmHg at 23°C). Because of its tendency to volatilize, the manufacturer initially recommended incorporation into soil immediately after application. Research in Kentucky showed that nitrapyrin was effective when sprayed directly on to N fertilizer particles, which were then surface-applied for no-tillage corn. The researchers believed that the immediate proximity of nitrapyrin and $NH_4^+$ ions together at the sites of nitrification facilitated its effectiveness, even when surface-applied.

Dicyandiamide (DCD), also known as cyanoquanidine, is a water-soluble, organic amide that is both a slow-release N source and a nitrification inhibitor. The composition of DCD is $C_2N_4H_4$, and it contains at least 65% N. It is a fine-grained, nonhygroscopic, crystalline powder, which is easily incorporated into solid, solution, or suspension fertilizers.

Widespread research on nitrapyrin and DCD shows generally that both are very effective in suppressing nitrification in soil for several weeks after application. The ratio of $NH_4^+$ to $NO_3^-$ is usually much higher in soils, and $NO_3^-$ is lower in drainage water from soils following application of ammonical fertilizers treated with nitrapyrin or DCD than the same applications without a nitrification inhibitor. A crop yield response to a nitrification inhibitor, however, occurs only if N was lost by leaching or denitrification and then only if those losses resulted in N deficiency sufficient to reduce yields without the nitrification inhibitor. Rarely are yield increases observed if nitrification inhibitors are used with much higher than optimum application rates of N. The environmental benefits of decreasing $NO_3^-$ leaching with a nitrification inhibitor would none the less be present and even more important where an excess of N was applied.

**Urease inhibitors** As in the case of nitrification inhibitors, several chemicals inhibit urease activity. Because mechanical incorporation is usually not practical, a urease inhibitor must suppress hydrolysis of broadcast-applied urea to the surfaces of soil and plant residues for no-tillage agriculture, or growing plants in pastures, forage fields, or lawns, and the effect must last long enough to allow rainfall to leach the urea molecules into the soil, where hydrolysis can occur with little risk of $NH_3$ volatilization loss from most soils.

**Table 4** Corn yield response to urease inhibitors with fertilizers surface applied for no-tillage corn

| | *Grain yields (Mg ha$^{-1}$)* | |
|---|---|---|
| *Fertilizer*[a] | *Pennsylvania*[b] | *Kentucky*[c] |
| UAN | 6.96 | 5.14 |
| UAN + ATS | 7.03 NS | – |
| UAN + NBPT | 7.40$^{*}$ | 5.52$^{*}$ |
| Urea | 6.65 | 5.27 |
| Urea + NBPT | 7.53$^{**}$ | 6.15$^{**}$ |
| Ammonium nitrate | 7.59 | 5.96 |

[a] UAN, urea ammonium nitrate solution; ATS, ammonium thiosulfate; NBPT, *N*-(*n*-butyl) thiophosphoric triamide.

[b] From Fox and Piekielek (1993).

[c] Unpublished data, Department of Agronomy, University of Kentucky.

NS, not significantly different ($P > 0.10$) by LSD.

Significantly different by LSD: $^{*}P \leq 0.10$, $^{**}P < 0.01$, respectively.

In many research studies (examples shown in Table 4), one urease inhibitor has been shown to be particularly efficacious for urea fertilizer management. Allied Chemical Corporation patented *N*-(*n*-butyl)thiophosphoric triamide (NBPT) as a urease inhibitor in 1985. Today, NBPT is marketed by Agrotain International as a technology entity that includes a practical means of treating solid urea or UAN solution with NBPT, or NBPT and DCD, at a fertilizer-blending facility and delivering the product in the field as a fertilizer application. The combination of NBPT and DCD provides urease and nitrification inhibition for a urea application.

## Environmental and Health Consequences of N Cycling

The principal environmental and health concerns associated with N cycling in soil are the potential for trace gas production and its effect on global warming and ozone depletion, the potential for nutrient loss, and the movement of N to environments where it can cause eutrophication and potential health problems.

### Global Warming and Ozone Destruction

Nitrous oxide and nitric oxide are trace gases formed during nitrification and denitrification transformations of N in soil. Molecule for molecule, $N_2O$ is 150 times more effective at adsorbing thermal radiation than is $CO_2$, the major greenhouse gas. Nitric oxide plays a well-known role in the destruction of stratospheric ozone as illustrated in the following reactions:

$$O_3 + h\nu \rightarrow O_2 + O\cdot \qquad [14]$$

$$N_2O + O\cdot \rightarrow 2NO \qquad [15]$$

$$NO + O_3 \rightarrow NO_2 + O_2 \qquad [16]$$

$$NO_2 + O\cdot \rightarrow NO + O_2 \qquad [17]$$

### Nutrient Losses

Denitrification and $NO_3^-$ leaching have been associated with the loss of up to 70% of the inorganic N in some fertilized fields. Ammonia volatilization occurring during hydrolysis of urea (eqn [2]) can result in losses of 10–50% of the N from surface-applied urea fertilizers. Ammonia volatilization due to urease activity is given as one explanation for the loss of 3–50% of the N in unincorporated manures such as poultry litter applied to neutral soils. Ammonia in high enough concentrations is a toxic gas, and its production in confined animal feeding operations (CAFOs) is a major concern. Unfortunately, venting such gases into the atmosphere is problematic for two reasons. First, the volatile N gases are odoriferous and generate nuisance complaints. Second, the volatile gases eventually return to soil or water as wet and dry deposition and can contribute to eutrophication in pristine environments.

### Groundwater Quality

Groundwater is the source of drinking water for more than half of the USA, and in rural areas it represents more than 95% of the water used for domestic purposes. Nitrate-N concentrations in streams and groundwater are typically less than $2\,\mathrm{mg\,l^{-1}}$ but have been steadily increasing in the USA and Europe. In the USA, for example, well-drained soils receiving high N inputs have median $NO_3^-$-N concentrations approaching $5\,\mathrm{mg\,l^{-1}}$ and up to 26% of sample wells have levels exceeding $10\,\mathrm{mg}\,NO_3^-\text{-N}\,\mathrm{l^{-1}}$, which is the US Environmental Protection Agency (USEPA) and Public Health Service drinking water standard. Many drinking water systems in Europe currently exceed the EC limit of $11.3\,\mathrm{mg\,l^{-1}}$.

The consequences of chronic consumption of $NO_3^-$-contaminated drinking water and vegetation can be severe and include methemoglobinemia (blue-baby syndrome) and increased potential for stomach cancer due to nitrosamine formation.

The consequences of nitrification and subsequent $NO_3^-$ leaching also have severe ecological consequences. The inability of growing plants to capture completely the N applied in fertilizer, and natural mineralization and nitrification in undisturbed soils opened to agriculture, grazing, and irrigation, has caused N pollution problems from the Chesapeake Bay in the eastern USA to the Baltic. The Gulf of Mexico, for example, receives drainage from approximately 41% of the continental USA via the Mississippi River and its tributaries. Due to nutrient leaching from agricultural watersheds, a hypoxic zone some 7000 square miles in size (approximately the same area as the state of New Jersey), which has significantly affected the shrimp and shellfish industries, has developed in the Gulf of Mexico off the Louisiana coast.

## Conclusion

The N cycle in soil is dynamic and inherently leaky. Off-site losses of N gases, soluble ions, and organic compounds are virtually impossible to eliminate completely. Nevertheless, there are several management practices for erosion control and fertilizer use that can minimize N losses from soil, and the cycle itself can be used to preserve environmental quality. For example, biological $N_2$-fixation rather than commercial fertilization will generally retain more soil N. Nitrates and nitrites in groundwater have been remediated through denitrification on increasingly larger scales to preserve groundwater quality and public health.

## List of Technical Nomenclature

| | |
|---|---|
| **Anoxia** | Conditions in which $O_2$ is absent. Synonymous with 'anaerobic' |
| **Autotrophy** | Growth using inorganic elements as C and energy sources |
| **Biomass** | Living organisms in soil, whether plants, animals, or microorganisms |
| **Eutrophic** | Excessive growth typically associated with aquatic systems in which one or more limiting nutrients such as N or P is provided |
| **Heterotrophy** | Growth using organic C as the source of C for biomass production and for energy |
| **Hydrolysis** | Addition of water to chemical bonds during chemical reactions |
| **Inhibition** | Suppression or slowing of a process for an indefinite time, usually a few weeks |

*See also:* **Enzymes in Soils; Greenhouse Gas Emissions; Microbial Processes:** Environmental Factors **Nutrient Management**

## Further Reading

Dixon JB and Weed SB (eds) (1989) *Minerals in Soil Environments*, 2nd edn. Soil Science Society of America Book Series 1. Madison, WI: Soil Science Society of America.

Fenchel T, King GM, and Blackburn TH (1998) *Bacterial Biogeochemistry*, 2nd edn. San Diego, CA: Academic Press.

Follett RF, Keeney DR, and Cruse RM (1991) *Managing Nitrogen for Groundwater Quality and Farm Profitability*. Madison, WI: Soil Science Society of America.

Fox RH and Piekielek WP (1993) Management and urease inhibitor effects on nitrogen use efficiency in no-till corn. *Journal of Production Agriculture* 6: 195–200.

Paul EA and Clark FE (1989) *Soil Microbiology and Biochemistry*. San Diego, CA: Academic Press.

Pearce RC, Zhan Y, and Coyne MS (1998) Nitrogen transformations in the tobacco float system. *Tobacco Science* 42: 82–88.

Rice CW and Tiedje JM (1989) Regulation of nitrate assimilation by ammonium in soils and in isolated soil microorganisms. *Soil Biology and Biochemistry* 21: 597–602.

Schlesinger WH (1997) *Biogeochemistry: an Analysis of Global Change*, 2nd edn. San Diego, CA: Academic Press.

Spalding RF and Exner ME (1993) Occurrence of nitrate in groundwater – a review. *Journal of Environmental Quality* 22: 392–402.

Stevenson FJ (ed.) (1982) *Nitrogen in Agricultural Soils*. Agronomy Monograph 22. Madison, WI: American Society of Agronomy.

Stevenson FJ and Cole MA (1999) *Cycles of Soil*, 2nd edn. New York: John Wiley.

Thorn KA and Mikita MA (2000) Nitrite fixation by humic substances: nitrogen-15 nuclear magnetic resonance evidence for potential intermediates in chemodenitrification. *Soil Science Society of America Journal* 64: 568–582.

Tiedje JM, Sexstone AJ, Myrold DD, and Robinson AJ (1982) Denitrification: ecological niches, competition, and survival. *Antonie van Leeuwenhoek* 48: 569–583.

US Environmental Protection Agency (1995) *Drinking Water Regulations and Health Advisories*. Washington, DC: Office of Water, USEPA.

Zumft WG (1997) Cell biology and molecular basis of denitrification. *Microbiology and Molecular Biology Reviews* 61: 533–616.

---

# Nitrates

**D S Powlson and T M Addiscott**, Rothamsted Research, Harpenden, UK

Under aerobic conditions nitrate ($NO_3^-$) is the final form of nitrogen produced in soil following the decomposition of soil organic matter or any added plant or animal material. It is the main chemical form in which nitrogen is taken up by plants, though most plants can also absorb N in the ammonium form ($NH_4^+$). **Figure 1** shows the central position of soil nitrate within the nitrogen cycle. Nitrate is a natural product that is almost always present in soil. In natural or seminatural ecosystems nitrogen is usually cycled very efficiently between soil microbes, soil solution, and plants so the concentration of nitrate in soil at any time is normally low and losses are small. By contrast, in order to grow large yields of agricultural crops, nitrogen is required in large quantities. For example, winter wheat growing under UK conditions would typically take up about 200 kg N ha$^{-1}$ and most agricultural crops, other than those grown under subsistence farming conditions, take up between 100 and 200 kg N ha$^{-1}$. Various practices are adopted in order to provide sufficient N and the use of inorganic fertilizers is particularly important, with almost 80 million tonnes of N being used annually worldwide (Table 1). Inorganic fertilizers either contain nitrate or forms of N such as ammonium or urea that are rapidly converted to nitrate by soil microorganisms. Alternatively N may be supplied by adding animal manure or by incorporating residues of plants rich in nitrogen, such as N-fixing legumes, which may have been grown as the previous crop, as a green manure or as part of a pasture which is then plowed up. In all cases, except in anaerobic conditions such as flooded rice, much of the N from these organic sources will be converted to nitrate before being absorbed by the crop. It is highly desirable to utilize N from organic materials but it is mistaken to assume that losses of nitrate are necessarily less than with inorganic fertilizers as the production of nitrate from organic sources is determined by the activity of soil microorganisms and not by crop demand – some nitrate is produced at times when crop uptake is small and the risk of nitrate loss is large.

The nitrate ion is extremely soluble in water and can easily be washed out of soil to natural waters. There are two main pathways by which nitrate ions dissolved in soil water are lost through the movement of water: (1) leaching – the vertical movement of water and its solutes through the soil profile to either groundwater or field drains; and (2) surface runoff – the horizontal movement of water and solutes over the soil surface until it reaches a river or lake. Surface runoff occurs where the soil is either compacted or is naturally poorly draining, e.g., due to high clay content. A variation on surface runoff is sometimes termed subsurface flow, in which water infiltrates the soil to some extent but then moves horizontally due to impeded drainage, for example due to compaction at the base of the plow layer in cultivated soil or because of a change in soil physical properties with depth.

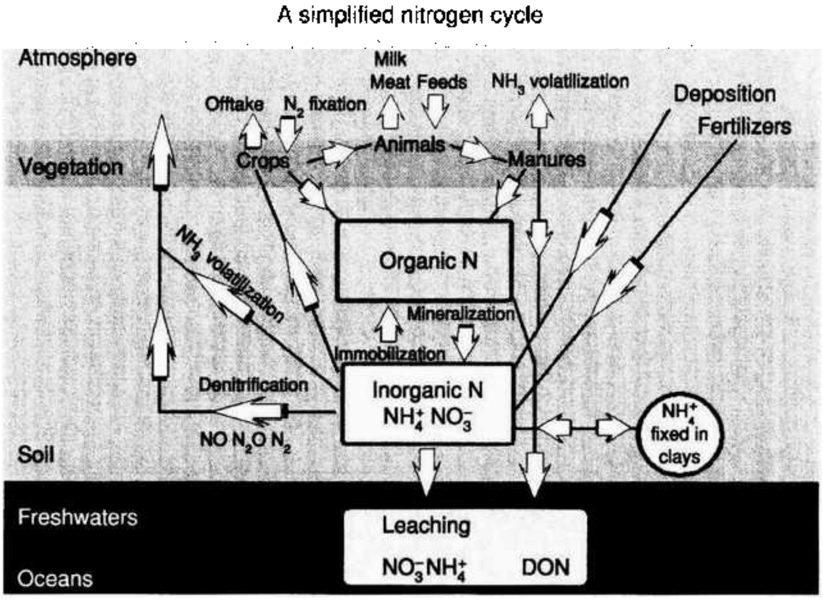


**Figure 1** Soil nitrate within the nitrogen cycle. Reproduced with permission from Davies DB (2000) *Soil Use and Management* 16: 142–144.

**Table 1** World consumption in 1994/1995 of nitrogen in various types of fertilizer

| *Type of fertilizer* | *N consumed globally (million tons)* | *% of world total N usage* |
|---|---|---|
| Ammonium nitrate (AN) | 6.58 | 9.0 |
| Calcium ammonium nitrate (CAN) | 3.69 | 5.1 |
| Urea | 31.57 | 43.3 |
| Ammonium sulfate | 2.37 | 3.2 |
| Injected ammonia gas (anhydrous ammonia) | 4.20 | 5.8 |
| N solutions | 3.79 | 5.2 |
| Other 'straights' (i.e., N not mixed with other nutrients) | 9.05 | 12.4 |
| N in compound fertilizers | 11.68 | 16.0 |
| Total | 72.93 | 100.0 |

Source: *World Fertilizer Consumption Statistics* (1997), No. 28, p. 17. Paris: International Fertilizer Industry Association.

Compared to natural ecosystems, the risk of nitrate loss through leaching, surface runoff, or denitrification is much greater in agricultural soils because: (1) nitrate concentration in soil is high for some periods of the year; and (2) at least with arable crops, soil is bare for part of the year so there is no removal of nitrate by plants. Any loss of N from agriculture is an economic loss for the farmer and a waste of a valuable resource for society, so much research over the last century has been directed at minimizing losses. Since about the 1970s issues of water pollution (some genuine and some not) have led to regulations and policies to limit the concentration of nitrate in water and this has become a main driver of research and a major influence on farming practice. In the European Union a limit of $50\,\mathrm{mg\,l^{-1}}$ of nitrate (equivalent to $11.3\,\mathrm{mg\,l^{-1}}$ of nitrate-N) has been imposed on any ground- or surface-water to be used as a source of drinking water and any surface water where eutrophication is considered to be a risk. In the USA the limit is $45\,\mathrm{mg\,l^{-1}}$ of nitrate, equivalent to $10\,\mathrm{mg\,l^{-1}}$ of nitrate-N. These limits have a significant influence on agricultural practices: for example, under the climatic and agronomic conditions of the UK, annual leaching losses from arable soils are rarely less than $20\,\mathrm{kg\,N\,ha^{-1}}$ and often twice this. But in the drier eastern regions, which are important for arable cropping, excess winter rainfall is often only 200 mm or less: under these conditions a nitrate loss of only $20\,\mathrm{kg\,N\,ha^{-1}}$ will cause the nitrate concentration limit in water to be exceeded.

In addition to issues of water pollution, nitrate can also be involved in the emission of environmentally damaging gases to the atmosphere. Under anaerobic conditions nitrate can be microbially reduced to the gas nitrous oxide ($N_2O$) which is a powerful greenhouse gas and also has a role in the destruction of ozone in the stratosphere; this provides a good reason for minimizing the amount of excess nitrate in the environment.

## Chemical Structure and Physical Properties

The nitrate ion, $NO_3^-$, has a symmetrical planar trigonal structure. It is negatively charged and so is termed an anion. Salts formed by the nitrate ion with cations (e.g., $Ca^{2+}$, $K^+$, $Na^+$) are generally soluble and, in soil, are always present in solution. These salts are dissociated so it is the nitrate ion that is the entity of interest. The practice of referring to nitrates in soil or water is incorrect – nitrate is a unique species, not a plural.

Being an anion, nitrate is attracted to positively charged surfaces. Some acid soils contain positive surface charge that can retain nitrate ions, but this is the exception. In the vast majority of soil worldwide, especially those used for agriculture, nitrate is not retained in this way and remains in soil solution and moves through the processes of mass flow or diffusion.

## Biological Transformations in Soil

Numerous organisms in soil are involved in the breakdown of organic matter to inorganic (mineral) materials, the process termed mineralization. During aerobic mineralization, C in organic materials is converted to $CO_2$ and N is initially converted to ammonium; this step of N transformation is termed ammonification and is carried out by a wide range of heterotrophic organisms including various fungi and bacteria. A much smaller range of autotrophic bacteria are responsible for the conversion of ammonium to nitrate via a two-step process, together termed nitrification. The first step is conversion of ammonium to nitrite; classically this is described as being mediated by bacteria of the *Nitrosomonas* species, but it is now known that the less-studied species *Nitrosolobus*, *Nitrosospira*, and *Nitrosococcus* are at least as important:

$$2NH_4^+ + 3O_2 \rightarrow 2NO_2 + 4H^+ + 2H_2O + \text{energy}$$

Nitrite rarely accumulates in soil, which is fortunate as it is toxic to plants and many microbes. It is rapidly oxidized in aerobic soils to nitrate by *Nitrobacter* species:

$$2NO_2^- + O_2 \rightarrow 2NO_3^- + \text{energy}$$

Although the nitrifying bacteria in soil cover a narrow group of genera, they are ubiquitous, so it can be assumed that any ammonium (whether from added fertilizer or from mineralization) will be converted to nitrate within a period of days to weeks provided the soil is aerobic and conditions of temperature, moisture, and pH are conducive for general microbial activity. The first step of nitrification produces $H^+$ ions and is thus an acidifying reaction. This is illustrated in Figure 2, which shows changes in soil pH over a period of 150 years in a long-term grassland experiment at Rothamsted, UK. In the treatment where N fertilizer is applied as ammonium sulfate, pH decreased from 5.7 to 3.7. In the treatment receiving the same quantity of N in the nitrate form there was little change in soil pH.

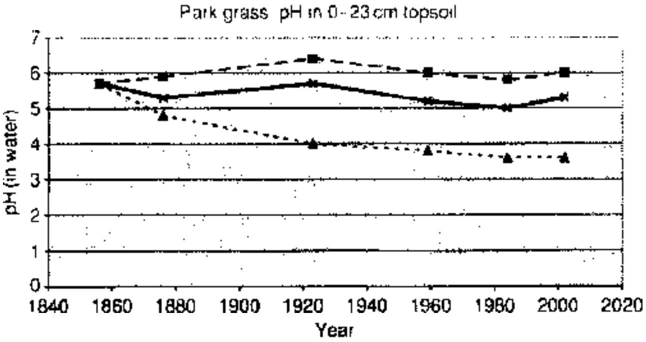


**Figure 2** Influence of the long-term application of ammonium sulfate fertilizer on soil pH in the Park Grass grassland experiment at Rothamsted, UK. Plots received N as either ammonium sulfate or sodium nitrate every year since 1856. Nitrification of ammonium in the ammonium sulfate treatment caused an acidification of soil but application of N already in the nitrate form did not. × Control; ■ sodium nitrate; ▲ ammonium sulfate.

Nitrate in soil can be destroyed by the action of denitrifying bacteria when oxygen is in short supply. Under anaerobic conditions the nitrate ion can be used instead of $O_2$ as the acceptor of free electrons produced during respiration. The products are a mixture of nitrous oxide ($N_2O$) and nitrogen gas ($N_2$) – the sequence can be represented as:

$$NO_3^- \rightarrow NO_2^- \rightarrow N_2O \rightarrow N_2$$

The extent to which the reduction proceeds is strongly influenced by the pH and temperature of the soil. At a low temperature and a pH less than 5, the ratio of $N_2O$ to $N_2$ is about 1, but at a temperature of 25°C or more and a pH greater than 6, most of the $N_2O$ is reduced to $N_2$. Soils are not usually uniform and even in mainly aerobic soils there are often small localized anaerobic zones in which denitrification occurs. This spatial variability, together with the variable ratio of $N_2O$ to $N_2$, makes denitrification very difficult to measure. It also implies that nitrification and denitrification may be occurring simultaneously close to each other in the soil. Nitrous oxide is also produced to some extent as a by-product of nitrification (Figure 3) and it is often difficult to distinguish between the sources. The environmental impacts of nitrous oxide are mentioned later.

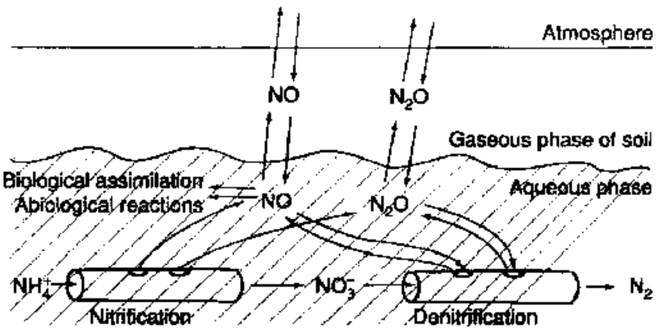


**Figure 3** Formation of nitrous oxide ($N_2O$) and nitric oxide (NO) through denitrification and nitrification. Reproduced with permission from Davidson EA (1991) Fluxes of nitrous oxide and nitric oxide from terrestrial ecosystems. In: Rogers JE and Whitman WB (eds) *Microbial Production and Consumption of Greenhouse Gases: Methane, Nitrogen Oxides, Halomethanes*, pp. 219–235. American Society for Microbiology.

## Forms of Nitrogen Fertilizer

In Europe ammonium nitrate or calcium ammonium nitrate (CAN) is widely used. These forms supply 50% of their N as pure nitrate and so instantaneously increase the quantity of nitrate in topsoil from a very low value, typically about 10 kg nitrate-N $ha^{-1}$, to 50 kg nitrate-N $ha^{-1}$ or more depending on the amount applied. A typical application of N to an arable crop in spring, at the beginning of the rapid crop growth period, might be 100–200 kg N $ha^{-1}$, but usually divided into two or more portions. The ammonium N is typically nitrified to nitrate over a period of 1–4 weeks. Worldwide, urea is by far the most commonly used form of nitrogen fertilizer. Before becoming available for plant uptake, the N in urea first has to be converted to ammonium (the process of urea hydrolysis) but this occurs within a few days. The ammonium is then nitrified within 2–3 weeks. Thus, in practice, the difference between urea and ammonium nitrate or CAN in terms of impact on soil nitrate is fairly small. These transformations can be studied by labeling the N in urea or either the ammonium or nitrate parts of ammonium nitrate with $^{15}N$, the heavy isotope of nitrogen. In some experiments in northern France in which $^{15}N$-labeled fertilizers were applied to field plots of winter wheat in April, intact urea molecules were virtually undetectable in soil within a week of application. All of the N derived from urea was either converted to ammonium or nitrate, absorbed by the crop, or immobilized by soil microorganisms within 2 weeks of application. Where ammonium nitrate was applied, the majority of the ammonium had been converted to nitrate or absorbed by plants or microorganisms within 2 weeks. For applications in March, when soil temperatures were lower, the transformations were slower but the trends the same.

A small but potentially significant difference between urea and nitrate-containing fertilizers is that the peak concentration of nitrate in soil is likely to be less with urea. Thus urea may give a decreased risk of nitrate loss by leaching or denitrification shortly after application. However, with urea some N may be lost as ammonia gas under some circumstances and slightly more may be absorbed by microorganisms as they tend to take up ammonium rather than nitrate if both are available. Overall agronomic experiments show only small differences in crop utilization of N from urea compared with nitrate-containing fertilizers.

A number of additives are available commercially that retard either urea hydrolysis or nitrification, so retaining a greater proportion of N from fertilizer in the ammonium form and decreasing the amount as nitrate. In principle these additives (urease inhibitors or nitrification inhibitors) should decrease the losses associated with nitrate by decreasing the quantity in soil. Whilst this can be useful for nitrate coming from fertilizer, the impact of inhibitors on overall nitrate loss is likely to be small because it is nitrate derived from the mineralization of soil organic matter that is usually the main source. Also, until now, their high cost has precluded their use from normal agricultural crops; they are used to some extent to make fertilizer N slowly available to certain high-value horticultural crops.

## Management of Nitrate in Agriculture

### Loss of Nitrate from Fertilizer During the Crop-Growing Period

The period immediately after N fertilizer application is a time of high risk of nitrate loss. During this period nitrate concentration in soil is high as a result of the added fertilizer and the mineralization and nitrification of N derived from soil organic matter. The concentration rapidly declines as crop uptake proceeds. One method of measuring the total loss of N from fertilizer during the growing season of a crop is to apply N fertilizer labeled with $^{15}N$, the heavy isotope of nitrogen. In this way N derived from fertilizer can be distinguished from N derived from soil organic matter or other sources in soil as these are predominantly composed of $^{14}N$. $^{15}N$ is a stable isotope, so is not radioactive and can be used in field or laboratory experiments without the need for special safety procedures. **Figure 4** shows the results of such an experiment in which about 170 kg N $ha^{-1}$ was applied in spring to autumn-sown wheat in south-east England. The N was applied as ammonium nitrate with both the ammonium and nitrate parts labeled with $^{15}N$, i.e. $^{15}NH_4$ $^{15}NO_2$. About two-thirds of the applied $^{15}N$-labeled N derived from fertilizer was recovered in the above-ground parts of the crop when it was harvested. A further 20% was retained in the soil, almost entirely as organic forms as this N had

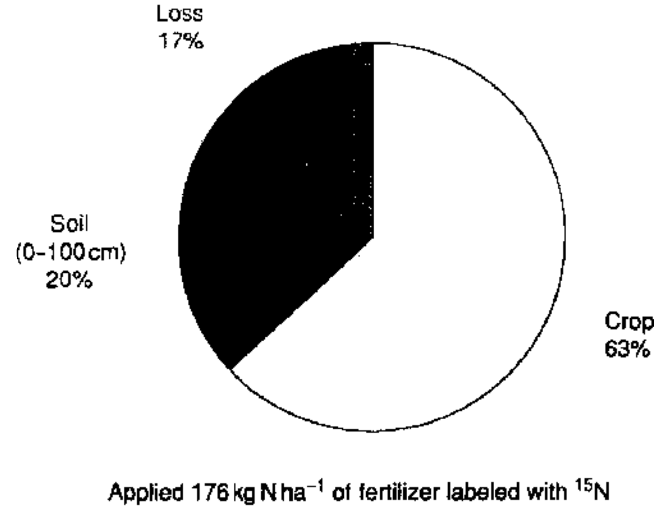


**Figure 4** Fate of $^{15}$N-labeled fertilizer (double-labelled ammonium nitrate, $^{15}NH_4$ $^{15}NH_3$) applied to winter wheat in south-east England. N was applied as a single dressing in spring (April) to a healthy crop of winter wheat, sown the previous autumn. The crop and soil were sampled at the time of harvest in August.

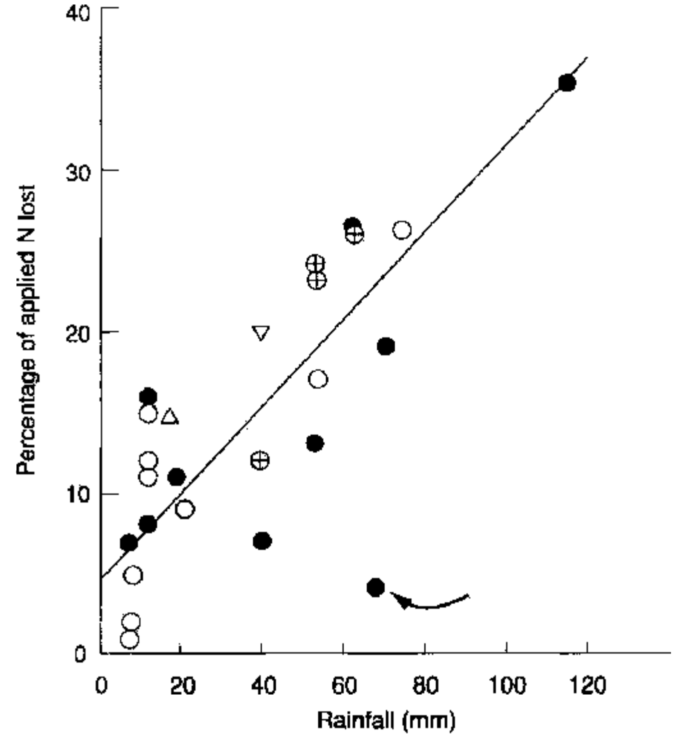


**Figure 5** Relationship between loss of $^{15}$N-labeled fertilizer applied to crops in spring and rainfall in the 3-week period following application. All circles (●, ⊕, ○) refer to winter wheat crops. Different symbols indicate crops receiving rates of fertilizer N known to be above (⊕) or below (○) that giving maximum grain yield. All other wheat crops indicated by ●. Arrowed result excluded from regression. Results included for experiments with forage maize (▽) and canola (oilseed rape, △) but used in calculating regression: Percentage of applied N lost = 5.0 + 0.264 (rainfall in 3 weeks after N application, mm). Reproduced with permission from Powlson DS, Hart PBS, Poulton PR, Johnston AE and Jenkinson DS (1992) Influence of soil type, crop management and weather on the recovery of 15N-labelled fertilizer applied to winter wheat in spring. *Journal of Agricultural Science, Cambridge* 118: 83–100.

been incorporated into wheat roots, soil organisms, and soil organic matter – this can also be seen in **Figure 6**, later. About 17% of the applied N had been lost – it could not be found in the crop or in soil to a depth of 1 m. This result is typical for healthy well-managed winter wheat growing in a temperate climate. In other experiments in similar environments losses ranged from less than 5% of applied fertilizer N to a maximum of 35%. Other studies indicated that in this environment dentrification was the major and most common cause of loss; but in years when rainfall was particularly heavy soon after N fertilizer application nitrate leaching became more important, as expected, especially in sandy soils. Taking together results from a series of similar field experiments in south-east England, there was a correlation between the percentage of the applied fertilizer N lost during the growing season and the total quantity of rain falling in the 3-week period following fertilizer application (**Figure 5**). This emphasizes the importance of the vulnerable period for loss when much nitrate is present in soil because crop uptake is not yet complete. It has also been found that any factors such as disease or weeds that inhibit crop growth (and uptake of N) tend to increase N losses from soil.

Although losses of fertilizer N can be relatively low in some situations, as described above, this is not universally the case. There are many examples of much lower recoveries, implying losses of 50% or sometimes even more. In some cases this is because of high rainfall in the period soon after N application – if this is an occasional event there is little that can be done to avoid it other than for the farmer to take note of local weather forecasts and for meteorologists to improve these as much as possible. There are, however, some measures that can be taken to decrease the risk of loss. For some crops and growing conditions it is common to apply N fertilizer at the time of sowing the crop, or even before. This means that nitrate is present in soil for a prolonged period as uptake by the crop will not start until seeds have germinated and the plants reached a certain size; depending on the crop and climatic region this could be as long as 2 months, allowing considerable opportunity for losses by leaching or denitrification unless it is a region where rainfall is very low at that time. Eliminating N applications at a time when the crop is not growing rapidly is one means of decreasing N losses and increasing the efficiency of use of nitrate derived from fertilizers.

An example of inappropriate timing of N application is shown in **Table 2**. In these experiments $^{15}$N-labeled potassium nitrate was applied to winter wheat at the time of sowing the crop in autumn. Under the climatic conditions of north-west Europe

**Table 2** Fate of $^{15}$N-labeled fertilizer applied to winter wheat at sowing time in autumn in south-east England

| *Soil type* | *Loss of autumn-applied $^{15}$N at the time of harvest (%)* |
|---|---|
| Silty clay loam | 46 |
| Silty clay loam | 63 |
| Silty clay loam | 60 |
| Sandy clay loam | 39 |
| Sandy loam | 78 |

wheat grows to a limited extent in autumn and early winter, typically taking up about 20 kg N ha$^{-1}$, and then no more until the main growth period in spring. Thus, in most conditions, little of the applied nitrate is taken up during the winter period and is exposed to loss, mainly by leaching or surface runoff as low temperature in winter limits denitrification. In maritime climates rainfall in winter exceeds evapotranspiration so water can move through the soil and cause leaching of nitrate, either continuously or intermittently. In continental climates soil is frozen during much of the winter so leaching does not occur. When the soil thaws in spring, a large amount of leaching (and/or surface runoff) occurs in a short period but the overall effect on nitrate loss is similar. Table 2 shows that between 39 and 78% of the autumn-applied N was lost from the crop/soil system, almost certainly by nitrate leaching. In the past it was thought that applying fertilizer N at the time of sowing was beneficial for early growth of the crop and 'seedbed' applications were recommended. It is now known that there is no benefit and it simply causes an unnecessary loss of N, and is a waste of money. A similar inappropriate timing of N fertilizer is the situation where N is applied at about sowing time of a crop and is immediately followed by irrigation. In this situation some of the nitrate in soil, not yet taken up by the newly established crop, can be leached below rooting depth or denitrified depending on soil and climatic conditions.

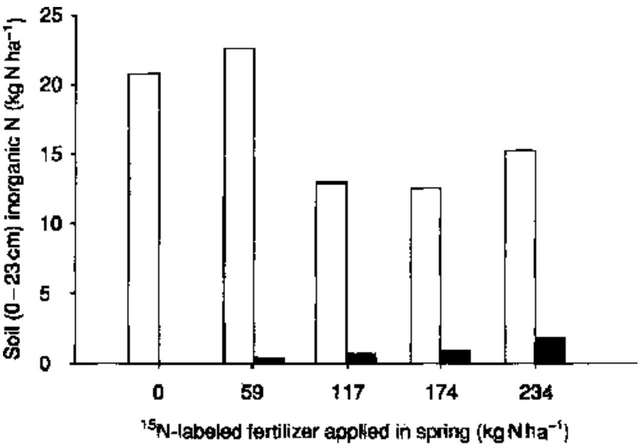


**Figure 6** Labeled and unlabeled nitrate in plow-layer soil (0–23 cm) at the time of harvesting winter wheat. A range of rates of fertilizer N, labeled with $^{15}$N, was applied in spring. Almost all the nitrate present in soil at the time of harvest was unlabeled, indicating that it was derived from the mineralization of soil organic N, not from fertilizer. Additional nitrate was present in deeper soil layers, almost all unlabeled. This soil-derived nitrate, plus additional nitrate from later mineralization, is the main source of nitrate leaching to water during the subsequent winter. □ unlabeled N; ■ labeled N. Reproduced with permission from Macdonald AL, Powlson DS, Poulton PR, and Jenkinson DS (1989) Unused fertilizer nitrogen in arable soils–its contribution to nitrate leaching. *Journal of the Science of Food and Agriculture* 46: 407–419. © Society of Chemical Industry.

### Losses of Nitrate Derived from Soil or Fertilizer Outside the Main Crop Growth Period

Although some losses of fertilizer-derived nitrate do occur during the growing period of agricultural crops, in most environments the largest losses by nitrate leaching or surface runoff occur when crops are not actively growing or are not present at all. The examples discussed above of fertilizer being applied at the time of sowing rather then just before the main growing period illustrate this. It is obviously important to know the source of nitrate in soil during these periods of high risk. It is commonly assumed that it is unused fertilizer but much research has shown this is not usually the case for the major arable crops in the world or for pasture. Short-season horticultural crops can be exceptions to this generalization as they are often given very large rates of fertilizer N and this can be used inefficiently, leaving significant residues unused in the soil.

Experiments using fertilizer labeled with $^{15}$N have been informative in identifying the sources of nitrate in soil in autumn and winter. **Figure 6** shows the quantities of nitrate in soil at the time of harvesting a winter wheat crop in England. The fertilizer applied in spring had been labeled with $^{15}$N, but the results show that very little of the nitrate left in soil after harvest in late summer was derived from this – the vast majority of the nitrate present at this time, and subject to leaching during the forthcoming winter, was unlabeled because it was derived from the mineralization of organic matter in the soil. Similar results have been obtained from numerous other studies with crops such as wheat, barley, and maize in Europe, North America, and elsewhere.

Winter losses of nitrate by leaching arise because the soil microbes responsible for mineralization and nitrification are active when conditions of temperature and moisture are suitable, which does not always coincide with crop uptake – this is illustrated in Figure 7. N mineralized and nitrified in spring does coincide with the major period of N uptake for most crops but that mineralized in autumn or early winter, before low temperature stops the process, is likely to remain in soil and be exposed to a high risk

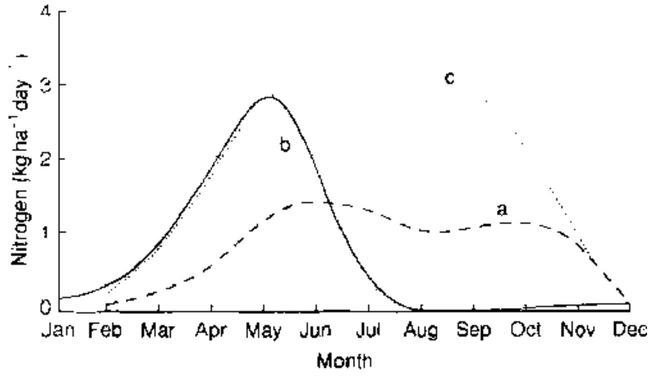


**Figure 7** Representation of the time course of (a) mineralization of soil organic N to produce nitrate. (b) N uptake by an autumn sown arable crop and (c) N uptake by grass. — wheat; ····· grass; - - - mineralization. Reproduced from Addiscott TM, Whitmore AP, and Powlson DS (1991) *Farming, Fertilizers and the Nitrate Problem*. CAB International.

of leaching. Leaching may occur intermittently during winter (as in maritime climates where soil is not frozen during winter) or in a shorter period after the spring thaw in continental climates, as described above for autumn-applied fertilizer. Unfortunately, many agricultural practices that are beneficial for crop nutrition and soil quality tend to worsen the problem of winter nitrate loss. For example, additions of organic manures lead to increased production of nitrate at all periods of the year – partly when it will be absorbed by the growing crop (spring and summer) but partly when it will be at high risk of loss due to low crop uptake and the movement of water through soil (autumn and winter). Similarly, plowing up pasture in a mixed grassland/arable rotation such as organic farming brings both benefits and problems. Nitrate produced during the growing season of the subsequent crop substitutes for inorganic fertilizer that would otherwise be needed, but that produced too late for uptake in summer or too early for spring uptake will be subject to leaching.

Winter leaching losses from grass can be less than from arable crops as the plants are present throughout the year and able to absorb nitrate later in the autumn than an arable crop and start earlier in the spring, as the grass will already have a well-established root system. In this sense, grassland is more like a natural ecosystem with plants present throughout the year. However, low leaching is generally only the case with cut grass. In grassland that is grazed by animals, or has slurry applied, mineralization of these organic inputs can lead to nitrate formation in late autumn that is not taken up by the grass and is at risk of being leached.

As a general principle, arable crops that are sown in autumn rather than spring minimize the opportunity for leaching losses during winter – some nitrate is taken up in autumn before leaching begins and roots are already present in spring. With a spring-sown crop, the soil is bare during winter and there is a period of several weeks after sowing in the spring before the crop takes up nitrate in significant quantity. This trend is generally shown when measurements of drainage are made from arable fields or nitrate losses are estimated from measurements of nitrate concentration in soil water made using porous ceramic cups. However, it cannot be assumed that the presence of an autumn-sown crop solves the problem as the extent to which the crop is an effective sink for nitrate is highly dependent on the time of sowing and on weather conditions. A crop sown rather late in autumn and slow in establishing may only absorb 10 kg ha$^{-1}$ of nitrate before winter leaching begins and thus have little impact on the amount of nitrate lost. By contrast, a vigorously growing early-sown crop may absorb 30–50 kg ha$^{-1}$ of nitrate and greatly decrease losses. Where spring crops are grown, a management practice sometimes used to decrease nitrate loss is a winter cover crop or ‘catch crop.’ This is a crop grown specifically to absorb nitrate during the period from late autumn to early spring that is incorporated into the soil before sowing the spring crop. Cover crops can sometimes be very effective at decreasing nitrate leaching, although they do introduce additional agronomic difficulties. Also, as the incorporated plant material is mineralized it is common for rather little of it to become available to the subsequent spring-sown crop – mineralization is usually fairly slow, occurring over several years, and some contributing to nitrate leaching in the next winter.

In a series of experiments using hydrologically isolated plots in central England (the Brimstone Farm experiment), varied crop rotations and management practices were compared for their influence on winter leaching of nitrate. Drains were installed under each plot so that water could be collected and analyzed for nitrate. Losses were greatest when soil was bare prior to a spring-sown crop. They could be decreased in that year by growing a cover crop but this led to increased leaching in later years. The only way to decrease nitrate loss greatly was to abandon arable cropping and convert the land to unfertilized grass. Overall it was concluded that, under the soil and environmental conditions of the site, continuous autumn cereals were the best compromise between decreased nitrate movement to water and profitable agriculture.

### Improving the Management of Inorganic Fertilizer

Although it is generally true that nitrate from mineralization of organic matter in soil, or added to soil, is the major source of nitrate losses, there are

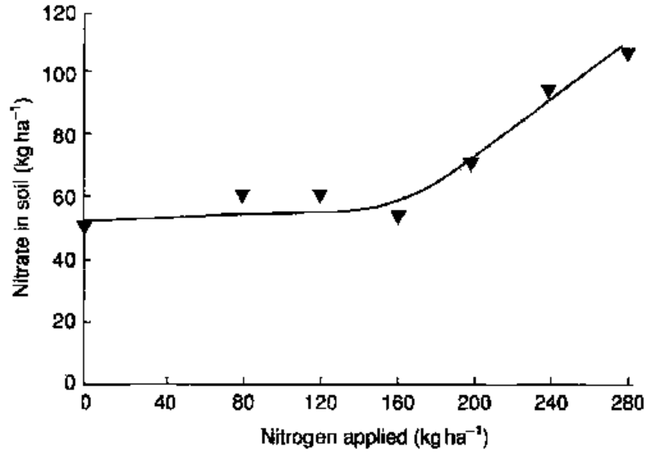


**Figure 8** Mean quantities of nitrate in soil to a depth of 90 cm after harvesting winter wheat at eight sites in central and eastern England. Reproduced with permission from Cheney K (1990) Effect of nitrogen fertilizer rate on soil nitrate nitrogen content after harvesting winter wheat. *Journal of Agricultural Science, Cambridge* 114: 171–176.

exceptions. If more inorganic fertilizer is applied to a crop than is required, some of the excess may be left in the soil and contribute to the pool of nitrate liable to be lost in the period after crop harvesting – an example is shown in **Figure 8**. The amount of nitrate in the soil profile (to a depth of 90 cm) was measured after harvesting winter wheat at eight sites in central and eastern England. A range of fertilizer N rates, including zero, had been applied in spring. **Figure 8** shows the mean nitrate quantities for the eight sites. For rates of fertilizer N between zero and 160 kg N ha$^{-1}$ (the average rate giving maximum crop yield), about 60 kg ha$^{-1}$ of nitrate-N remained in soil but when this application was exceeded, residual nitrate rapidly increased to over 100 kg ha$^{-1}$ of nitrate-N. It should be noted that, even where no fertilizer N had been applied, 50 kg ha$^{-1}$ of nitrate-N remained in the soil, so small decreases in applications below a rate likely to give approximately maximum yield will have little influence on residual nitrate and nitrate leaching.

**Figure 9** shows the measured amounts of nitrate lost in drainage water during winter from plots of the long-term Broadbalk wheat experiment at Rothamsted, UK. Each plot has received a specific rate of N fertilizer annually for many years. Up to the rate of N fertilizer corresponding to maximum grain yield (and also the calculated economic optimum rate), nitrate losses were about 10 kg N ha$^{-1}$ or less but after this fertilizer N rate was exceeded, losses increased greatly. Other studies have shown that a plot receiving organic manure annually loses the largest amount of nitrate. To avoid excessive amounts of fertilizer being applied it is necessary to estimate in advance two quantities, both of which are difficult to predict: (1) the likely requirement of N by the specific crop; and (2) the amount of N likely to become available to the crop from mineralization or other sources. Estimating crop N requirement is best achieved by referring to the yields achieved on that field or region in the past. Of course, the value obtained will only be an estimate as weather in the coming growing season will be unknown at the time. For estimating N supply from sources other than fertilizer the following approaches are currently used or being developed:

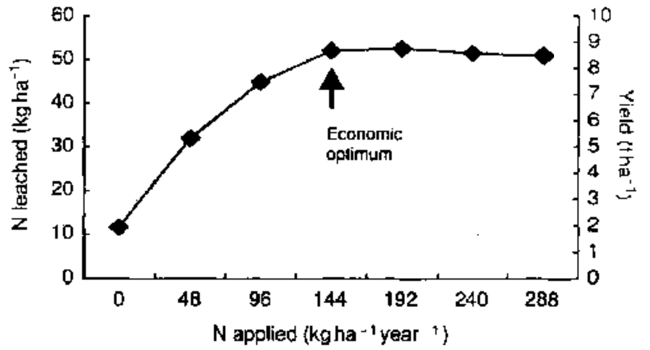


**Figure 9** Nitrate lost in drainage water from plots of the long-term Broadbalk Experiment at Rothamsted that have received uniform applications of inorganic N fertilizer for many years. ◆, wheat grain yield; , N leached. Reproduced from Goulding K (2000) Nitrate leaching from arable and horticultural land. *Soil Use and Management* 16: 145–151.

1. Using computer models of N cycle processes to estimate the production of inorganic N from mineralization based on soil type, past cropping, past inputs of organic manures and either average or expected climatic conditions for the site. Such models also include estimates of likely N losses based on soil type, timing of agronomic operations, and weather. Modeling systems may also include a model for crop growth or this may be estimated separately. Decision support systems based on computer models have been developed in many regions of the world over the last 20 years and, in principle, can be highly site-specific. The availability of home computers and computerized farm record-keeping has made this approach realistic in many countries. The models used range from simple nutrient balances to complex descriptions of the main N cycle processes. The more complex the model, the more detailed data are required to run it – a judgment has to be made between simplicity and complexity.
2. Using tables of typical N supply values based on runs of models for a range of conditions or measurements from specific experiments. This system can never be as site-specific as a modeling system but is often very effective, low-cost, and requires a minimum of specialist personnel.
3. Measurements of inorganic N in the soil profile before the main crop-growing period. This is often termed the $N_{min}$ method. Assumptions have to be made regarding the rooting depth of the crop and amount of inorganic N that will be lost and so not

be available to the crop. In regions with fairly stable weather conditions (e.g., continental areas within Europe) and soils that contain few stones and so are easy to sample (e.g., loess), this method has proved successful. The measurements can also be used to adjust predictions based on models or tables.

4. Variable N applications. In some situations, where fields are large and soils are not uniform, it may be helpful to adjust N applications according to the variations in soil type and crop growth – giving larger applications where growth can reasonably be expected to be most vigorous and small applications where it is constrained. The technology for variable applications is now readily available but the limitations to its use are: (1) insufficient understanding of the factors influencing crop growth in a particular year; and (2) insufficient data on soil variability. Maps of yield variation from previous years are easy to obtain with modern combine harvesters but it is often difficult to interpret their significance for fertilizer use in future years.

### Regulations Affecting N Use in Agriculture

In many parts of Europe and North America concern about nitrate entering water has led to regulations being imposed on agricultural practices. In the European Union areas subject to regulation are termed 'nitrate-vulnerable zones' (NVZs). The regulations are based on current understanding of the causes of nitrate loss, especially the need to minimize nitrate in soil during the vulnerable winter period. The measures adopted in NVZ regulations include the following:

1. Closed periods for N applications. No applications of organic or inorganic forms of N during specified periods in winter appropriate for climate and growing conditions in the area (about 5 months, e.g., 1 September to 1 February for inorganic fertilizer to arable land in the UK). The closed season is shorter for grassland, recognizing that N uptake is possible over a longer period than for arable crops.

2. Limits on quantities of N applied. The farmer is required to use a recognized method for assessing crop requirement and this must not be exceeded. For organic manures the maximum allowed application decreases after 4 years, recognizing that the continued use of organic manure will increase the amount of mineralizable N in soil. For example, in the UK the maximum for grassland is initially 250 kg ha$^{-1}$ of total N in manure but decreases to 170 kg ha$^{-1}$.

3. Application controls. Limitations on applications when weather conditions could lead to runoff to adjacent surface waters, e.g., because soil is frozen or waterlogged. Also ensuring that applications are made accurately and evenly.

4. Slurry storage. Ensuring that a farm with animals has sufficient capacity to store slurry during the closed period.

5. Record-keeping. Keeping adequate records on N applications, cropping and animal numbers for at least 5 years.

## Nitrate and Health

One of the main drivers for agricultural measures to decrease losses of nitrate from soil to water has been the claim that nitrate in drinking water is a cause of two serious diseases. The evidence for the link has never been strong and medical research now indicates that it does not exist at all. There is also evidence that some intake of nitrate has a positive effect on health.

### Stomach Cancer

In the past it was claimed that nitrate ($NO_3^-$) from drinking water or vegetables was reduced to nitrite ($NO_2^-$) by bacteria in the mouth and that when nitrite entered the stomach it could react with amines from the breakdown of proteins to form N-nitrosamines, which are known to be carcinogenic. One early epidemiological study appeared to support this as it showed a higher incidence of stomach cancer in an area of the UK having high concentrations of nitrate in drinking water. However, a later and much larger study, published in the mid-1980s, contradicted this result. Nitrate concentration in saliva was measured in people from two areas in the UK, one with an unusually high incidence of stomach cancer and one with an unusually low incidence. The results were the opposite of what would be predicted if nitrate was associated with stomach cancer – nitrate concentrations in saliva were about 50% higher in the low-cancer-risk area. Such evidence in itself can neither prove nor disprove a link but it clearly indicates that any link must be weak and outweighed by other factors. Several other studies have shown a negative correlation between nitrate and incidence of stomach cancer. One study was based on the number of deaths of workers in a factory manufacturing ammonium nitrate fertilizer over a 35-year period. Because of the ammonium nitrate dust in the factory, the workers had a larger intake of nitrate than the general population and this was shown by increased concentrations of nitrate in their saliva. However, the number of deaths from stomach cancer among these workers between 1946 and 1981 was exactly the number expected on the basis of incidence in the general population. It can be concluded that there is now no medical evidence to suggest any link between nitrate intake and stomach cancer.

### Methemoglobinemia (Blue-Baby Syndrome)

Methemoglobinemia or blue-baby syndrome is a serious disease that can kill infants less than 1 year old. Nitrite derived from the reduction of nitrate can react with hemoglobin and block the site that normally transports oxygen around the body. The first report of the problem in the USA in 1945 referred to 'well-water methemoglobinemia,' and all the cases identified in a 1991 review were caused by water from wells, 98% of which were described as 'privately dug.' It is now thought that the condition only occurs when water is contaminated by gastroenteritic bacteria from sewage or manure heaps, as can occur if a shallow well is dug on farm premises. In such cases nitrate and bacteria both contaminate the water. The condition is currently known in parts of Africa where surface waters are used for human consumption but are contaminated by animal excreta, introducing both bacteria and nitrate. The publisher of the original report was well aware that both bacteria and nitrate were involved in the condition, but the possible role of the bacteria seems to have been played down subsequently. It is now seems likely that it is the bacterial contamination that is the main factor. There are no reports of blue-baby syndrome associated with bacteriologically clean water from a mains supply.

### Beneficial Effects of Nitrate on Health

Recent medical research has indicated that nitrate can protect humans and animals from gastroenteritis. It has shown that nitrate is reduced to nitrite by denitrifying bacteria in the mouth, similar to those in soil. It also now appears that the human metabolic system actively secretes nitrate into saliva for this purpose if the intake of nitrate from water or vegetables is insufficient. The proposed reaction between nitrite and amines in the stomach (mentioned above in relation to the proposed link with stomach cancer) probably does not occur because the acid conditions in the stomach destroy nitrite and nitric oxide, NO, is formed. A great deal is now known about the physiological role of NO, one of which is the killing of bacteria such as *Salmonella* and *Escherichia coli* which can cause gastroenteritis. In this way it is now thought that nitrate is a key precursor for NO and thus plays an important role in the body's defense against stomach disease.

## Impacts of Nitrate on Ecology of Surface Waters

Any nitrate escaping from agricultural soil to streams, rivers, or lakes can increase the growth of water plants and thus alter the ecology of aquatic systems. This is undesirable in itself and can also interfere with fisheries and navigation for shipping by clogging the waterway with large quantities of water weed. Nitrate can also contribute to the growth of algae in surface waters which are unsightly and some of which are toxic. When algae die the bacteria decomposing them use oxygen dissolved in the water, thus deoxygenating the water with adverse consequences for fish and other organisms. In freshwaters phosphate is usually the limiting factor controlling the growth of algae, but nitrate is also required. In river estuaries and sheltered seas (e.g., the Baltic Sea and the Gulf of Mexico), nitrate is probably the controlling factor. Thus the impact of nitrate on the ecology of natural waters is a genuine and significant reason for controlling losses from agriculture.

## Conclusions

Nitrate is the most common form of inorganic nitrogen in soil, being the final product of the decomposition of organic materials under aerobic conditions. It is the form normally taken up by plants, though most plants can also take up ammonium. If sufficient food is to be grown for the current world population of over 6 billion and its expected growth during the twenty-first century, greatly increased amounts of nitrogen will be needed. It will be necessary to use increasing amounts of inorganic fertilizers in addition to better utilization of organic manures and other wastes and improvement in the use of legumes to supply nitrogen to crops. This continued intensification of agriculture will inevitably lead to further losses of nitrate to natural waters. The considerable amount of research conducted over recent decades can help to decrease these losses, both by increasing the efficiency of use of inorganic fertilizers and by better retaining nitrate from organic matter mineralization within the soil/plant system. However, losses can never be entirely eliminated as the nitrate ion is soluble and is always at some risk of being washed out of soil. It can also be converted to nitrous oxide, a powerful greenhouse gas. The ecological impacts of nitrate on aquatic ecosystems can be serious so, for a range of environmental reasons, it is important to minimize nitrate losses from agricultural soils as much as possible. This is in addition to the economic and sustainability reasons for avoiding waste of a vital resource. However, the proposed adverse health impacts of nitrate in water that have been the main driver for limiting nitrate loss now appear to be unfounded. Indeed, the opposite appears to be the case – there is evidence from medical research that some intake of nitrate has beneficial effects on health through the production of nitric oxide in the stomach and its role in the control of gastroenteritis.

*See also:* **Eutrophication**; **Greenhouse Gas Emissions**; **Isotopes in Soil and Plant Investigations**; **Nitrogen in Soils:** Cycle; Nitrification; Plant Uptake; Symbiotic Fixation; **Pollution:** Groundwater

## Further Reading

Addiscott TM, Whitmore AP, and Powlson DS (1991) *Farming, Fertilizers and the Nitrate Problem.* Wallingford: CAB International.

Benjamin N (2000) Nitrates in the human diet – good or bad? *Annales de Zootechnologie* 49: 207–216.

Catt JA *et al.* (1998) Strategies to decrease nitrate leaching in the Brimstone Farm experiment, Oxfordshire, UK, 1988–1993: the effects of winter cover crops and unfertilized grass leys. *Plant and Soil* 203: 57–69.

Chency K (1990) Effect of nitrogen fertilizer rate on soil nitrate nitrogen content after harvesting winter wheat. *Journal of Agricultural Science, Cambridge* 114: 171–176.

Davidson EA (1991) Fluxes of nitrous oxide and nitric oxide from terrestrial ecosystems. In: Rogers JE and Whitman WB (eds) *Microbial Production and Consumption of Greenhouse Gases: Methane, Nitrogen Oxides, Halomethanes*, pp. 219–235. Washington, DC: American Society for Microbiology.

Dykhuizen RS, Fraser A, Duncan C, Golden M, Benjamin N, and Leifert C (1999) Antimicrobial effect of acidified nitrite on gut pathogens: the importance of dietary nitrate in host defences. In: Wilson WS, Ball AS, and Hinton RH (eds) *Managing Risks of Nitrates to Humans and the Environment*, pp. 295–316. Cambridge: Royal Society of Chemistry.

Goss MJ, Howse KR, Christian DG, Catt JA, and Pepper TJ (1998) Nitrate leaching: modifying the loss from mineralized organic matter. *European Journal of Soil Science* 49: 649–659.

Goulding K (2000) Nitrate leaching from arable and horticultural land. *Soil Use and Management* 16: 145–151.

Goulding KWT, Poulton PR, Webster CP, and Howe MT (2000) Nitrate leaching from the Broadbalk wheat experiment, Rothamsted, UK, as influenced by fertilizer and manure inputs and weather. *Soil Use and Management* 16: 244–250.

Hatch D, Goulding K, and Murphy D (2002) Nitrogen. In: Haygarth PM and Jarvis SC (eds) *Agriculture, Hydrology and Water Quality*, pp. 7–27. Wallingford: CAB International.

Jarvis SC (2000) Progress in studies of nitrate leaching from grassland. *Soil Use and Management* 16: 152–156.

Macdonald AL, Powlson DS, Poulton PR, and Jenkinson DS (1989) Unused fertilizer nitrogen in arable soils – its contribution to nitrate leaching. *Journal of the Science of Food and Agriculture* 46: 407–419.

Macdonald AJ, Poulton PR, Powlson DS, and Jenkinson DS (1997) Effects of season, soil type and cropping on recoveries, residues and losses of 15N-labelled fertilizer applied to arable crops in spring. *Journal of Agricultural Science, Cambridge* 129: 125–154.

Powlson DS (1993) Understanding the soil nitrogen cycle. *Soil Use and Management* 9: 86–94.

Powlson DS (1999) Fate of nitrogen from manufactured fertilizers in agriculture. In: Wilson WS, Ball AS, and Hinton RH (eds) *Managing Risks of Nitrates to Humans and the Environment*, pp. 42–57. Cambridge: Royal Society of Chemistry.

Powlson DS (1997) Integrating agricultural nutrient management with environmental objectives – current state and future prospects. *Proceedings No. 402.* York: The Fertiliser Society.

Powlson DS, Hart PBS, Poulton PR, Johnston AE, and Jenkinson DS (1986) Recovery of $^{15}$N-labelled fertilizer applied in autumn to winter wheat at four sites in eastern England. *Journal of Agricultural Science, Cambridge* 107: 611–620.

Recous S, Fresnau C, Faurie G, and Mary B (1988) The fate of labelled $^{15}$N urea and ammonium nitrate applied to a winter wheat crop. *Plant and Soil* 112: 205–214.

Wilson WS, Ball AS, and Hinton RH (eds) (1999) *Managing Risks of Nitrates to Humans and the Environment.* Cambridge: Royal Society of Chemistry.

# Nitrification

**J I Prosser**, University of Aberdeen, Aberdeen, UK

## Introduction

Nitrification is the oxidation of reduced forms of nitrogen, ultimately to nitrate. It is carried out by three microbial groups: (1) autotrophic ammonia oxidizers, (2) autotrophic nitrite oxidizers, and (3) heterotrophic nitrifiers. Autotrophic ammonia and nitrite oxidizers are characterized by their ability to oxidize ammonia sequentially to nitrate under aerobic conditions, but these organisms possess a wider metabolic diversity. This has significant consequences for their ecology, for the occurrence and kinetics of nitrification, and for interactions with other organisms. The major sources of reduced forms of soil nitrogen are excretion and decomposition of organic nitrogen, derived from animals and plants, and application of ammonia-based fertilizers. Many microorganisms and plants require ammonium for growth, while others assimilate nitrate. For both groups, nitrification is important in regulating the supply or loss of nitrogen from the environment. While ammonia has a propensity to bind to soil particles, its conversion to nitrate leads to significant losses of soil nitrogen through leaching and conversion to gaseous forms through denitrification. Nitrification also contributes to soil acidification, with consequent increases in mobilization of toxic metals, particularly in heavily fertilized and poorly buffered soils. In addition, nitrifiers are involved in production and conversion of a number of greenhouse gases.

## Autotrophic Ammonia-Oxidizing Bacteria

### Taxonomy

Autotrophic ammonia oxidizers are Gram-negative bacteria, traditionally placed within the *Nitrobacteriaceae* and characterized by their ability to oxidize ammonia. Further classification is based on a limited number of characteristics, including cell shape, internal membrane structures, flagella, inclusion bodies, salt tolerance, urease activity, and substrate affinity and inhibition. Confirmation of taxonomic groupings and species descriptions is by DNA:DNA hybridization of pure cultures. The type strain, *Nitrosomonas europaea*, has been the subject of the majority of biochemical, physiological, and genomic studies.

More recently, analysis of sequences of 16S ribosomal ribonucleic acid (rRNA) and *amo* A genes (encoding the active subunit of ammonia monooxygenase) has enabled phylogenetic classification of cultured ammonia oxidizers. This places *Nitrosococcus oceanus* in the $\gamma$-proteobacteria and other known autotrophic ammonia oxidizers in seven lineages within a monophyletic group within the $\beta$-proteobacteria: *Nitrosospira*, *Nitrosomonas europaea/Nitrosococcus mobilis*, *Nitrosomonas communis*, *Nitrosomonas marina*, *Nitrosomonas oligotropha*, *Nitrosomonas cryotolerans*, and *Nitrosomonas* sp. Nm143. Difficulties in cultivating ammonia oxidizers limit available physiological data, but links exist between these phylogenetic groups, their physiological characteristics, and their environmental origin. Analysis of 16S rRNA gene sequences also indicates evolutionary relationships with photosynthetic bacteria and other groups possessing intracytoplasmic membrane structures, suggesting that ammonia oxidizers evolved from photosynthetic organisms. 16S rRNA sequence analysis suggests only peripheral associations with methane oxidizers, but similarities exist between *amo*A genes of $\gamma$-proteobacterial ammonia oxidizers (*N. oceanus*), and *pMMO* (particulate methane monooxygenase), suggesting a common ancestral origin.

### Biochemistry

Ammonia oxidation by $\beta$-proteobacteria involves two steps:

$$NH_3 + O_2 + 2H^+ + 2e^- \rightarrow NH_2OH + H_2O$$
$$NH_2OH + H_2O \rightarrow NO_2^- + 5H^+ + 4e^-$$

In the first step, ammonia is oxidized to hydroxylamine by a membrane-bound ammonia monooxygenase. The reaction is endergonic and requires oxygen and a source of reducing equivalents. Hydroxylamine is then oxidized to nitrite by hydroxylamine oxidoreductase, using oxygen from water in an exergonic reaction. Two electrons are required for ammonia oxidation, while the remainder pass down the electron transport chain. The $NO_2^-/NH_3$ redox potential is $-340$ mV and reduced nicotinamide adenine dinucleotide (NADH) is synthesized by reverse electron flow using adenosine triphosphate (ATP). Hydrogen ions generated by the second step lead to acidification of unbuffered growth medium, limiting sustained growth.

An alternative mechanism for the first step has recently been proposed, involving nitrogen dioxide and tetroxide, based on studies of *N. eutropha*:

$$NH_3 + N_2O_4 + 2H^+ + 2e^- \rightarrow NH_2OH + 2NO + H_2O$$
$$2NO + O_2 \rightarrow 2NO_2(N_2O_4)$$

This mechanism does not appear to involve ammonia monooxygenase, and studies have also demonstrated reduction in lag phase and the stimulation of growth, activity, and cell yield in cultures supplied with nitric oxide. Any reduction in lag phase will provide ammonia oxidizers with advantages in soil where ammonia supply is intermittent.

### Carbon Metabolism

Organic carbon is obtained by nitrifiers via carbon dioxide fixation, and all strains studied as of 2003 possess type I Rubisco genes, but with sequence variation between strains. Rubisco enzymes in some strains are located in carboxysomes. Energy generated by ammonia oxidation is only just sufficient for growth, particularly as carbon dioxide fixation requires significant reducing equivalents, and generation of reducing equivalents consumes ATP.

Ammonia monooxygenase has broad substrate specificity, and ammonia oxidizers can oxidize and assimilate, but not grow on, a number of organic compounds, including methanol, bromoethane, ethylene, propylene, cyclohexane, benzene, and phenol. Acetate, formate, and pyruvate can increase biomass yield by as much as 130%, but higher concentrations inhibit growth. The ecological implications for this are not clear, but assimilation of soil organic carbon may provide a selective advantage, while ammonia oxidizers may be important in biodegradation of organic pollutants in oligotrophic environments, where alternative carbon sources are scarce.

Ammonia monooxygenase is similar to the membrane-bound methane monooxygenase of methanotrophic bacteria. Ammonia oxidizers can oxidize but not grow on methane and carbon monoxide, and the enzymes have common inhibitors. Incorporation of methane is inhibited by ammonia, and vice versa, and complete methane oxidation may be prevented by inhibition of hydroxylamine oxidation by formaldehyde, produced during methane oxidation. In general, ammonia oxidizers have higher oxidation rates and lower $K_m$ values for ammonia than methanotrophs, while the converse is true for oxidation and affinity for methane. Competition between the two groups for ammonia and methane therefore appears unlikely, but high levels of ammonia-based fertilizer may suppress methanotroph activity. Despite these considerations, ammonia oxidizer 16S rRNA and *amoA* sequences have been found in $^{13}$C-labeled nucleic acids extracted from grassland soils treated with $^{13}$C-labeled methanol and methane, suggesting their role in soil methane oxidation. Ammonia oxidizers may also have a role in the oxidation of another greenhouse gas, carbon monoxide. They have greater affinity for carbon monoxide than carboxydobacteria, and may provide an important carbon monoxide sink. Carbon monoxide oxidation only occurs during ammonia oxidation, which is required as an energy source, even though carbon monoxide and ammonia compete for the ammonia monooxygenase enzyme.

### Ammonia Oxidation under Anaerobic Conditions

Although traditionally considered to be aerobic, autotrophic ammonia oxidizers maintain activity at low oxygen concentrations and under anaerobic conditions. *N. europaea* possesses a nitrite reductase and can denitrify nitrite under anaerobic conditions using hydrogen or organic compounds as electron donors. (They can also denitrify under oxic conditions, using ammonia as electron donor.) This results in the production of nitric and nitrous oxides, and dinitrogen gas. At atmospheric oxygen concentrations, nitrogen oxides constitute 0.1–0.5% of the total N oxidized, but proportions rise to 2.5–10% at low oxygen concentrations.

Anaerobic ammonia oxidation by *N. eutropha* has recently been demonstrated in which nitrogen dioxide or nitrogen tetroxide acts as an electron acceptor, generating nitric oxide and hydroxylamine. The latter is converted to nitrite with release of dinitrogen gas. In addition, anaerobic ammonia oxidation is carried out by autotrophic planctomycetes, which utilize nitrite as an electron acceptor, generating nitrate and nitrogen, with hydrazine and hydroxylamine as intermediates. The process is sensitive to oxygen and nitrite, and the calculated mass balance is:

$$\begin{aligned}\mathrm{NH_4^+} &+ 1.32\mathrm{NO_2^-} + 0.066\mathrm{HCO_3^-} + 0.13\mathrm{H^+} \rightarrow 0.26\mathrm{NO_3^-}\\ &+ 1.02\mathrm{N_2} + 0.066\mathrm{CH_2O} + 0.5\mathrm{N_{0.15}} + 2.03\mathrm{H_2O}\end{aligned}$$

However, neither of these anaerobic processes has been detected in soil, where their significance is unclear. Both require substrates which are produced under oxic conditions (nitrogen dioxide and nitrite) but the processes may occur at oxic–anoxic interfaces, such as in biofilms formed on soil particulate material.

## Autotrophic Nitrite-Oxidizing Bacteria

Nitrite oxidizers are classified into four genera within the proteobacteria: *Nitrobacter* (the most studied; $\alpha$-proteobacteria), *Nitrococcus* ($\gamma$-proteobacteria), *Nitrospira* ($\delta$-proteobacteria), and *Nitrospina* (forming its own subdivision). *Nitrobacter* strains are related to several denitrifying organisms, while *Nitrobacter* and *Nitrococcus* possess intracytoplasmic membranes and are related to photosynthetic bacteria. $\delta$-Proteobacterial nitrite oxidizers do not contain intracytoplasmic membranes and are not related to photosynthetic organisms. Only *Nitrobacter* and *Nitrosospira* have been detected in soil. Nitrite oxidation is catalyzed by nitrite oxidoreductase, the $\mathrm{NO_2^-/NO_3^-}$ redox potential is 430 mV, and reducing equivalents are generated by reverse electron flow, consuming ATP:

$$\mathrm{NO_2^- + H_2O \rightarrow NO_3^- + 2H^+ + 2e^-}$$

Nitrite oxidizers fix $\mathrm{CO_2}$ for growth but, unlike ammonia oxidizers, many are capable of mixotrophic and heterotrophic growth on a range of simple organic substrates. Heterotrophic growth is slow (doubling times of 30–150 h) but leads to greater biomass yields. Nitrite oxidoreductase enables heterotrophic growth under anaerobic conditions, with nitrate as the electron acceptor and production of nitrite and ammonia, and nitrogen oxides. This metabolic versatility may explain greater abundance of nitrite oxidizers than ammonia oxidizers in soil.

## Heterotrophic Nitrification

Heterotrophic nitrification is the oxidation of inorganic and organic reduced forms of N, to nitrate, by a wide range of fungi and heterotrophic bacteria. In some organisms, the mechanism is similar to that in autotrophic ammonia oxidizers and is linked, in some strains, to aerobic denitrification. The second mechanism, termed fungal nitrification, is linked to lignin degradation and involves reaction of reduced organic compounds with hydroxyl radicals produced in the presence of hydrogen peroxide and superoxide. There is little evidence that heterotrophs gain energy or other benefit from nitrification, and cellular rates of heterotrophic nitrification activity are significantly lower than for autotrophs. Nevertheless, it may be important in acid soils or where C:N ratios and heterotroph biomass are high.

## Analysis of Nitrifiers and Nitrification

### Nitrification Activity and Rates

Nitrification rates in soil are traditionally measured as ammonia consumption or nitrite plus nitrate production during incubation of soil or soil slurry. Zero-order kinetics result from short-term measurements at nonlimiting ammonia concentrations and measure potential nitrification, which can be used to estimate nitrifier abundance. Short-term measurements at limiting substrate concentrations generate Michaelis–Menten kinetics. Long-term incubation of nonenriched soil leads to exponential increases in nitrite plus nitrate owing to unrestricted growth, at nonlimiting substrate concentrations, or growth at submaximal and decreasing rates at limiting substrate concentrations. Nitrate-production kinetics are frequently fitted to the logistic equation. Kinetics can be interpreted differently, however, if nitrifiers are considered to be distributed heterogeneously, as clusters growing on discrete sources of ammonium within the soil. If nitrification is measured in individual soil aggregates, the observed nitrate production curves fit well with the cumulative distribution curves for lag periods. Bulk nitrate concentrations therefore reflect not cumulative product of a uniformly growing population but the continued activation of clusters of nitrifying bacteria. The increase in nitrate concentration then depends on the distribution of lag periods prior to the onset of nitrification. Further evidence for heterogeneous distribution of ammonia oxidizers in soil is provided by molecular analysis of communities in grassland soils. Regular application of inorganic nitrogen fertilizer reduces heterogeneity in ammonia oxidizer communities, in association with similar reduction in the heterogeneity of soil pH and ammonia concentration. Analysis of nitrification kinetics during incubation is therefore important in providing information on soil population levels and substrate concentrations. Growth parameters are more easily obtained from laboratory cultures and typical values are given in **Table 1**.

Other nitrogen cycle processes, e.g., denitrification, immobilization, and ammonification, will influence concentrations of ammonium, nitrite, and nitrate. Rigorous measurement of soil nitrification rates, and distinction between autotrophic and heterotrophic nitrification therefore necessitate the use of metabolic blocks and $^{15}N$-based techniques. The former involve measurement of nitrate production in the presence of specific nitrification inhibitors e.g., nitrapyrin and acetylene. The second involves applying $^{15}N$ as $NH_4^+$ or $NO_3^-$ and following changes in the concentration of label in the respective pools.

### Enumeration

Dilution-plate enumeration of viable nitrifiers is not feasible because of poor growth on solid media and enumeration is traditionally achieved using the most probable number (MPN) method. This involves multiple inoculation from dilutions of a soil suspension into mineral-salts medium containing ammonium and a pH indicator. After incubation for several (typically 4) weeks, growth is assessed by nitrite production, ammonium loss, and/or acidification, and cell concentrations are calculated using statistical tables. The requirement for laboratory cultivation leads to underestimation of cell concentrations. In addition, growth may be biased by ammonium concentrations in the medium, with inhibition of some strains at high concentrations (**Figure 1**). MPN counts for typical soils range between $10^3$ and $10^5$ cells $g^{-1}$

**Table 1** Reported ranges for kinetic constants derived from pure cultures of ammonia and nitrite oxidizers

| | *Ammonia oxidizers* | *Nitrite oxidizers* |
|---|---|---|
| Maximum specific growth rate ($h^{-1}$) | 0.02–0.088 | 0.018–0.058 |
| Doubling time (h) | 8–35 | 12–39 |
| Biomass yield (g biomass $mol^{-1}$ substrate) | 0.42–1.72 | 1.11–1.51 |
| Cell yield (cells $mol^{-1}$ substrate) | 1.85–10.6 | |
| Carbon yield (ratio of $CO_2$ fixed: $NO_2^-$ or $NO_3^-$ produced) | 0.014–0.096 | 0.013–0.031 |
| Cell activity (fmol $NO_2^-$ or $NO_3^-$ $cell^{-1}$ $h^{-1}$) | 0.9–31.3 | 5.1–42 |
| Biomass activity (fmol $NO_2^-$ or $NO_3^-$ g $biomass^{-1}$ $h^{-1}$) | 4–200 | 15.7–25.2 |
| Saturation constant for growth (mmol $l^{-1}$ $NH_4^+$ or $NO_2^-$) | 0.12–14 | 0.045–0.178 |
| Saturation constant for activity (mmol $l^{-1}$ $NH_4^+$ or $NO_2^-$) | 0.05–0.07 | 1.6–3.6 |

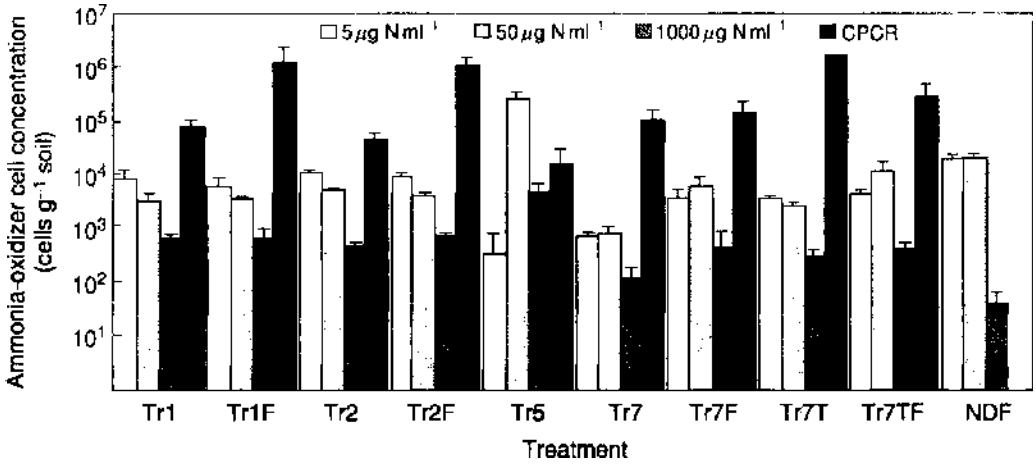


**Figure 1** Enumeration of ammonia-oxidizer cells in long-term ecological research (LTER) soils using the most probable number (MPN) method with medium containing 5, 50, or 1000 mg of $NH_4^+$-N ml$^{-1}$ and competitive polymerase chain reaction (C-PCR). Treatments are: Tr1, conventional tilling; Tr2, no tilling; Tr5, *Populus* perennial cover crop; Tr7, historically tilled; NDF, natural deciduous forest. Error bars represent the standard errors. Suffixes 'T' and 'F' indicate tillage and fertilization, respectively. (Reproduced with permission from Phillips CJ, Harris D, Dollhopf SL *et al.* (2000) Effects of agronomic treatments on the structure and function of ammonia oxidising communities. *Applied and Environmental Microbiology* 66: 5410–5418.)

soil for ammonia oxidizers, with nitrite oxidizers slightly more abundant. Quantitative molecular techniques (e.g., competitive PCR (cPCR), real-time PCR) avoid cultivation bias and detect the total nitrifier populations, giving cell concentrations 1–2 orders of magnitude greater than MPN counts (**Figure 1**).

## Enrichment and Isolation

Autotrophic ammonia-oxidizing bacteria are readily enriched by inoculation of an inorganic medium, containing ammonium with soil, and incubation for several weeks. Isolation of pure cultures is more difficult, due to faster growth of heterotrophs on organic carbon contaminants or by-products of nitrifier growth. This typically takes several months and cultures frequently do not survive continued subculturing. Laboratory selection also reduces the likelihood of obtaining dominant or representative environmental strains, but enrichment of a strain is necessary for unequivocal evidence of its presence and viability in a particular soil.

## Soil Communities and Diversity

Information on 16S rRNA and *amo*A gene sequences from cultured ammonia oxidizers has led to the design of oligonucleotide polymerase chain reaction (PCR) primers specific for β-proteobacterial autotrophic ammonia oxidizers. These have been used to amplify ammonia-oxidizer gene fragments from nucleic acids extracted directly from soil. Gene fragment sequences may then be analyzed by cloning, sequencing of clone library representatives, and phylogenetic analysis. Alternatively, fingerprinting techniques such as denaturing or temperature gradient gel electrophoresis (DGGE, TGGE) and terminal restriction fragment length polymorphism (T-RFLP) analysis may be used. These approaches do not require laboratory cultivation and thereby eliminate many of the disadvantages of cultivation-based techniques, and difficulties in purifying and identifying ammonia-oxidizing bacteria.

The application of molecular techniques has greatly expanded our knowledge of nitrifier communities in soils. Phylogenetic analysis of amplified sequences reveals considerable diversity within soil ammonia oxidizers. Duplicate sequences and sequences identical to pure cultures are rarely obtained, although this partly reflects the few pure culture sequences in databases and, potentially, sequencing and cloning errors. Sequences are sometimes homologous to those of laboratory cultures enriched from the same environment and the majority fall within the lineages obtained with sequences from cultivated organisms. One exception is the possible subdivision of *Nitrosospira* into four clusters, and there is evidence of differences in physiological characteristics of pure and enriched cultures representative of these different *Nitrosospira* clusters.

Soils are dominated by β-proteobacterial ammonia oxidizers, with γ-proteobacterial sequences apparently restricted to marine environments. The links between 16S rRNA-defined lineages and physiological characteristics aid in the interpretation of the distribution of these organisms under different soil conditions. This approach also indicates that *Nitrosospira* may be more abundant than *Nitrosomonas* in soil. If so, more studies are required of *Nitrosospira* physiology, rather than basing interpretation on information on *Nitrosomonas europaea*.

Fingerprinting techniques, particularly DGGE, have been used to study the effects of environmental factors and soil treatments on ammonia-oxidizer communities. This provides more rapid assessment and greater replication than cloning and/or sequencing approaches. In addition, bands of interest on DGGE gels may be excised, cloned, and sequenced and information on identity obtained by comparison with database sequences. These approaches can also be used to assess quantitative or semiquantitative changes in ammonia-oxidizer community structure.

Comparison of 16S rRNA gene sequences in enrichment cultures with those amplified from nucleic acids extracted from the same soil provides an indication of selection during laboratory cultivation. The relative abundances obtained using these two approaches differ but, in one study, 16% of sequences were common to both enrichment cultures and clones analyzed. Molecular analysis of cultures obtained during MPN enumeration also indicates selection, with frequent domination by *Nitrosomonas* strains in samples from environments which appear, from molecular analysis, to be dominated by *Nitrosospira*. Molecular- and cultivation-based techniques therefore provide different views of diversity and species composition within soil ammonia-oxidizer communities. However, molecular techniques are also subject to bias and additional techniques are required. In this respect, direct detection of cells using fluorescent *in situ* hybridization (FISH) with specific ammonia-oxidizer probes would be useful. This approach, however, is limited by autofluorescence and the relatively low cell concentrations of ammonia oxidizers in soil.

## Influence of Soil Characteristics on Ammonia-Oxidizer Communities

### Soil pH

Ionization of ammonia at low pH values reduces substrate availability for ammonia oxidizers, and laboratory batch growth of ammonia oxidizers does not occur at pH values of less than 6.5. However, autotrophic ammonia oxidizers with neutral pH optima are frequently isolated from acid soils, and autotrophic nitrification has been demonstrated in soils with pH values as low as 3.5. In some situations, nitrification in acid soils is due to heterotrophic nitrification. Coupled denitrification–nitrification by heterotrophic nitrifiers is unlikely, due to the requirement for high substrate concentrations, and heterotrophic nitrification involving ammonia monooxygenase-based reactions suffers from reduced ammonia availability at low pH. However, acidophilic fungal heterotrophic nitrifiers have been characterized and fungal nitrification is significant in some acid soils. Nevertheless, use of specific inhibitors of autotrophic ammonia oxidation and $^{15}$N-based studies have demonstrated active, autotrophic ammonia oxidation in acid soils.

One explanation for this phenomenon is the possession by some ammonia oxidizers of urease activity, enabling growth on urea as a sole nitrogen source. Urease activity and growth on urea are independent of pH in the range 4–8, possibly through uptake by passive diffusion. Laboratory studies show that, under these conditions, ammonia oxidation can proceed at pH values as low as 4, until urea is exhausted, after which it ceases, even if ammonium is present in solution. This has been demonstrated in a strain of *Nitrosospira* that cannot oxidize ammonium at pH values less than 7 (**Figure 2**). Urease activity therefore provides

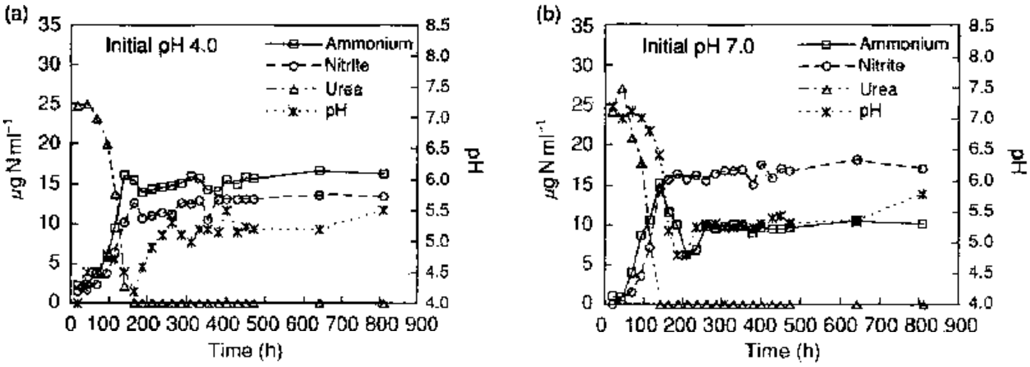


**Figure 2** Changes in pH and in concentrations of urea, ammonium, and nitrite during growth of *Nitrosospira* strain NPAV in liquid batch culture on poorly buffered medium containing urea, at initial pH values of (a) 4 and (b) 7. (Reproduced with permission from Burton SAQ and Prosser JI (2001) Autotrophic ammonia oxidation at low pH through urea hydrolysis. *Applied and Environmental Microbiology* 67: 2952–2957.)

the potential for acidophilic ammonia oxidation, but is not a prerequisite for ammonia oxidizers isolated from acid soils.

Autotrophic ammonia oxidizers appear to be protected from low pH in soil when occurring at high cell concentrations and associated with particulate matter. High cell concentrations can be achieved in aggregates, which when inoculated into liquid medium can nitrify at low pH, and nitrification at pH 4 has been demonstrated for cells immobilized in alginate. Biofilm formation on surfaces such as glass slides, glass beads, ion-exchange resin beads, and clay minerals can reduce the pH minimum for growth and activity of *N. europaea* by up to 2 pH units. Surface growth protection increases with the cation exchange capacity of the particulate material and lag phases prior to growth of stationary-phase or starved cells are reduced for attached cells. Protection from low pH may result from preferential growth and nitrification at the clay surface, due to adsorption of ammonium. It is greatest when the only source of ammonium is that bound to clay minerals. This has been demonstrated by growth on ammonia-treated vermiculite (ATV), in which ammonia is fixed to the vermiculite by high-temperature exposure to ammonia. In the presence of ATV, *N. europaea* growth occurs without a lag phase and with greater nitrite yield than for suspended cells, but with no decrease in the pH of the medium. This, and the other clay effects, suggests localized buffering; ammonium released from the clay surface is utilized by adsorbed cells and $H^+$ ions are effectively inactivated by exchange for ammonium ions. It is not clear why biofilm growth should reduce the lag phase, but this presents a significant ecologic advantage in competing for transient supplies of ammonium. A possible mechanism involves cell–cell signaling molecules, acyl homoserine lactones, which will accumulate at high cell concentrations. When *N. europaea* biofilms established on glass beads are starved and then resupplied with ammonia, oxidation occurs immediately. However, supply of ammonia to suspended, nonbiofilm cells in liquid medium results in a lag period, prior to ammonia oxidation. The lag period increases with the length of the preceding starvation period but is significantly reduced by addition of acyl homoserine lactones. Signaling compounds are produced by a wide range of soil microorganisms and affect a number of ecologically important processes. These compounds may therefore mediate interactions between nitrifiers and other soil microorganisms, particularly in biofilms and in the rhizosphere, where biomass concentrations are sufficient to lead to high concentrations of signaling molecules.

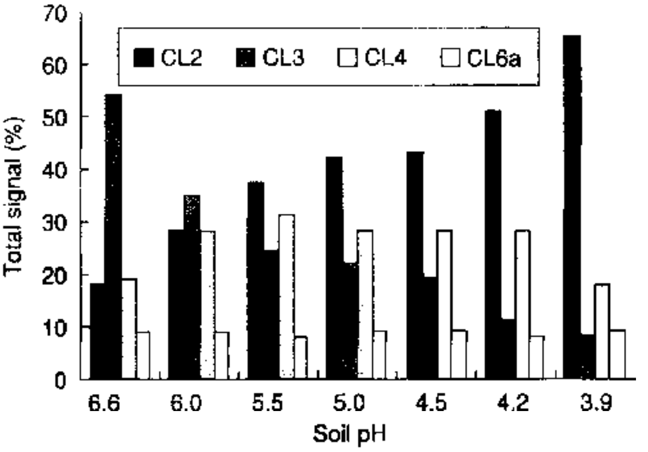


**Figure 3** The influence of soil pH on the relative abundances of *Nitrosospira* clusters 2 (CL2), 3 (CL3), and 4 (CL4), and *Nitrosomonas* cluster 6a. Abundances are calculated as relative intensities of denaturing gradient gel electrophoresis (DGGE) analysis of ribosomal ribonucleic acid (rRNA) genes amplified from DNA extracted from soil maintained at a range of pH values, using ammonia oxidizer-specific primers. (Reproduced with permission from Stephen JR, Kowalchuk GA, Bruns MA *et al.* (1998) Analysis of $\beta$-subgroup ammonia oxidiser populations in soil by DGGE analysis and hierarchical phylogenetic probing. *Applied and Environmental Microbiology* 64: 2958–2965.)

The development of molecular techniques has enabled assessment of the influence of soil pH on ammonia-oxidizer communities. In one study, DGGE analysis was carried out on $\beta$-proteobacterial 16S rRNA gene fragments amplified from DNA extracted from soils maintained for 36 years at pH values in the range 3.9–6.6. Relative abundances of different clusters were assessed as relative intensities of DGGE bands and showed a greater proportion of *Nitrosospira* cluster-2 sequences in acid soils and a greater proportion of *Nitrosospira* cluster 3 at higher pH (**Figure 3**). Other clusters detected were *Nitrosospira* cluster 4, which may have increased in relative abundance at pH 5.5, and *Nitrosomonas* cluster 6a, which was independent of pH. Molecular studies of agricultural, grassland, and forest soils provide further evidence for the association of *Nitrosospira* cluster-2 sequences with acid soils. However, lack of information on distinctive physiological characteristics specific to representatives of this group, and of physiological diversity within ammonia oxidizers in general, prevents establishment of links between species diversity, functional diversity, and mechanisms for nitrification at low pH.

## Nitrogen Fertilization

Sustained nitrogen fertilization and liming increase nitrifier population size, as indicated by cell counts (MPN and cPCR) and nitrification potential measurements. In addition, evidence from molecular studies of nitrifiers in long-term ecological research (LTER) sites subjected to a number of different fertilization and cultivation conditions indicates an influence on the ammonia-oxidizer communities. Tilled soils were dominated by *Nitrosospira* cluster-3 sequences, while never-tilled soils contained *Nitrosospira* clusters 3 and 4, and *Nitrosomonas* cluster 6. Similar effects of fertilization on the relative proportions of *Nitrosospira* cluster-3 and -4 sequences have been found in chalk grassland soil. Little is known about the persistence of changes after perturbations.

Nitrogen supplied as inorganic fertilizer or, in grazing systems, as animal urine leads to localized, high ammonia concentrations. These concentrations are sufficient to inhibit ammonia-sensitive groups of ammonia oxidizers, while permitting growth of more tolerant strains. Soil microcosm experiments have demonstrated that the kinetics of nitrification of ammonia derived from synthetic sheep urine applied to grassland soil depends on the relative abundances of ammonia-sensitive and ammonia-tolerant strains. Low numbers of the latter result in significant lag periods before nitrification, impact significantly on maintenance of high levels of ammonia for plant growth, and reduce nitrogen losses associated with leaching and denitrification of nitrate.

### Inhibition of Soil Nitrification

Specific inhibitors of autotrophic nitrification (e.g., nitrapyrin, dicyandiamide, etridiazole) have been used in the field to reduce losses of fertilizer nitrogen and to reduce nitrate pollution. Inhibitors, particularly acetylene, are also used as metabolic blocks when measuring nitrification rates. Higher concentrations of these compounds are required for inhibition in soil than in laboratory culture because of inactivation and degradation and because attachment of nitrifiers to soil surfaces protects cells from inhibition. Protection is greater with expanding clays with high cation exchange capacity. In many mature coniferous soils, nitrogen is present as ammonium rather than nitrate and net nitrification rates are low. This has led to suggestions of allelopathic inhibition of nitrification by polyphenolic compounds released by roots, which can inhibit laboratory growth of nitrifiers. However, use of $^{15}$N-based techniques has demonstrated high gross nitrification rates in these soils and low nitrate levels are more likely due to greater immobilization of nitrate. There is little direct evidence for specific effects of plant diversity and community structure on ammonia-oxidizer communities, although plants can exert indirect effects through their influence on soil ammonium concentrations.

## Summary

Nitrifying bacteria convert the most reduced form of soil nitrogen, ammonia, into its most oxidized form, nitrate. In itself, this is important for soil ecosystem function, in controlling losses of soil nitrogen through leaching and denitrification of nitrate. Nitrifiers also contribute to other important processes, including nitrous oxide production, methane oxidation, degradation of organic compounds, and carbon monoxide oxidation. The development of $^{15}$N-based techniques has increased significantly our ability to dissect soil nitrogen transformations and their rates, while molecular techniques now enable characterization of soil nitrifier community structure and changes in species composition. These approaches are increasing our understanding of the ecology of soil-nitrifying bacteria and provide the potential for determining relationships between diversity, community structure, and ecosystem function in this important group of organisms.

*See also:* **Nitrogen in Soils:** Cycle; Nitrates; Plant Uptake

## Further Reading

Belser LW (1979) Population ecology of nitrifying bacteria. *Annual Review of Microbiology* 33: 309–333.

De Boer W and Kowalchuk GA (2000) Nitrification in acid soils: microorganisms and mechanisms. *Soil Biology and Biochemistry* 33: 853–866.

Koops H-P and Pommerening-Röser A (2001) Distribution and ecophysiology of the nitrifying bacteria emphasizing cultured species. *FEMS Microbiology Ecology* 37: 1–9.

Koops H-P, Purkhold U, Pommerening-Röser A, Timmermann G, and Wagner M (2003) The lithoautotrophic ammonia-oxidizing bacteria. In: Dworkin D, Falkow S, Rosenberg E, Schleifer K-H, and Stackebrandt E (eds) *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community.* 3rd edn, release 3.13. New York: Springer-Verlag.

Kowalchuk GA and Stephen JR (2001) Ammonia-oxidizing bacteria: a model for molecular microbial ecology. *Annual Review of Microbiology* 55: 485–529.

Laanbroek HJ and Woldendorp JW (1995) Activity of chemolithotrophic nitrifying bacteria under stress in natural soils. *Advances in Microbial Ecology* 1: 275–304.

Prosser JI (ed.) (1986) *Nitrification*. Oxford, UK: IRL Press.

Prosser JI (1989) Autotrophic nitrification in bacteria. *Advances in Microbial Physiology* 30: 125–181.

Prosser JI and Embley TM (2002) Cultivation-based and molecular approaches to characterisation of terrestrial and aquatic nitrifiers. *Antonie van Leeuwenhoek* 81: 165–179.

Prosser JI, Embley TM, and Webster G (2002) The influence of selection pressures on species diversity, functional gene diversity and activity of ammonia oxidising bacteria. In: Hails R, Beringer JE, and Godfray HCJ (eds) *Genes in the Environment*, pp. 187–202. Oxford, UK: Blackwell Publishing.

Schmidt I, Slickers O, Schimid M *et al.* (2002) Aerobic and anaerobic ammonia oxidizing bacteria – competitors or natural partners? *FEMS Microbiology Ecology* 39: 175–181.

# Plant Uptake

**A Hodge**, University of York, York, UK

## Introduction

Nitrogen (N) is essential for life because it is a component of nucleic acids and amino acids and thus peptides and proteins. In addition, plants also require N for chlorophyll. In most terrestrial ecosystems, N is the primary limiting nutrient. In soils, N is present predominantly in organic form, and inorganic N is made available by the mineralization of organic N to ammonium ($NH_4^+$) and subsequent oxidation to nitrate ($NO_3^-$). Both these inorganic N sources can be taken up by plant roots, but roots have to compete with soil microorganisms as well as other plant roots for this N. Recent evidence suggests that plants may be able to access simple organic N sources directly without the need for the mineralization process to release inorganic N, but these N sources also have to be captured in a competitive environment. The distribution of N sources in soil, particularly organic N forms, is both spatially and temporally heterogeneous. Plant roots respond to this heterogeneity in a number of ways in order to capture N from the soil system.

## Plant–Microbial Competition for N

Mineralization of organic material to $NH_4^+$ is carried out by a wide range of soil animals and microorganisms. However, mineralization and the subsequent conversion of $NH_4^+$ to $NO_3^-$ are generally considered to be the rate-limiting steps in the terrestrial N-cycle. Due to their rapid turnover times and rapid adaptation rate, microorganisms are generally thought to utilize inorganic N sources first, with the plant only being able to access what is left over after microbial uptake. Exceptions are plants such as legumes and alder which form symbiotic associations with rhizobium and *Frankia* bacteria, respectively. These symbionts form nodules on the roots which are capable of fixing dinitrogen and thus supply the plant with N directly. The majority of plants, however, do not form $N_2$-fixing symbiotic associations and must rely on N uptake via their roots or mycorrhizas.

Most transformations of N also depend on the supply of carbon (C), and thus these two nutrient cycles are closely linked. The C:N ratio of the substrate determines the rates at which N (and C) are released and mineral N made available for plant uptake. Also influential are the composition of the decomposer microbial community and its requirements for N. Fungi tend to have a higher substrate assimilation efficiency and a lower N requirement per unit of C assimilated than bacteria. Thus, if the decomposer population is predominantly fungal, more N may be released than if it was predominantly bacterial. However, fungi also tend to turn over less rapidly than bacterial populations, so the N contained in their structural components will be locked up for a longer time than bacterial N. Generally net mineralization, resulting in the release of N, occurs if the C:N ratio of the substrate being decomposed is lower than approx. 25:1. At higher C:N ratios, net immobilization occurs and the microbial decomposer community requires additional N from the soil system to decompose the substrate.

Net immobilization of N reduces the N available for the plant and can reduce overall plant growth. Although other factors influence the release of N, for example lignin content of the substrate, there is a strong inverse relationship between the amount of plant N captured from organic substrates and the C:N ratio of the substrate, as shown in **Table 1**.

In addition to microorganisms soil animals also play an important role in the mineralization process through physical, chemical, and biological processes. Soil animals can physically break up and redistribute organic materials in the soil. They also ingest organic materials which are subsequently broken down by a range of enzymes and/or by microflora present in their guts. Much of the C they consume is used for metabolic processes and the excreted waste products, in the form of $NH_4^+$, urea, or amino acids, are a much more readily available form of N for plant uptake than the organic material originally ingested. Soil animals feed on microorganisms as well as other soil animals. The plant benefits from the regulation of the size and activity of the microbial population, and from the release of more N into the soil–plant system. An increase in plant N capture of up to 75% has been reported as a result of protozoa grazing on soil microorganisms. Thus, although microorganisms may

**Table 1** The percentage N captured by plants from organic material (*Lolium perenne* shoots) of varying C:N ratio

| *Plant species* | *C:N ratio* | *Duration (days)* | *N from organic material captured by plant (%)* |
|---|---|---|---|
| *Lolium perenne*[a] | 12:1 | 30 | 22 |
| *Lolium perenne*[b] | 12:1 | 70 | 26 |
| *Poa pratensis*[b] | 12:1 | 70 | 26 |
| *Lolium perenne*[c] | 21:1 | 49 | 11 |
| *Lolium perenne*[d] | 31:1 | 39 | 5 |
| *Poa pratensis*[d] | 31:1 | 39 | 3 |
| *Lolium perenne*[e] | 31:1 | 56 | 9 |
| *Poa pratensis*[e] | 31:1 | 56 | 5 |

[a] Data from Hodge A, Stewart J, Robinson D, Griffiths BS, and Fitter AH (2000) Plant N capture and microfaunal dynamics from decomposing grass and earthworm residues in soil. *Soil Biology and Biochemistry* 32: 1763–1772.

[b] Hodge A, Stewart J, Robinson D, Griffiths BS, and Fitter AH (2000) Spatial and physical heterogeneity of N supply from soil does not influence N capture by two grass species. *Functional Ecology* 14: 645–653.

[c] Hodge A, Stewart J, Robinson D, Griffiths BS, and Fitter AH (2000) Competition between roots and soil micro-organisms for nutrients from nitrogen-rich patches of varying complexity. *Journal of Ecology* 88: 150–164.

[d] Hodge A, Stewart J, Robinson D, Griffiths BS, and Fitter AH (1998) Root proliferation, soil fauna and plant nitrogen capture from nutrient-rich patches in soil. *New Phytologist* 139: 479–494.

[e] Hodge A, Robinson D, Griffiths BS, and Fitter AH (1999) Why plants bother: root proliferation results in increased nitrogen capture from an organic patch when two grasses compete. *Plant, Cell and Environment* 22: 811–820.

initially sequester inorganic N, protozoa grazing ensure that this N is released back into the root-soil environment and becomes available for root uptake.

## Inorganic N

The fate of N into different pools in the soil can be measured by addition of materials labeled (or enriched) with the stable isotope $^{15}N$. Short time scales (i.e., a few hours to days) are used to measure flow rates between pools, because recycling of labeled material over longer incubations may occur and lead to underestimations of the total flux through the pool. Such short-term studies on partitioning of labeled inorganic N sources added to soils have been used to measure the competition between microorganisms and plants (Table 2). Generally, a larger percentage of the $^{15}N$-labeled inorganic N material is detected in the microbial biomass than plants, indicating that, at least in the short term, microorganisms are the superior competitors. These types of studies also indicate how rapid the flux is through the inorganic pools in soil. Thus the size of the pool, which is typically measured to estimate available N, does not reflect the N that is actually available to the plants and soil microorganisms if that pool has a high flux rate.

The studies summarized in Table 2 added N compounds enriched with $^{15}N$ in order to trace the uptake of N by the different groups of organisms (i.e., microorganisms or plants). Another method used by some investigators is to measure differences in the natural abundance of $^{15}N$ among different pools of N in soil to determine the source of N that plants or microbes are utilizing. This technique, however, is complicated by fractionation of $^{15}N$ by biochemical processes and mixing of N among the various soil pools. In addition, there is evidence that the $\delta^{15}N$ of plants is affected by their mycorrhizal status. Thus, some investigators have concluded that plant $\delta^{15}N$ values cannot be used to identify the N source that a plant is accessing in the soil. Although natural abundance techniques have considerable potential, many technical difficulties need to be resolved before the technique can be effectively used in a wide range of environments.

Some plants, such as members of the Ericaceae, prefer to take up $NH_4^+$ over $NO_3^-$. There is some evidence to suggest that these plants may lack the capability to reduce large quantities of $NO_3^-$ in their tissues. These plants generally grow in acidic soils, where nitrification of $NH_4^+$ to $NO_3^-$ is inhibited and $NH_4^+$ accumulates, and this may explain their inorganic N source preference. If $NH_4^+$ accumulates inside the plant it becomes phytotoxic, thus $NH_4^+$ is rapidly transformed to amide or amino acid form after absorption. For most plants, $NO_3^-$ is the major form of N absorbed and nitrate is then reduced to nitrite ($NO_2^-$) and to $NH_4^+$ by the linked enzymes nitrate and nitrite reductase. When $NO_3^-$ is mixed with a small amount of $NH_4^+$ (rather than just $NO_3^-$ alone), plant growth generally increases. These conditions reflect the availability of inorganic N in most fertile soils but of course this will also depend on soil type. In soil solution, $NH_4^+$ concentrations are generally lower than $NO_3^-$ concentrations because of rapid transformation to $NO_3^-$ or because $NH_4^+$ is strongly held on to negatively charged particles such as soil organic materials and clays, and so unavailable for plant uptake. The uptake of a particular inorganic N source may also vary depending on the time of

**Table 2** Data from field studies following the partitioning of inorganic N between plants and the microbial biomass after addition of $^{15}N$-labeled sources

| | | | $^{15}N$ detected (%) | |
|---|---|---|---|---|
| *Plant species* | *Added material* | *Duration* | *Plants* | *Microorganisms* |
| *Grassland*[a] | $^{15}NH_4^+$ | 1 d | 9–11 | 46–61 |
| *Temperate forest*[b] | $^{15}NO_3$ | 1 d | 20–26 | 38–50 |
| Well drained | $NH_4{}^{15}NO_3$ | 2 d | 0.19 | 20 |
| | $^{15}(NH_4)_2SO_4$ | 2 d | 0.04 | 21 |
| Poorly drained | $NH_4{}^{15}NO_3$ | 2 d | 0.01 | 0 |
| | $^{15}(NH_4)_2SO_4$ | 2 d | 0.01 | 0 |
| *Allium tricoccum*[c] in a hardwood forest floor | $^{15}(NH_4)_2SO_4$ | 2 d | 1 | 22 |
| | $K^{15}NO_3$ | 2 d | 3 | 32 |

[a] Data from Jackson LE, Schimel JP, and Firestone MK (1989) Short-term partitioning of ammonium and nitrate between plants and microbes in an annual grassland. *Soil Biology and Biochemistry* 21: 409–415.

[b] Groffman PM, Zak DR, Christensen S, Mosier A, and Tiedje JM (1993) Early spring nitrogen dynamics in a temperature forest landscape. *Ecology* 74: 1579–1585.

[c] Zak DR, Groffman PM, Pregitzer KS, Christensen S, and Tiedje JM (1990) The vernal dam: plant–microbe competition for nitrogen in northern hardwood forests. *Ecology* 71: 651–656.

year, as it is well established that $NO_3^-$ uptake is more sensitive to low temperature than is $NH_4^+$.

In addition, plants can absorb N through their foliage. Foliar additions of N in the form of urea, ammonium nitrate, and potassium nitrate are added to some crops, although this practice is not widespread. Plant foliage can also absorb nitrogenous gases such as NO, $N_2O$, and particularly ammonia ($NH_3$) from the atmosphere or volatilized from the soil surface. Losses of N via $NH_3$ volatilization can be considerable, i.e., approx. 50% of the N in animal slurries and manures can be lost in this way. Ammonia volatilization is increased under dry conditions, high temperatures, high air movement and in soils of alkaline pH and low organic matter or clay content. The presence of plants helps to reduce $NH_3$ losses because their roots compete for $NH_4^+$ in the soil, preventing its transformation and subsequent loss as $NH_3$. Also their foliage directly absorbs volatilized $NH_3$ from the soil surface. Foliar absorption of $NH_3$ is of considerable importance to the plant's N uptake in some terrestrial N cycles. However, cell membranes are highly permeable to $NH_3$ and the local, rapid increase in pH within the plant cell may irreversibly damage it. Thus although this mechanism can contribute to the plant's total N uptake, at high absorption levels $NH_3$ becomes toxic to the plant.

Another input of N is emission of $NH_3$ and other gases including nitrogen oxides ($NO_x$; i.e., NO and $NO_2$) and $N_2O$ from intensive agricultural and industrial sources, and their subsequent deposition elsewhere via wet (largely in the form of nitrate and ammonium) or dry (largely in the form of $NO_x$ and ammonia) deposition. This redistribution of N can have a positive effect on plant productivity in N-limited systems. However, excessive levels can cause soil acidification and adversely affect plant diversity.

## Plant Inorganic N Uptake and Rhizosphere pH

The form of inorganic N taken up by the plant can strongly influence the pH of the rhizosphere. To maintain the electrochemical gradient between root cells and soil (cells typically have electrochemical potentials of less than $-100\,mV$) which drives cation uptake, plants taking up N as nitrate will release bicarbonate ($HCO_3^-$) and/or hydroxyl ($OH^-$) ions, causing the rhizosphere pH to increase. In contrast, plants taking up ammonium as the main form of N will tend to release protons, causing the rhizosphere pH to decrease. However, even when grown on nitrate as the sole source of N, the rhizosphere pH of some plants (such as chickpea) can decrease. This decrease is almost certainly due to a higher ratio of cation to anion uptake. Leguminous plants with active $N_2$-fixing nodules also tend to have a large cation to anion uptake ratio and so a tendency to excrete protons and so decrease the pH of their rhizosphere soil. Thus the consequences for the rhizosphere pH will depend not only on the form of N (i.e., $NH_4^+$, $NO_3^-$ or $N_2$ fixation) the plant is taking up but also on the plant species present and the buffering capacity of the soil.

## Organic N

In some ecosystems such as the arctic, alpine, and boreal forest, annual plant N uptake is higher than should be available through inorganic N capture alone. This has led to the conclusion that many of the plant species in these ecosystems are accessing organic N sources directly without the prior need for microbial mineralization. Forms of organic N in these soils range from the complex (e.g., humus) to the simple (i.e., amino acids and urea), although most of the evidence for plant uptake of intact organic N comes from studies on amino acid N. Some amino acids are present in the dissolved organic nitrogen (DON) fraction, but the majority are protected from uptake by adsorption on to soil particles or humic substances. Environmental fluctuations such as freeze–thaw and drying–rewetting events can result in the disruption of these soil particles and some of this bound amino N may subsequently be released. These environmental events can also disrupt microbial cells, leading to the death of a large fraction of the microbial community and a competitive advantage for the plant providing the plant itself is not also adversely affected by the same environmental events. The advantage to the plant is due not only to a decline in the number of microbial competitors but an increase in N supply through the release of bound amino N forms from soil particles and from the lysis of the microbial cell contents.

Studies using plants grown in solution culture or excised roots have demonstrated that uptake of organic N can occur at levels comparable with, or even in excess of, N uptake from inorganic sources. However, in such studies the uptake of organic N is maximized because, unlike in soil, these N compounds are freely available rather than physically or chemically bound and the plant does not have to compete with microorganisms for these N sources. Although a number of amino acid transporters have been identified on plant roots, amino acids in soil have to move across the rhizosphere, a zone of high microbial activity. Many of these microbes also possess amino acid transporters. The exception is with plants that form mycorrhizal associations, in which case uptake by the fungal hyphae may occur. In these cases, transport across the hyphosphere (the volume of soil influenced by the external mycelium of the mycorrhizal fungus) instead of the rhizosphere occurs. How significant organic N is as a source of N to the nonmycorrhizal plant is questionable. Nevertheless studies conducted in soil microcosms or under field conditions have demonstrated that uptake of intact amino acids can occur regardless of the mycorrhizal status of the plant. The amino acid is generally labeled with both $^{13}C$ and $^{15}N$ (i.e., 'dual'-labeled) so that uptake of intact organic N can be distinguished from uptake of mineralized N. If N was being taken up by the plant in mineralized form, $^{15}N$ levels would be enriched in the plant tissue but there would be no $^{13}C$ enrichment; whereas, if N was being taken up in organic form, both $^{13}C$ and $^{15}N$ would be enriched in the plant tissue. The uptake of the amino acid glycine has been used widely to demonstrate such an uptake mechanism (**Table 3**). However, glycine is a relatively poor substrate for microbial growth and the amounts of glycine added in such experiments generally far exceed levels normally found in the soil solution. Thus, what has been observed in these studies may not be applicable to ambient conditions. Moreover, although the uptake of glycine in intact organic form is high in plants harvested a few hours or days after glycine addition, in the longer term, plants probably capture more total N from glycine in the form of mineralized N (**Table 3**). Further progress in understanding plant uptake of organic N sources requires advances in analytical techniques and also measurement of amino acids at more realistic concentrations.

## Root Responses

Root systems are modular which allows them to be very plastic in their response to prevailing environmental factors, including the distribution of nutrient resources. When N availability is low, plants generally increase their root weight ratio (RWR; ratio of root weight to total plant weight) because more biomass is allocated to the root system. However, the RWR is not necessarily the best root trait to measure in order to assess the ability of plants to capture water, N, and nutrients in general. The RWR will include old lignified roots not actively involved in nutrient uptake *per se*, and also changes in the architecture of the root system in response to nutrient availability can occur independent of changes in root biomass. Uptake of nutrients is more closely related to root length than root mass. Roots can only absorb ions from the soil that are transported to them by diffusion and mass flow. Beneficial physiological changes in ion uptake rates by roots can enhance uptake of mobile ions such as nitrates. In contrast, more immobile ions (e.g., phosphate) are limited more by diffusion to the root rather than uptake at the root surface so would be less affected by changes in root uptake rates. The diffusion coefficient for $NO_3^-$ in the soil environment is approx. $10^{-10}\,m^2\,s^{-1}$ and, in contrast, the diffusion coefficient for phosphate is approx. $10^{-13}$–$10^{-15}\,m^2\,s^{-1}$. The lower diffusion coefficient of phosphate is due to insoluble complexes formed by phosphate ions with cations such as $Al^{3+}$, $Fe^{3+}$, and $Ca^{2+}$. Diffusion coefficients for $NH_4^+$ range between approx. $10^{-10}$ and $10^{-12}\,m^2\,s^{-1}$, but the availability of $NH_4^+$ is further reduced by sorption in soil. Thus, the distribution of $NO_3^-$ in soils should be more uniform than

**Table 3** Uptake of intact glycine from various studies

| *Plant species* | *Harvest time after addition of glycine* | *Glycine added taken up in intact form (%)* | *Mycorrhizal status* | *Comments* |
|---|---|---|---|---|
| *Vaccinium myrtillus*[a] | Few hours to 7 days | 91 | Ericoid mycorrhizal | Glycine injected into the mor layer of a boreal forest. Rates of glycine uptake were similar to those of [$^{15}$N]ammonium |
| *Deschampsia flexuosa*[a] | | 64 | Arbuscular mycorrhizal | |
| *Pinus sylvestris*[a] and *Picea abies*[a] | | 42 | Ectomycorrhizal | |
| *Trifolium hybridum*[b] | 21 h | 23 | Arbuscular mycorrhizal | Plants grown in pots in the field. None of the plant species took up glycine at greater rates compared with inorganic N |
| *Trifolium pratense*[b] | | 23 | Arbuscular mycorrhizal | |
| *Phleum pratense*[b] | | 19 | Arbuscular mycorrhizal | |
| *Ranunculus acris*[b] | | 20 | Not tested | |
| *Triticum aestivum*[c] | 4–8 h | 20 | Nonmycorrhizal | Plants grown in the field |
| *Lolium perenne*[d] | 3 h | 80 | Nonmycorrhizal | Plants grown in solution culture |
| *Plantago lancelota*[e] | 41 days | Negligible | Nonmycorrhizal and arbuscular mycorrhizal | Plants grown in pots. Uptake of intact glycine was negligible, but 72% of N from glycine was detected in the plants |

[a] Data from Näsholm T, Ekblad A, Nordin A, Giesler R, Högberg M, and Högberg P (1998) Boreal forest plants take up organic nitrogen. *Nature* 392: 914–916.

[b] Näsholm T, Huss-Danell K, and Högberg P (2000) Uptake of organic nitrogen in the field by four agriculturally important plant species. *Ecology* 81: 1155–1161.

[c] Näsholm T, Huss-Danell K, and Högberg P (2001) Uptake of glycine by field grown wheat. *New Phytologist* 150: 59–63.

[d] Thornton B (2001) Uptake of glycine by non-mycorrhizal *Lolium perenne*. *Journal of Experimental Botany* 52: 1315–1322.

[e] Hodge A (2001) Arbuscular mycorrhizal fungi influence decomposition of, but not plant nutrient capture from, glycince patches in soil. *New Phytologist* 151: 725–734.

that of phosphate and, in principle, even a relatively low root length density (RLD) should be enough to capture sufficient N from the soil when present in the form of $NO_3^-$. In actuality, plants have to capture N (and indeed other nutrients) in competition with other plant roots as well as soil microorganisms, also a large number of inputs in natural systems occur in organic form via leaf litter, dead roots, and animals, leading to a heterogeneous distribution of resource supply both spatially and temporally. Thus, the distribution of resources in soil, even of $NO_3^-$, is not uniform but patchy.

Root systems can respond to nutrient-rich zones or patches through proliferation of roots as demonstrated in the classic experiments conducted by Drew and colleagues in the 1970s. Barley roots exposed to a localized supply of nitrate, ammonium, and phosphate increased both the initiation (i.e., number) and extension (i.e., length) of first- and second-order laterals within that localized zone. Moreover shoot weight was only slightly depressed when these nutrients were supplied locally compared with controls receiving a uniformly high supply of these nutrients throughout their entire root system. However Drew's experimental system involved plants grown in sand units supplied with nutrient solutions and did not reflect the complexity of the soil environment. Nevertheless, the study demonstrated that roots respond, often spectacularly, to inorganic N as well as phosphate-rich zones (**Figure 1**).

Although root proliferation has been widely reported, it has only recently been demonstrated to have a benefit in terms of N capture in soil. When two different plant species are grown in competition for resources from an organic N-rich patch (grass shoots dual-labeled with the stable isotopes $^{15}$N and $^{13}$C) added to soil, plant $^{15}$N capture was related to RLD produced in the patch. No $^{13}$C enrichments were detected in the plant tissue, thus N must have been taken up as mineralized N. Thus, when in competition for N, even mineralized N, which can be highly mobile in soil, RLD is important, because the supply of N from the organic patch is finite. In contrast, in the absence of plant–plant competition, the supply of N can be thought of as invariant and RLD is less important.

The timescales to construct new roots in these nutrient-rich zones are generally weeks rather than days (**Figure 2**). In contrast, microbial populations can respond much more rapidly in time scales of hours to days. Thus, root proliferation responses are considered to be more important in terms of plant–plant competition rather than plant–microbial competitive interactions. High-order laterals produced upon root proliferation also tend to have higher N concentrations and thus are more expensive to

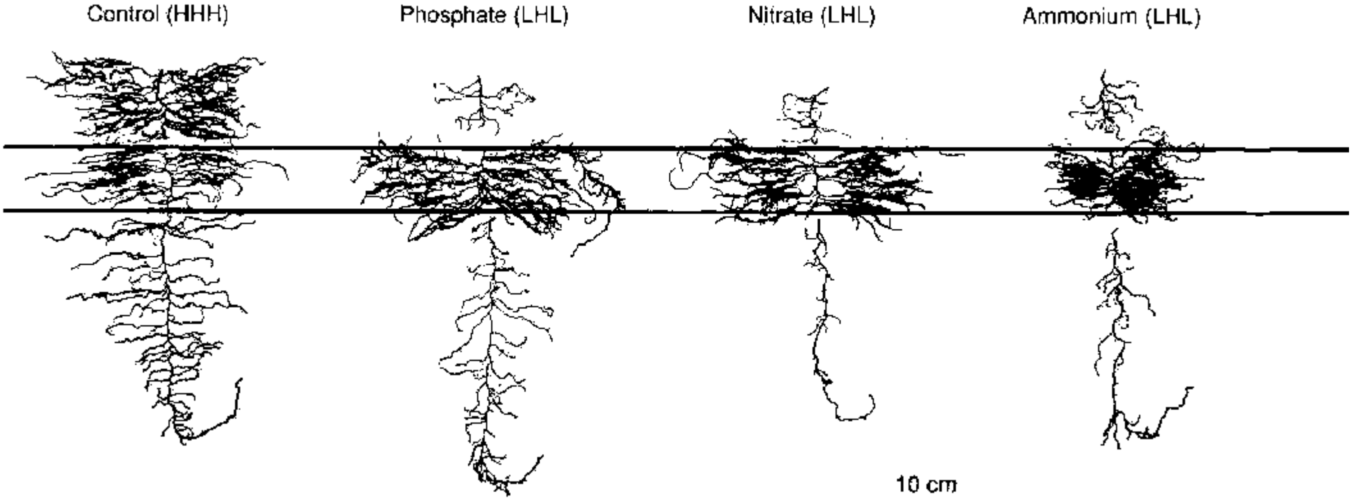


**Figure 1** The proliferation of primary and secondary laterals of barley (*Hordeum vulgare*) root systems when supplied with a 100-fold greater concentration of phosphate, nitrate, or ammonium to their middle part of their root system (indicated by the lines). H, high; L, low, which refer to the nutrient concentrations of phosphate, nitrate, or ammonium received by the top, middle, and bottom of the root system, respectively. The control which received uniformly high levels of nutrients to all parts of its root system is shown as a comparison. Reproduced with permission from Drew MC (1975) Comparison of the effects of a localized supply of phosphate, nitrate, ammonium and potassium on the growth of the seminal root system, and the shoot, in barley. *New Phytologist* 75: 479–490, with kind permission of the trustees of the *New Phytologist*.

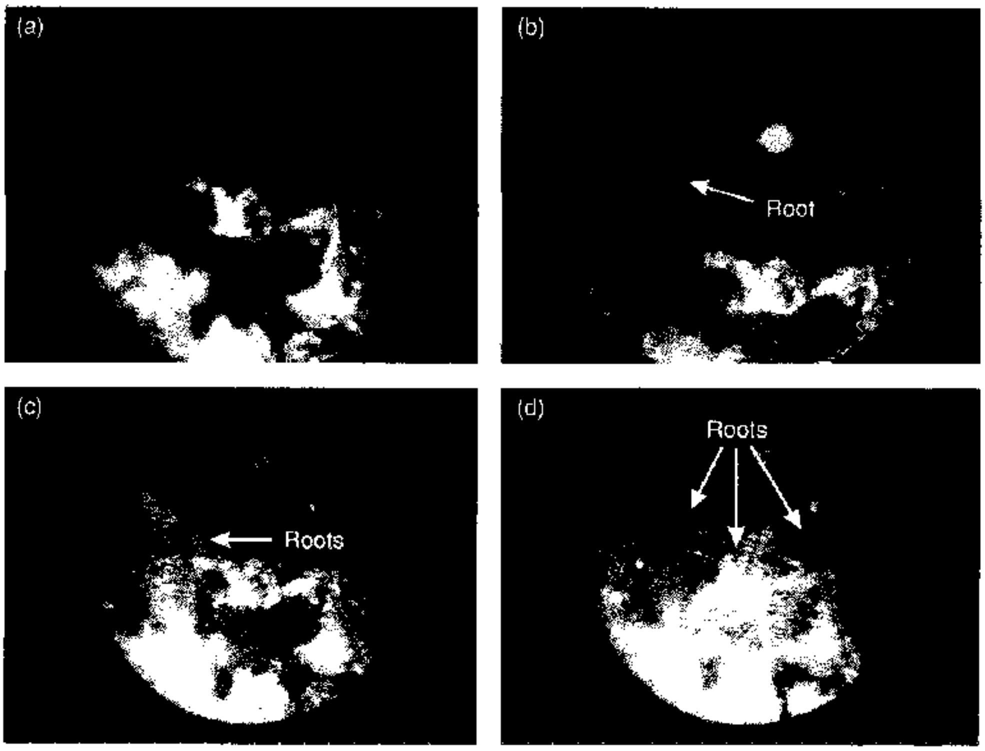


**Figure 2** (see color plate 44) Root growth and proliferation in an N-rich organic material added as a spatially discrete patch and comprising lyophilized algal cell organic material patch (C:N ratio of 3.2:1). (a) One day after the patch was added, no roots in the patch zone were observed; (b) 21 days after the patch was added, a single root was seen growing into the N-rich zone. (c) The same image 25 days and (d) 48 days after the patch was added: increased root production and root-hair development in the N-rich zone was clearly observed. After 49 days, the plants were harvested and found to contain 54% of the N orginally added in the algal cell patch in their tissues. Adapted from *Trends in Plant Science*, vol. 5, Hodge A, Robinson D, and Fitter A. Are microorganisms more effective than plants at competing for nitrogen? pp. 304–308. © 2000, with kind permission from Elsevier Science.

construct but presumably the benefits of exploiting a nutrient-rich patch and the capture of N from it outweigh the disadvantages of these roots' construction costs. Moreover, increased root turnover, i.e., a decrease in root longevity, is often observed in both fertile soils and fertile patches, but again presumably the advantages of having a young population of roots actively involved in nutrient-uptake processes outweigh the disadvantages of high root turnover rates.

## Mycorrhizal Contribution to Plant N Uptake

Most plants form mycorrhizal associations under natural conditions. Mycorrhizal symbiosis is discussed in detail elsewhere so only the benefit of mycorrhizas for plant N uptake will be considered here. Mycorrhizal associations may be expected to enhance N uptake from inorganic N sources because, by virtue of the extension of the fungal mycelium in the soil, they should be better able to compete with the soil microbial community for these N sources compared with plant roots. Many of the fungi involved in the ectomycorrhizal and ericoid mycorrhizal associations have been demonstrated to have a preference for $NH_4^+$ rather than $NO_3^-$ when grown in pure culture. The reasons for this are probably the same as the preference for inorganic shown by plant roots in that it reflects the dominant inorganic N source under which these fungi usually form their associations. In addition, the arbuscular mycorrhiza (AM), ericoid, and ectomycorrhizal associations have all enhanced N capture for their host plant when supplied with $NH_4^+$. Mycorrhizas are probably more effective at enhancing N capture when supplied with $NH_4^+$ than $NO_3$ because of the differences in mobilities of these ions. As $NH_4^+$ is much less mobile than $NO_3^-$, $NH_4^+$ uptake can be enhanced in the same manner that phosphate (which is even more immobile) is enhanced by mycorrhizal colonization, whereas the very mobile $NO_3^-$ diffuses readily to the root surface. However, under certain circumstances, such as drought conditions, when ion mobility is restricted, mycorrhizas have also been demonstrated to enhance $NO_3^-$ uptake for their host plant. In addition, AM colonization of legume roots has been demonstrated to have a positive effect on N nutrition for these plants when their hosts are in a tripartite symbiosis with $N_2$-fixing *Rhizobium* spp, but in this case the enhanced N nutrition of the plant is believed to be a consequence of increased plant P status and uptake of other micronutrients which may have indirectly enhanced the $N_2$-fixation process in addition to improving overall plant growth. AM mycorrhizal fungi have also been suggested to play a role in the redistribution of N within a plant community via fungal connections from legume to nonlegume plants, allowing transfer of N-containing products to the nonlegume host. However, it is unlikely that such a mechanism alone would be sufficient to meet the N demand of the nonlegume host.

The ectomycorrhiza and ericoid mycorrhizas also enhance N nutrition for the host plant by accessing organic N sources. These fungi produce hydrolytic enzymes which mineralize complex N forms, such as chitin and proteins, and then absorb the $NH_4^+$ liberated. Uptake of N from simple organic sources such as amino acids and peptides has also been demonstrated. However, there is considerable variability among species and even among different strains of the same species in their ability to acquire N from organic sources. For example ectomycorrhizal fungi such as *Suillus bovinus*, *Paxillus involutus*, and *Hebeloma crustuliniforme* are classified as 'protein fungi' because their growth yields on organic N are the same, or in excess of, yields on ammonium, whereas the so-called 'nonprotein fungi' *Laccaria laccata* and *Lactarius rufus* grow poorly and have low yields on organic N (gliadin and bovine serum albumin, BSA) sources. Once the more-complex organic N sources are cleaved by the mycorrhizal fungi, this may aid capture of N in more simple organic form by either different mycorrhizal fungi or plant roots. The ability of AM associations to use organic N sources to any great extent is controversial, although it has been shown that AM fungi can access such N sources and transfer the N to their host plant.

There is much current interest in the uptake of intact organic N sources by plants; however, its significance as a major N-acquiring system in most ecosystems is still unclear. For plants, the availability of N whether in inorganic or organic form will be determined by the activities of the microbial community in soil and, thus, plant uptake of N cannot be studied independently of these microorganisms. Plants are effective at competing for N with microorganisms because of their much slower tissue turnover times. Also, by remobilization of N from senescing tissues, plants can reduce their losses of N back into the soil system.

*See also:* **Acid Rain and Soil Acidification**; **Buffering Capacity**; **Carbon Cycle in Soils:** Dynamics and Management; **Foliar Applications of Nutrients**; **Isotopes in Soil and Plant Investigations**; **Mycorrhizal Fungi**; **Nitrogen in Soils:** Cycle; Symbiotic Fixation; **Organic Residues, Decomposition**; **Rhizosphere**; **Root Architecture and Growth**

## Further Reading

Drew MC (1975) Comparison of the effects of a localized supply of phosphate, nitrate, ammonium and potassium on the growth of the seminal root system, and the shoot, in barley. *New Phytologist* 75: 479–490.

Fitter A, Hodge A, and Robinson D (2000) Plant response to patchy soils. In: Hutchings MJ, John EA, and Stewart AJA (eds) *The Ecological Consequences of Environmental Heterogeneity*, pp. 71–90. Oxford, UK: Blackwell Science.

Hodge A, Robinson D, Griffiths BS, and Fitter AH (1999) Why plants bother: root proliferation results in increased nitrogen capture from an organic patch when two grasses compete. *Plant, Cell and Environment* 22: 811–820.

Hodge A, Robinson D, and Fitter A (2000) Are microorganisms more effective than plants at competing for nitrogen? *Trends in Plant Science* 5: 304–308.

Hutchings MJ and de Kroon H (1994) Foraging in plants: the role of morphological plasticity in resource acquistion. *Advances in Ecological Research* 25: 159–238.

Lipson D and Näsholm T (2001) The unexpected versatility of plants: organic nitrogen use and availability in terrestrial ecosystems. *Oecologia* 128: 305–316.

Marschner H (1995) *Mineral Nutrition of Higher Plants*, 2nd edn. London, UK: Academic Press.

Robinson D (1994) The responses of plants to non-uniform supplies of nutrients. *New Phytologist* 127: 635–674.

Smith SE and Read DJ (1997) *Mycorrhizal Symbiosis*. San Diego, CA: Academic Press.

# Symbiotic Fixation

**J I Sprent**, University of Dundee, Dundee, UK

## Introduction

Nitrogen-fixing bacteria are known to form symbiotic associations with some members of all major groups of plants, as well as with some fungi. Although there are numerous reports of nitrogen-fixing bacteria occurring in animals, for example termite guts, the significance to their hosts remains to be proven. In global terms, nodulated plants (both legume and actinorhizal) fix most nitrogen, but many of the other symbioses are very important within their own ecosystems. All nonnodulated nitrogen-fixing symbioses have cyanobacteria as their endosymbiont.

Common features to most symbiotic systems that have been studied in detail are: first, that given a supply of combined nitrogen (nitrate, ammonium, organic N) from the soil, they will tend to use this in preference to fixing $N_2$; in most cases this partially or completely excludes the development of the symbiosis, a situation frequently described in agricultural systems, but which is likely to occur also in natural ecosystems; second, that the host plant assimilates the ammonium produced by the nitrogen-fixing endophyte. Thus a series of control mechanisms is invoked whereby the plant host supplies organic carbon, amino compounds, and mineral elements in exchange for ammonium. The only exception is in lichens, where the fungal component is probably responsible for supplying minerals from the soil or other substratum, and carbon fixation is carried out by either a green alga or the nitrogen-fixing endophyte (see **Table 1**).

Where estimates of fixed nitrogen are given, these are subject to large errors because of limitations of available methods for use on a field scale.

## Cyanobacterial Symbioses

Table 1 lists the major types of symbiosis between cyanobacteria and plants. Nitrogen-fixing cyanobacteria usually have nitrogenase located in specialized cells known as heterocysts, and this is true of almost all symbiotic forms studied.) Exceptions to this observation are those listed in **Table 1** as having 'coccoid forms' of endophyte. Shortly after infection, heterocyst frequency may be similar to that of free-living forms, but subsequently it increases markedly as the endophyte fixes nitrogen using carbon fixed by the host, making photosynthesis by vegetative cells unnecessary. Further, in many systems, the cyanobacteria do not receive light, although they continue to produce photosynthetic pigments. Only in two membered lichens (**Table 1**) does the cyanobacterium need to fix both $CO_2$ and $N_2$.

Another common feature to all except some diatoms is that water is necessary for infection by motile filaments (hormogonia) of the cyanobacterium. In bryophytes *Azolla* and *Gunnera*, these invade cavities in the host plant: *Azolla* also has a means of dispersing its endophyte during spore production. Within cavities, host cells make close contact with their symbionts and often undergo structural modifications enabling the efficient interchange of nutrients. The need for water may be one of the reasons why cyanobacterial symbioses are limited in number. Only lichens are found in very dry areas.

### Lichens

Identification of the fungi involved in lichen formation is difficult, but it is likely that about 20% of known fungi can be found in lichens. Of these, the vast majority are ascomycetes. No phycomycetes are known to form lichens, although one coenocytic species, *Geosiphon pyriforme*, produces bladders that can house *Nostoc punctiforme*. This interesting association may

**Table 1** Major categories of nitrogen-fixing symbioses involving cyanobacteria

| *Category* | *Genus of host* | *Genus of cyanobacterium* | *Comments* |
|---|---|---|---|
| Diatoms | *Bacteriastrum* | *Richelia* | Includes both marine and freshwater genera. Marine genera may be important in vertical translocation of organic carbon in oceans |
| | *Chaetoceros* | *Richelia* | |
| | *Denticulata* | Coccoid form | |
| | *Epithemia* | Coccoid form | |
| | *Hemiaulus* | *Richelia* | |
| | *Neostreptotheca* | Coccoid form | |
| | *Rhopalodia* | Coccoid form | |
| | *Rhizosolenia* | *Richelia* | |
| | *Rupertia* | *Nostoc* | |
| | *Streptotheca* | Coccoid form | |
| Lichens | *Collema* | *Nostoc* | Lichens are important in boreal and temperate forests, in arid areas, and on rocks, often in coastal areas. They may have only two components (fungal, cyanobacterial), or three, when a green alga is also present. In the former the cyanobacterial member fixes both $CO_2$ and $N_2$, in the latter, only $N_2$. Some lichens have unicellular endophytes |
| | *Dendriscocaulon* | *Scytonema* | |
| | *Dictyonema* | *Scytonema* | |
| | *Ephebe* | *Fischerella* | |
| | *Leptogium* | *Nostoc* | |
| | *Lichina* | *Calothrix* | |
| | *Lobaria* | *Nostoc* | |
| | *Massalongia* | *Scytonema* | |
| | *Nephroma* | *Nostoc* | |
| | *Pannaria* | *Nostoc* | |
| | *Parmelia* | *Nostoc* | |
| | *Peltigera* | *Nostoc* | |
| | *Placopsis* | Uncertain | |
| | *Placynthium* | *Dichothrix* | |
| | *Polychidium* | *Scytonema* | |
| | *Pseudocyphellaria* | *Nostoc* | |
| | *Solorina* | *Nostoc* | |
| | *Stereocaulon* | *Nostoc* | |
| | *Sticta* | *Nostoc* | |
| Hornworts | *Anthoceros* | *Nostoc* | The cyanobacteria are located in pockets in the thallus |
| | *Phaeoceros* | *Nostoc* | |
| | *Notothylas* | *Nostoc* | |
| | *Dendroceros* | *Nostoc* | |
| Thalloid liverworts | *Blasia* | *Nostoc* | The cyanobacteria are located in pockets in the thallus |
| | *Cavicularia* | *Nostoc* | |
| Ferns | *Azolla* | *Anabaena* | Although '*Anabaena*' is in common use, it is really a *Nostoc* and is found in leaf pockets |
| Cycads | *Bowenia* | *Nostoc* is very common, other reports include *Calothrix* | All extant cycads are symbiotic. Cyanobacteria are found in specially modified roots, usually referred to as 'coralloid,' but sometimes called 'nodules' |
| | *Ceratozamia* | | |
| | *Cycas* | | |
| | *Dioon* | | |
| | *Encephalartos* | | |
| | *Lepidozamia* | | |
| | *Macrozamia* | | |
| | *Microcycas* | | |
| | *Stangeria* | | |
| | *Zamia* | | |
| Angiosperms | *Gunnera* | *Nostoc* | All extant species are symbiotic. Unlike all the above, except some diatoms, cyanobacteria are intracellular |

well represent another type of symbiosis for which more examples remain to be discovered.

In some areas, lichens are responsible for significant inputs of fixed nitrogen into soils, either directly or indirectly. Two of the most widely studied systems are the boreal forests and arctic tundra, where nitrogen-fixing lichens (often thalloid *Sterocaulon* spp.) may be the dominant component of the ground flora, as in parts of Scandinavia and northern Canada. Although the total amount of nitrogen fixed may seem low by agricultural standards (1–2 compared with up to 400 kg N ha$^{-1}$ per year), when set against the requirements of growth at low temperatures, it may be sufficient not only for the lichens, but, following mineralization, for other organisms in the ecosystem. Similar arguments can be applied to the epiphytic lichen *Lobaria oregana* found on old-growth Douglas fir in the Pacific Northwest of the USA, where it contributes 3–4 kg N $ha^{-1}$ per year, which may enter the soil via stem flow. In dry areas, both cold and hot, ranging from Antarctica to the tropics, lichens and free-living cyanobacteria may be among the few photosynthetic organisms found. Lichens can survive years of desiccation then resume growth and both $CO_2$- and $N_2$-fixing activity. However, little fixed nitrogen may be released into the soil, as the lichens are part of a tightly controlled ecosystem where the presence of bacteria and other organisms suggest that the entire nitrogen cycle occurs within it. Similar arguments apply to crustose lichens occurring on rocks (and buildings).

### Bryophytes

This category includes the hornworts and thalloid liverworts of **Table 1**. They are found in all parts of the world on bare, nitrogen-deficient soils that are very moist but not flooded. Botanically they are interesting because of their life cycle and the means of transport between the gametophyte and sporophyte generations. **Figure 1**a shows the main features of a hornwort, *Anthoceros*. A relevant question relating to this association is: Why are there so few bryophytes with nitrogen-fixing symbionts? Current molecular evidence suggests that extant symbiotic forms arose from an ancient branch of the bryophyte lineage. Many modern bryophytes, particularly mosses, have adapted to live in harsh environments, including dry areas. As water is essential for the cyanobacteria to enter pockets of the developing thallus, there may have been a selective disadvantage to symbiotic forms.

### Ferns

*Azolla* is an unusual fern in being both aquatic and having a nitrogen-fixing endophyte (**Figure 1b**). Being aquatic, it is not strictly relevant to a discussion on soils, but, as it has been and still is widely used as a nitrogen fertilizer for wetland rice, it will be mentioned. Its agricultural use has been the subject of many reviews. These have taken different standpoints

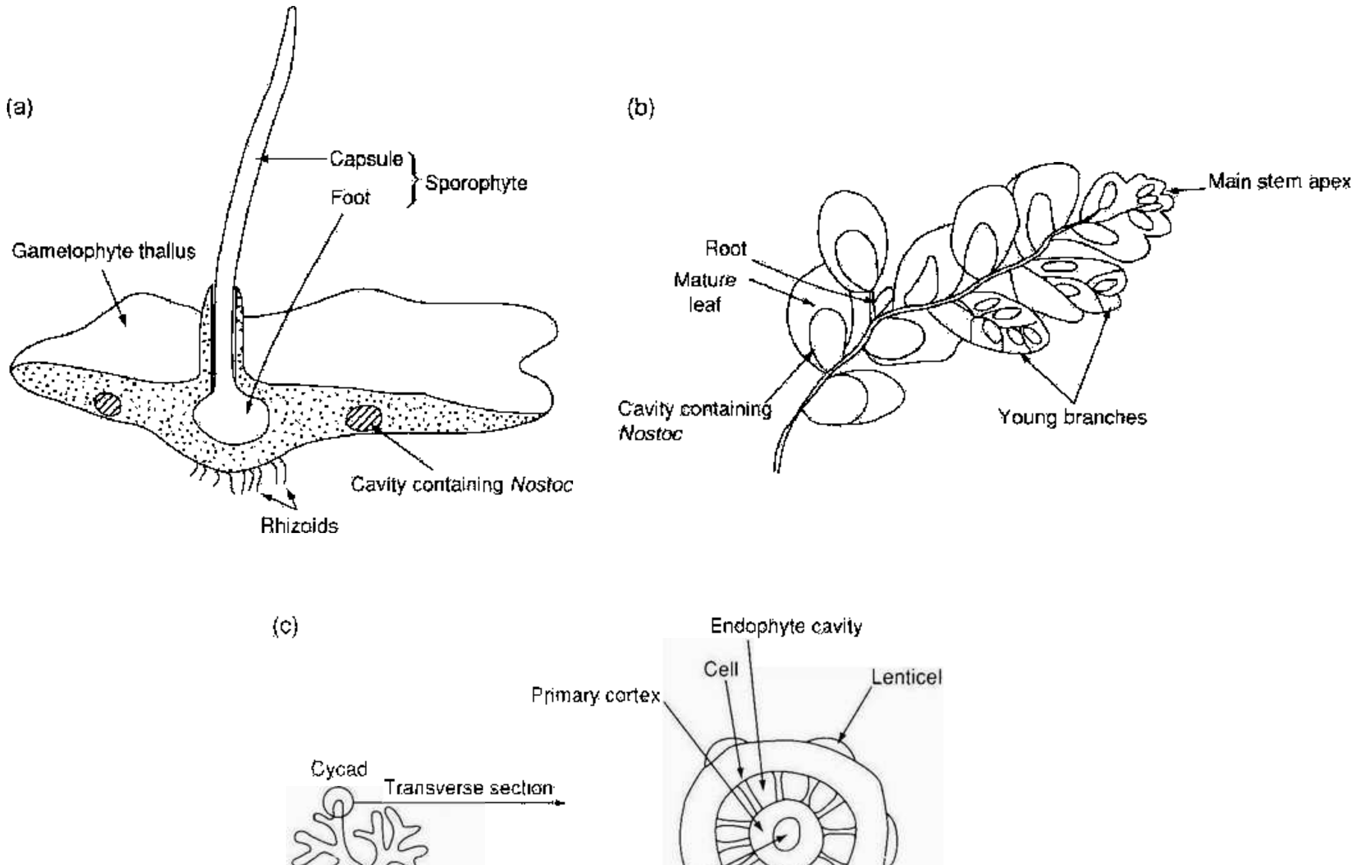


**Figure 1** Some cyanobacterial symbioses with plants: (a) the hornwort *Anthoceros*, approximately ×10 natural size; (b) *Azolla*, ×10 natural size; (c) generalized cycad, approximately natural size; and magnified transverse section. © J Sprent.

as to the practicality of the use of *Azolla* in modern rice-growing, although there is general agreement that it can yield significant amounts of easily mineralized N. However, in many areas of the world where *Azolla* has been introduced, for example South Africa, it can also be a nuisance by blocking water channels. Recently attention has been given to control of *Azolla* by insects and other agents, and in its use for feeding fish, poultry, and pigs.

Why only one genus of fern is known to have a nitrogen-fixing symbiosis is not clear. Current fossil evidence is that *Azolla* evolved comparatively recently compared with other ferns, suggesting that it may have been a 'one-off' event to fill a particular need.

### Cycads

Many texts state that cycads are only found naturally in the southern hemisphere, but this is not true. Although their frequency may be greater in the southern hemisphere, cycads occur in tropical and warm temperate areas of both hemispheres. This distribution can be traced through the fossil record, which shows that cycads were around at least 230 million years ago, before Pangaea broke up into Gondwana (which moved south) and Laurasia (which moved north). It is possible that the comparative rarity of natural stands of cycads in the northern hemisphere relates to destruction of their habitat. Certainly several of them are listed as endangered. Biologically, cycads are particularly interesting in that all extant and fossil forms appear to be symbiotic and yet no evolutionarily related groups of plants (for example gymnosperms) have nitrogen-fixing members. There must have been pressure toward symbiosis at an early stage in the evolution of cycads. This may have been a shortage of combined nitrogen, but evidence supporting this hypothesis is not yet clear. Many of the most abundant cycads today live in dry environments with low-nutrient soils, but others are found in rainforests and seasonally dry areas.

Hormogonia apparently invade gaps in the cortex of cycad root and stimulate the formation of a much-branched structure, the coralloid root (**Figure 1c**). Within this, *Nostoc* is found in cavities between the primary and secondary cortex. It is very easy to find coralloid roots in cycads grown as pot plants, as they are near to and often protrude from the surface of the soil. However, in other cycads, for example *Macrozamia reidlii*, coralloid roots are found well below ground level (down to 1 m). This species is common in areas of Western Australia that are prone to fire. *M. reidlii* can regenerate after fire from subterranean tubers and the deep-seated coralloid roots remain functional. In some eucalypt forests, cycads apparently provide an important source of nitrogen, equivalent to that of understorey legumes in other eucalypt forests. The largest genus of cycads is *Encephalartos*, widely distributed in a variety of habitats in east and southern Africa. Some species are fire-resistant, others frost-tolerant, and one at least is deciduous. Many standard textbooks describe cycads as 'living fossils.' Whilst they may not be as dominant as in former times, in areas such as parts of Australia and South Africa and in the everglades of Florida they are still a prominent component of natural vegetation. However, as they are widely collected for use in horticulture, some species are now threatened.

The list of cycad genera given in **Table 1** is conservative. Other genera have been proposed, of which the most often cited is *Chigua*, with two species native to Colombia; however, some workers believe that these should be included in *Zamia*. The division of cycads into families is also an area of current uncertainty. However, as far as soils are concerned, it is probably safe to assume that, at any site where cycads grow, there will be an input of symbiotically fixed nitrogen.

### Angiosperms

*Gunnera* is the only genus of angiosperm known to have a symbiotic association with a cyanobacterium, in this case *Nostoc punctiforme*. Taxonomically, *Gunnera* appears to be isolated, placed in its own family, and far removed from other angiosperms with nitrogen-fixing symbioses. There are approximately 40 species and all contain *Nostoc* in modified glands (sometimes referred to as nodules) at the base of petioles. Plants are herbs that vary in size from those with leaves over 1 m in height to small creeping forms. In common with other angiosperms, the endophyte of *Gunnera* is intracellular, in that it is contained within the volume surrounded by a plant cell wall, but extracellular in that it is confined within a membranous sac (now often called a symbiosome). Transport of material into and out of the glands is via phloem, unlike nodulated plants in which outward transport is via the xylem. *Gunnera* appears to lack nitrate reductase, so is unable to use nitrate. All studies suggest that *Nostoc* is capable of fixing enough nitrogen for the needs of its host and hence, in the longer term, is likely to be a good source of N for soil mineralization. There is abundant fossil pollen evidence that *Gunnera* is of ancient origin and its present distribution, largely in the southern hemisphere, is restricted to areas with high humidity. This may have been its downfall if it was unable to adapt to the increasingly dry conditions that spread over much of the world since angiosperms first evolved. However, *Gunnera* appears to be very important in some of its present locations, where nodulated legumes and actinorhizal plants are scarce or absent. An extreme case of this is in the Malvinas (Falkland Islands) off the coast of Argentina, where *Gunnera magellanica* is the only native angiosperm with a nitrogen-fixing symbiont. Here it is common and widespread, being found from sea level to all elevations, and clearly has a vital role to play.

## Nodulated Angiosperms

These fall into two categories, based on the nature of the endophyte. *Frankia*, a filamentous bacterium belonging to the actinomycetes, nodulates a variety of plants (**Table 2**). Legumes and the Ulmaceous genus *Parasponia* are nodulated by unicellular bacteria from the $\alpha$- and $\beta$-proteobacteria, collectively known as rhizobia (**Table 3**). Recent molecular evidence has placed all nodulated plants in a distinct group, known as the Rosid I clade. However, the clade also contains many genera that lack the ability to nodulate. If, as is generally assumed, all members of the clade have some predisposition to form nodules, why is it only expressed in certain genera? The implications of this for the evolution of nodulation are an area of intense current interest, especially as the internal structure of nodules in the two groups is different.

A problem common to all aerobic nitrogen-fixing organisms, including the symbiotic *Frankia* and rhizobia, is to reconcile the need for oxidative phosphorylation to fuel the nitrogen fixation with the fact that the enzyme is inactivated by oxygen. Nodules have a number of ways of controlling the rate of oxygen diffusion to the active enzyme so that nitrogen fixation is balanced with the supply of fixed carbon. In some cases this means reducing the concentration of oxygen from the 20% found in air, by means of fixed and/or variable diffusion barriers. Once a sufficiently low level (about 1%) is reached, there may be a need to maintain a constant flux of oxygen; this is achieved by plant forms of hemoglobin. It has been known for many years that hemoglobin gives most active

**Table 2** Genera of plants known to be nodulated by *Frankia*. These are grouped according to family (numbers in parentheses indicate the total number of genera in the family); families are arranged in four groups, separated by an additional line space, according to their position in the Rosid 1 clade of angiosperms. The first group also contains the Ulmaceae, in which *Parasponia*, the only genus outside the legumes known to nodulate with rhizobia, is placed. It should be noted that there are six groups in this clade. Of the other three, two contain no nitrogen-fixing members and the other, the legumes. All groups have nonnodulating members

| *Family* | *Nodulating genera* | *Comments* |
|---|---|---|
| Elaeagnaceae (3) | *Elaeagnus* | Where known in this group, infection is intercellular and no nodule rootlets are formed. Vesicles in the first two families are spherical and septate, with the exception of *Ceanothus*, where they lack septa. In the Rosaceae, vesicles are elliptical and aseptate. Within the whole group, Rosaceae is sister to the other two families |
| | *Hippophae* | |
| | *Comptonia* | |
| Rhamnaceae (55) | *Ceanothus* | |
| | *Colletia* | |
| | *Discaria* | |
| | *Kentrothamnus* | |
| | *Retanilla* | |
| | *Talguenea* | |
| | *Trevoa* | |
| Rosaceae (100) | *Cercocarpus* | |
| | *Chamaebatia* | |
| | *Cowania* | |
| | *Dryas* | |
| | *Purshia* | |
| Betulaceae (6) | *Alnus* | In this group, infection is through root hairs. Nodule rootlets are present in all genera except *Alnus* |
| Myricaceae (3) | *Comptonia* | |
| | *Myrica* | |
| Casuarinaceae (5) | *Allocasuarina* | Vesicles are absent from *Allocasuarina* and *Casuarina* nodules. *Quadrangula* has recently been proposed for two species previously included in *Casuarina*. In all other genera vesicles are septate, but vary in shape from spherical to club-shaped (Myricaeae) |
| | *Casuarina* | |
| | *Ceuthostoma* | |
| | *Quadrangula* | |
| | *Gymnostoma* | |
| Coriariaceae (1) | *Coriaria* | Details of infection in this group are still unclear. Nodule rootlets occur in *Datisca*. Vesicles are filamentous and aseptate. Nodules on these genera have several other distinct features |
| Datiscaceae (3) | *Datisca* | |

legume nodules their pink internal color, but it is now apparent that the pigment is also a necessary part of the functioning of some actinorhizal nodules. Species living in waterlogged conditions may also need an anatomical pathway that allows air to pass through a series of spaces from above the water level to the nodule. This may be achieved by negatively geotropic nodule rootlets, as in *Myrica*, or by a continuous layer

**Table 3** Currently recognized bacterial genera and species known to form nitrogen-fixing root nodules in association with legume roots (sometimes stems). All except those indicated are in the α-proteobacteria, but not all are in the same branch. The genera *Agrobacterium* and *Allorhizobium* may be incorporated into *Rhizobium*

| *Genus* | *Species* | *Examples of legumes nodulated and comments* |
|---|---|---|
| *Agrobacterium* | ? | Some nodulating bacteria appear to belong to this genus, but their exact nature has not been confirmed |
| *Allorhizobium* | *undicola* | *Neptunia natans* |
| *Azorhizobium* | *caulinodans* | *Sesbania rostrata* stems |
| *Blastobacter* | Not determined | *Pterocarpus erinaceous*, *P. lucens* |
| *Bradyrhizobium* | *elkanii* | *Glycine max* |
| | *japonicum* | *Glycine max* |
| | *liaoningense* | *Glycine max* |
| | sp. (host genus) | Several potential species e.g., sp. (*Lupinus*) are currently defined by their hosts |
| *Burkholderia* | Two spp.? | *Aspalathus carnosa*<br>*Machaerium lunatum*<br>In the β-proteobacteria |
| *Devosia* | *neptuniae* | *Neptunia natans* |
| *Mesorhizobium* | *amorphae* | *Amorpha fruticosa* |
| | *chacoense* | *Prosopis alba* |
| | *ciceri* | *Cicer arietinum* |
| | *huakuii* | *Astragalus sinicus* |
| | *loti* | *Lotus corniculatus* |
| | *mediterraneum* | *Cicer arietinum* |
| | *plurifarium* | *Acacia senegal* |
| | *tianshanense* | *Glycyrrhiza pallidiflora* |
| *Methylobacterium* | *nodulans* | Some *Crotalaria* spp. |
| *Ralstonia* | *taiwanensis* | *Mimosa* spp. In the β-proteobacteria. Now renamed *Wautersia taiwanensis* |
| *Rhizobium* | *etli* | |
| | bv. phaseoli | *Phaseolus vulgaris* |
| | bv. mimosae | *Mimosa affinis* |
| | *gallicum* | |
| | bv. gallicum | *Phaseolus vulgaris* |
| | bv. phaseoli | *Phaseolus vulgaris* |
| | *galegae* | |
| | bv. officinalis | *Galega officinalis* |
| | bv. orientalis | *Galega orientalis* |
| | *giardinii* | |
| | bv. giardinii | *Phaseolus vulgaris* |
| | bv. phaseoli | *Phaseolus vulgaris* |
| | *hainanense* | *Desmodium sinuatum* |
| | *huautlense* | *Sesbania herbacea* |
| | *leguminosarum* | |
| | bv. phaseoli | *Phaseolus vulgaris* |
| | bv. trifolii | *Trifolium pratense* |
| | bv. viciae | *Pisum sativum* |
| | *mongolense* | *Medicago ruthenica* |
| | *tropici* | *Phaseolus vulgaris*. This is a complex species, with several types within it |
| | *yanglingense* | *Gueldenstaedtia multiflora* |
| *Sinorhizobium* | *arboris* | *Acacia senegal* |
| | *fredii* | *Glycine max* |
| | *kostiense* | *Acacia senegal* |
| | *medicae* | *Medicago truncatula* |
| | *meliloti* | *Medicago sativa* |
| | *saheli* | *Sesbania pachycarpa* |
| | *terangae* | |
| | bv. acaciae | *Acacia laeta* |
| | bv. sesbaniae | *Sesbania rostrata* roots |

of aerenchyma, as in a number of wetland legume genera.

Within the two groups of nodulated angiosperms, individual genera and species are nodulated by different strains of either *Frankia* or rhizobia, in other words there is specificity between the partners. In order for nodulation to proceed, the endophytes must recognize and be recognized by their hosts, gain entry into roots (occasionally stems in legumes), by-pass the host's defense reaction, and stimulate production of nodule tissue. The recognition process involves a series of exchanges of key molecules produced by the host and the endophyte, following which various nodulation genes in each partner are induced. Some cell division in host roots often starts toward the end of this recognition pathway. Entry may be gained via root hairs, as in many legumes and actinorhizal genera such as *Alnus* or by penetration between epidermal cells or at the point of emergence of lateral roots (adventitious roots in stem-nodulating legumes). Subsequent to the initial infection, nodule development in legumes and actinorhizal plants proceeds in different ways. The common feature of all actinorhizal nodules and those induced by rhizobia on *Parasponia* is that the vascular system is central and the infected tissue cortical. In legume nodules the vascular system is cortical and the infected tissue is central. Even so, there is increasing evidence that similar nodulation genes may be involved in both types, consistent with a common predisposition for nodulation.

### Actinorhizal Plants

Plants nodulated by *Frankia* are collectively referred to as actinorhizal. Technical difficulties, coupled with a lack of funding for nonagricultural species has resulted in research on these plants lagging behind that on nodulated legumes. However, progress is now rapid and the important role of these plants is increasingly being recognized. Table 2 lists the currently agreed actinorhizal genera, and this list is unlikely to be radically altered, as close relatives of the nodulating genera have been extensively examined in almost all areas of the world. However, there have been isolated (unconfirmed) reports of nodules on members of the familes Cucurbitaceae and Zygophyllaceae, both of which are in the Rosid I clade. Molecular evidence has grouped actinorhizal plants in a way that fits well with the known variations in infection pathways and nodule characters, as can be seen in Table 2. All nodules have indeterminate growth (Figure 2), most are perennial, and older parts may become very woody, with the active regions confined to the tips of nodule lobes. It is now generally agreed that nodulation with *Frankia* has evolved on a number of different occasions. One question that is now being addressed is whether or not there is a relationship between particular host plants and the type of *Frankia* with which they nodulate. Although host–endophyte specificity has been shown in a number of cases, it varies widely. On balance it is probably true to say that there is more evidence for coevolution between host plants and *Frankia* than there is for legumes and rhizobia.

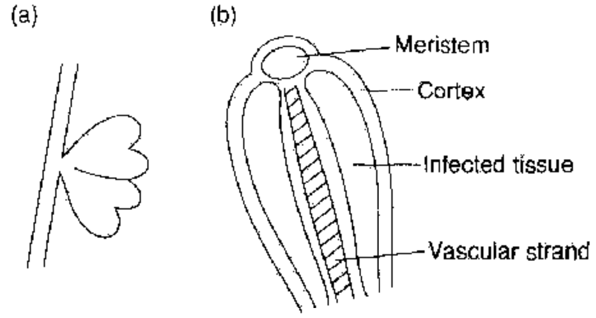


**Figure 2** Actinorhizal nodules. (a) A young nodule as found in *Alnus*. Nodules branch repeatedly, forming a woody coralloid structure up to 8–10 cm in diameter; (b) longitudinal section through the tip of a lobe of a generalized nodule; in some genera this also produces a nodule rootlet (**Table 2**). The detailed structure of the infected tissue varies greatly with genus, and not all nodules have lobes which are symmetrical in longitudinal section. © J Sprent.

Growing *Frankia* in pure culture from crushed nodules has proved a difficult task and one that has still not been achieved for all host genera. Molecular methods have enabled more detailed studies of *Frankia* in soil to be undertaken and it now seems likely that, as with rhizobia, it has a saprophytic stage that may tide it over during periods when no compatible host is present. Being filamentous, *Frankia* has the potential for differentiation. This may occur in the production of vesicles and of spores. Vesicles are usually swollen structures formed at the ends of filaments under conditions of low oxygen concentration. Where present they are the site of nitrogenase activity and the thickness of their lipid-based walls is adjusted to prevent oxygen reaching the enzyme. The location of vesicles within the infected host cells varies considerably between genera. Some strains of *Frankia* form spores within nodules; others do not. Evidence that spores are effective perennating structures is not strong and spore-positive nodules often show lower nitrogenase activity.

Actinorhizal plants are mainly woody in nature, most being shrubs or trees. They are found worldwide from the arctic (*Dryas*, an unusual genus in that it nodulates in some areas but not others) to the tropics (members of the Casuarinaceae), on high mountains (Himalayan alder) and in lowland bogs (*Myrica gale*). The major type of vegetation from which they seem to be absent is tropical rain forest. Many are drought-resistant (*Cercocarpus* and other shrubs from the rain shadow of the US Rocky Mountains), a factor often coupled with saline tolerance (some species of *Casuarina*). They are used in a variety of ways, including forestry (both as a nitrogen input, e.g., *Ceanothus* and as a source of timber, e.g., *Alnus*, *Casuarina*), for land reclamation (*Casuarina*) for animal browse (e.g., *Purshia*), as an agricultural crop (*Hippophäe*), and for pharmaceutical purposes (*Myrica*). The amounts of nitrogen fixed when they are growing in their natural environment or when planted is related to soil and climatic conditions, but is similar in magnitude to that of legumes in similar environments (up to approximately 100 kg N ha$^{-1}$ per year). As with legumes, optimum nitrogen fixation may need formation of other structures such as endo- and/or ectomycorrhizas and cluster (proteoid) roots.

### Legumes

The Leguminosae, also known as the Fabaceae, is the third-largest family of flowering plants. However, only a small number have been exploited by humans, and our knowledge of nitrogen fixation (if any) by most of them ranges from nonexistent to sparse. A great deal of effort has gone into describing legumes from around the world, but not many plant taxonomists look beneath their feet, generally using only characters such as flowers, fruit, and habit that are visible above ground. Most taxonomists agree that the family should be divided into three subfamilies, Caesalpinioideae, Mimosoideae, and Papilionoideae, with a possible fourth currently being considered. It is also widely understood that nodulation is rare in the first, common in the second, and virtually always present in the third. Recent work has not altered this understanding (**Table 4**), but has identified some closely related groups, mostly trees, from each subfamily that lack the ability to nodulate. All known naturally nonnodulating legumes are from tropical and subtropical regions, although some are cultivated at higher latitudes. Their distribution within the family is consistent with the suggestion that nodulation in legumes arose at least twice, possibly three or four times.

Legume nodules are found in a variety of shapes, sizes, and structures (**Figure 3**). As with actinorhizal plants, nodule features tend to be uniform with taxonomic groupings. Some have determinate growth, with only a transient meristem; these are relatively short-lived. Most are indeterminate, with a distinct apical meristem, often showing branching. These are usually longer-lived and may be perennial. They are generally better able to withstand adverse conditions, because they can resume growth and produce new nitrogen-fixing tissue after periods of stress such as drought or low temperature. Studies with agricultural species such as clovers, alfalfa (lucerne), and beans generally show that nodulated plants are less able to withstand environmental extremes than the same species grown in soil containing combined nitrogen. Many commercial cultivars respond well to 'starter nitrogen,' i.e., a small dose of nitrogen fertilizer at planting. To some extent this reflects the fact that until recently most plant breeders raised crops on very fertile soils and thus may have been inadvertently selecting against efficient nitrogen fixation. Around the beginning of the twenty-first century, there was a major effort to raise legume cultivars that would produce economic yields on less-fertile soils and under environmental stresses such as drought and salinity. However, it is also important to go back to the field and look for wild legumes that grow well in extreme environments (**Table 5**). The only conditions where nodulated legumes do not seem to occur naturally is in sea water, although some may grow well in saline areas and others submerged in fresh water. In human terms, perhaps even more importantly, it is now being recognized that there are many nodulated legumes that have been used by native peoples for millennia that can provide food often under very adverse conditions. Early settlers in the USA made great use of tubers of *Apios americana*, a species used by American Indians. In dryland Africa, many tubers are used for food and even water under dry conditions. Whilst most of these

**Table 4** Genera in legume subfamilies which can and cannot nodulate and for which there is no information. Since legume taxonomy is currently undergoing many rearrangements, all numbers must be regarded as very approximate. Many of the genera for which there are no reports are monotypic. Only three genera are known so far in which there is good evidence for both nodulating and nonnodulating species, *Acacia* and *Pentaclethra* (Mimosoideae), and *Pterocarpus* (Papilionoideae)

| | *Number of genera* | | | |
|---|---|---|---|---|
| *Subfamily* | *Total* | *Nodulated* | *Nonnodulated* | *Unknown* |
| Caesalpinioideae | 157 | 8 | 72 | 76 |
| Mimosoideae | 78 | 38 | 9 | 31 |
| Papilionoideae | 479 | 353 | 27 | 99 |

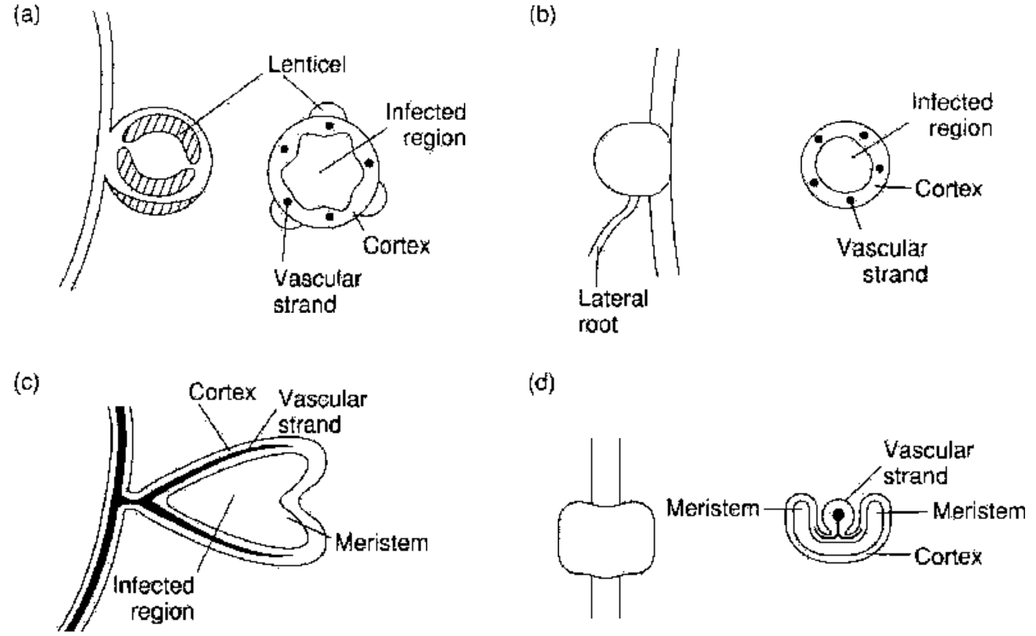


**Figure 3** Legume nodules, drawn from field material, and showing the major types of nodule formed. (a, b, d) Left, whole nodules; right, transverse sections through these nodules; (c) a longitudinal section. (a) Determinate nodule of the desmodioid type, as found in soybean and related genera. Size varies from 2–7 mm in diameter. Lenticels on the surface facilitate gaseous exchange; (b) determinate nodule of the aeschynomenoid type, always associated with lateral or adventitious roots. They are found in many genera, including peanut and important tree genera such as *Dalbergia* and *Pterocarpus*, and on stems of some legumes, such as certain species of *Aeschynomene*. Size is usually between 2 and 6 mm in diameter. The infected tissue lacks the uninfected cells present in desmodioid nodules; (c) indeterminate nodule, drawn from a species of *Acacia*. This is the commonest type of nodule and occurs in all three subfamilies. It is characterized by an apical meristem, and the infected tissue has a gradation from young, inactive cells through active, nitrogen-fixing cells to (in older nodules) an inactive senescent region. They vary greatly in ratio of length to breadth and in the degree of branching (if any). Size varies from 2–3 mm in length to complex coralloid structures which may be more than 5 cm in diameter; (d) lupinoid nodules. These are not common, but are typical of many species of *Lupinus*. They have a lateral meristem which gradually encircles the root, forming a collar. As with other indeterminate nodules, there is a gradient of cells in the infected region. © J Sprent.

**Table 5** Some environmental constraints to nitrogen fixation by legumes used in agriculture and examples of wild legumes which appear adapted to such conditions

| *Constraint* | *Wild legumes showing tolerance* |
|---|---|
| Low temperature | *Astragalus alpinus*, *Hedysarum arcticum*, *Oxytropis sordida* |
| Shade | Some *Inga* spp., *Ormosia countinho*, *Poecilanthe hostmannii* |
| Flooding | *Discolobium pulchellum* (requires flooding for nodulation), some *Astragalus* spp., some *Neptunia* spp. |
| Salinity | *Halimodendron halodendron*, *Prosopis tamarugo*, *Vigna marina* |
| Drought | Some *Acacia* spp., *Faidherbia albida*, some *Vigna* spp. |
| Low available P | Some *Acacia* spp., *Crotalaria grahamiana*, *Cajanus cajan* |
| Low soil pH | Some *Astragalus* spp., some *Lupinus* spp. |
| High soil Al | *Dioclea guianensis*, some *Stylosanthes* spp. |

can nodulate, there are exceptions, such as the marama bean, *Tylosema esculentum*. If 'old' new crops are to be sustainably exploited, it is essential to know: First, can it nodulate; and, second, will it nodulate under the conditions under test?

The soil contains a large number of bacterial genera known to nodulate legumes (Table 3). What is not clear from Table 3 is that one strain of rhizobia may nodulate one or many hosts, and that some hosts are nodulated by one strain, others by many. These variations raise a number of theoretical and practical questions. Whereas early studies placed great emphasis on host–rhizobial specificity, assuming that the two partners had coevolved, this is now much less clear. Most of the early work on legume nitrogen fixation was carried out on temperate crop plants and in these the 'rules' usually apply. The main exception is the common bean *Phaseolus vulgaris*, now known to nodulate with several different rhizobial genera (Table 3). Since about 1989, there has been a major thrust toward examining rhizobia from tropical areas. This is when the rules start to break down. New phenomena such as stem-nodulating legumes, photosynthetic rhizobia, new combinations of growth rate and acid/alkali production, and many more have been discovered. Many examples of broad host-range rhizobia have been discovered, adding to the classic one, originally referred to as the 'cowpea miscellany.' At that stage, broad host range and slow growth were thought to be associated, but there is now a fast-growing strain, NGR 234, isolated from Papua New Guinea, that can nodulate at least 70 genera, including some from each of the three legume subfamilies. This and other findings have challenged the idea of coevolution. We now know that clusters of bacterial genes, called 'symbiotic islands,' are capable of lateral transfer in the soil between genera, and this may account for the growing number of new rhizobial genera which are being discovered. It may also help to explain the close relationship between some rhizobia and plant and animal pathogens.

If it is possible to make any generalizations about nodulated legumes, it may be that, when radiating into higher latitudes, legumes tend to be herbaceous, always to nodulate, and to have a recognizable level of host–rhizobial specificity. Nodules are also usually indeterminate, a benefit in areas with a markedly seasonal climate. In low latitudes, there are many nonnodulating legumes and, of the nodulating ones, many are nodulated by a wide range of rhizobia, producing nodules that vary greatly in their ability to fix nitrogen (effectivity). From the bacterial perspective, infectivity may be more important than effectivity.

## The Importance of Nitrogen-Fixing Symbioses to Soil Fertility

One of the problems with trying to estimate biological nitrogen fixation is the lack of suitable methods for field use, especially where trees are involved. As a result, published estimates for global nitrogen fixation vary widely. The true value may lie in the range 100–200 Tg N per year. In recent years, the availability of cheap fertilizer N has reduced both the amount of biologically fixed nitrogen in agriculture and also the amount of field research on the subject in the developed world. However, the rise in the organic movement and fears about nitrate pollution have provided a stimulus for renewed study. Developing countries have no doubt about the significance of nitrogen fixation for the well-being of their peoples and their economies. It is now very clear that optimization of nitrogen fixation in these countries requires much more research, as many of the procedures used in developed countries are inappropriate. To take a simple example, the diversity of rhizobia found in tropical soils poses major problems for the production of effective, competitive inoculant strains.

What is the fate of nitrogen fixed by symbiotic systems? Initially it is to supply the partners, but in the real world plants do not live on their own. We have little idea of how much, if any, of nitrogen fixed by plants that are also mycorrhizal (which most, including bryophytes, are) is exchanged for other nutrients (phosphorus, fixed carbon, trace elements) with nonfixing plants via hyphal connections. If this occurs, then selective harvesting of, for example, large nodulated trees could have a major effect on other plants in the ecosystem. Harvesting crops can also lead to loss of soil N if the amount removed in the crop exceeds that fixed. Even the benefits of crop rotation have been reassessed. A concomitant of nitrogen reduction by nitrogenase is the simultaneous production of hydrogen.) Although the energy so wasted may be partly recovered in some systems by an uptake of hydrogenase, there is little evidence that plants lacking such a system are at a disadvantage. A recent suggestion is that hydrogen evolution in the soil alters the microbial flora in such a way that plants may benefit from improved nutrition. This could help to explain some of the benefits of crop rotation which cannot be accounted for by improved soil fertility and/or reduction of pest and pathogen load.

*See also:* **Nitrogen in Soils:** Plant Uptake

## Further Reading

Allen ON and Allen EK (1981) *The Leguminosae: A Source Book of Characteristics, Uses and Nodulation.* Madison: University of Wisconsin Press/and London, UK: Macmillan.

Benson DR and Silvester WB (1993) Biology of *Frankia* strains, actinomycete symbionts of actinorhizal plants. *Microbiological Reviews* 57: 293–319.

Berg RH, Langwenstein B, and Silvester WB (1999) Development in the *Datisca–Coriaria* nodule type. *Canadian Journal of Botany* 77: 1334–1350.

Bergman B, Johansson C, and Söderbäck E (1992) The *Nostoc–Gunnera* symbiosis. *New Phytologist* 122: 379–400.

Cleveland CC, Townsend AR, Schimel DS *et al.* (1999) Global patterns of terrestrial biological nitrogen ($N_2$) fixation in natural ecosystems. *Global Biogeochemical Cycles* 13: 623–645.

Giller KE (2001) *Nitrogen Fixation in Tropical Cropping Systems*, 2nd edn. Wallingford, UK: CABI.

Huss-Danell K (1997) Actinorhizal symbioses and their $N_2$ fixation. *New Phytologist* 136: 375–405.

Jones DL (1993) *Cycads of the World.* Washington, DC: Smithsonian Institution Press.

Rai AN, Söderbäck E, and Bergman B (2000) Cyanobacterium-plant symbioses. *New Phytologist* 147: 449–481.

Soltis DE, Soltis PS, Chase ME *et al.* (2000) Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atp*B sequences. *Botanical Journal of the Linnean Society* 133: 381–461.

Spaink HP, Kondorosi A, and Hooykaas PJJ (eds) (1998) *The Rhizobiaceae – Molecular Biology of the Model Plant-Associated Bacteria.* Dordrecht, the Netherlands: Kluwer Academic.

Sprent JI (2001) *Nodulation in Legumes.* Kew, UK: Royal Botanic Gardens.

Sprent JI (2002) Knobs, knots and nodules. *New Phytologist* 153: 2–7.

Sprent JI and Sprent P (1990) *Nitrogen-Fixing Organisms: Pure and Applied Aspects.* London, UK: Chapman and Hall.

Wagner GM (1997) *Azolla*: a review of its biology and utilization. *Botanical Review* 63: 1–26.

Wall LG (2000) The actinorhizal symbiosis. *Journal of Plant Growth Regulation* 19: 167–182.

**Nitrogen Fertilizers** *See* **Fertilizers and Fertilization**

# NUCLEAR WASTE DISPOSAL

**G W Gee, P D Meyer, and A L Ward**, Pacific Northwest National Laboratory, Richland, WA, USA

## Introduction

Nuclear wastes are by-products of nuclear weapons production and nuclear power generation, plus residuals of radioactive materials used by industry, medicine, agriculture, and academia. Their distinctive nature and potential hazard make nuclear wastes not only the most dangerous waste ever created by humanity, but also one of the most controversial and regulated with respect to disposal. Nuclear waste issues, related to uncertainties in geologic disposal and long-term protection, combined with potential misuse by terrorist groups, have created uneasiness and fear in the general public and remain stumbling blocks for further development of a nuclear industry in a world that may soon be facing a global energy crisis.

## Categories

### High-Level Waste

The category 'high-level waste' (HLW) consists of either used (spent) nuclear fuel or the waste generated by reprocessing spent nuclear fuel. HLW generates much heat and radiation and requires remote handling and significant shielding, using materials such as lead, concrete, soil, or water, to prevent excessive radiation exposure to workers. Spent-fuel waste consists of fuel elements removed primarily from commercial reactors, after 3 or 4 years, when their useful life is over and they no longer generate electricity efficiently. The spent fuel assemblies from nuclear power plants are stored on-site, where they are typically submerged in large, water-filled pools, awaiting final disposal. In contrast, reprocessing wastes are stored mainly at nuclear weapons production facilities in liquid form in large underground storage tanks. **Figure 1** shows construction of large (1 million gallons, approx. 3.6 million liters) storage tanks buried at the Hanford Site in Washington state, USA. In some cases, the liquid wastes have been calcined into a dry granular form. In the USA, all HLW liquids will be solidified (vitrified) and shipped in secure canisters to a deep geologic repository. Similar plans exist for other countries.

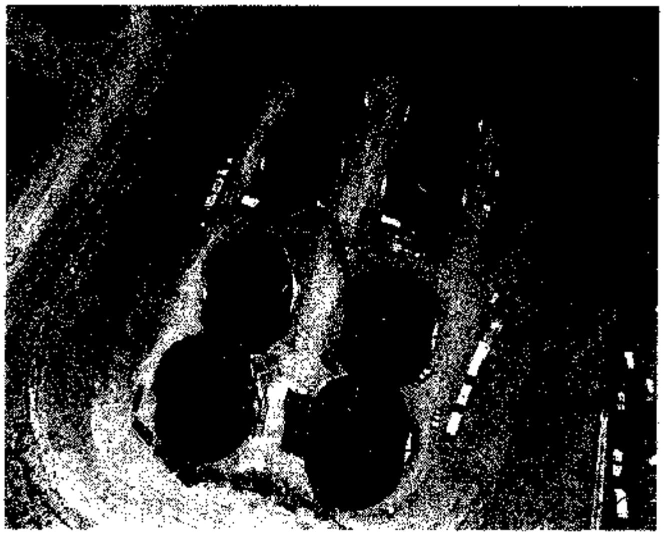


**Figure 1** Underground million-gallon (3.6 million liter) storage tanks for nuclear waste at the Hanford site in Washington State, USA.

### Transuranic Waste

Transuranic waste (TRU) is derived from the reprocessing of spent fuel and the use of plutonium in fabricating nuclear weapons. TRU is waste contaminated with alpha-emitting radionuclides of atomic number greater than 92 (i.e., greater than uranium, hence the term 'transuranic') and half-lives greater than 20 years in concentrations greater than 3700 Bq g$^{-1}$. TRU typically generates less heat and emits less radiation than HLW, but it remains toxic for centuries and thus requires the same long-term isolation as HLW. TRU in the USA is stored at the Waste Isolation Pilot Plant near Carlsbad, New Mexico.

### Low-Level Waste

The category 'low-level waste' (LLW) includes all radioactive waste other than HLW and TRU. Uranium mill tailings, a special form of LLW, are discussed separately. Most LLW contains relatively short-lived nuclides and has less radioactivity than found in HLW. LLW comes from a wide range of institutions, including nuclear power plants, government laboratories, test reactors, hospitals, laboratories, and industrial plants. LLW takes a wide variety of forms, including materials from medical research and treatments; contaminated rags and protective clothing; paper towels, filters and filter residues; and hand tools and equipment. Radiation from LLW can be high enough to require shielding for handling and shipping. **Figure 2** show disposal of LLW in concrete (shielded) containers at the Hanford LLW disposal site, USA.

LLW is classified into four groups according to degree of hazard. LLWs that can be disposed of by near-surface burial are classed as A, B, and C types, ranging from the lowest hazard (A) to the highest (C). The US Nuclear Regulatory Commission (NRC) provides requirements for packaging and the form of burial for each of these classes. LLW too hazardous to bury in near-surface facilities is termed 'Greater-Than-Class-C' waste and requires disposal in a deep geologic repository. LLW also includes radioactive materials from natural sources, including radium-226, radium-228, radon-222, plus other radioactive elements from terrestrial sources. Discrete and diffuse sources are part of the LLW inventory. Discrete waste sources include spent industrial gauges, radium needles (used in medical equipment), and exchange resins (e.g., those that separate radium and other radionuclides from contaminated waters). Diffuse LLW can include such things as coal ash, phosphate waste from fertilizer production, and contaminated water and drilling equipment from oil and gas production. Some LLW has been reused for construction backfill and road-building activities and thus spread over long distances. While these radiation levels are relatively low, they produce conditions that can be a detriment to human health. As an example, uranium tailings high in radium can produce significant levels of elevated radon-222 gas. People in homes with basements surrounded by tailings or other radium sources can be exposed to radon gas levels that exceed permissible limits (e.g., 0.74 Bq m$^{-2}$ s$^{-1}$).

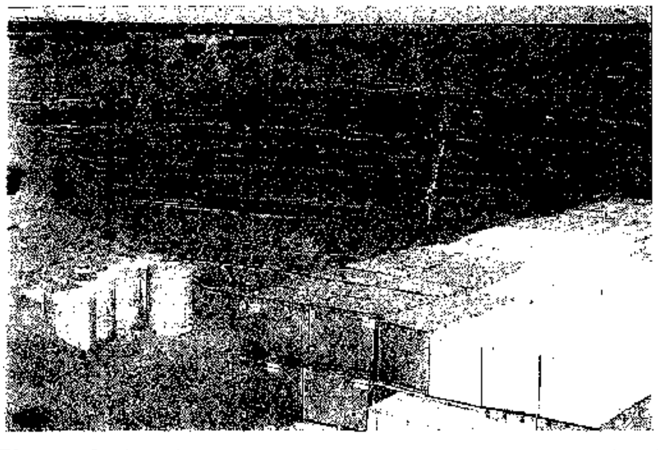


**Figure 2** Low-level waste storage at the Hanford low-level waste facility in Washington State, USA.

### Uranium Mill Tailings

This special category of LLW consists of residues in the form of fine sands and silts generated from the mining and milling of uranium ore. Large volumes of uranium tailings exist (more than 118 million cubic meters in the USA alone) containing low concentrations of naturally occurring radioactive materials, including thorium-232, radium-226, and its decay product, radon-222. Radon gas is the largest single radiation hazard from uranium mill tailings. Soil covers placed over uranium tailings piles have generally reduced the probability of radon escaping from them. Other environmental concerns include elevated uranium, sulfate, nitrate, and ammonia concentrations in groundwater beneath tailings piles.

### Mixed Waste

Waste that contains both hazardous chemicals and radioactive wastes (generally those in the LLW category) is termed 'mixed waste.' In the USA, mixed waste is subject to both NRC and Environmental Protection Agency (EPA) regulations. Sites where chemical contaminants such as dense, nonaqueous-phase liquids (DNAPLs, e.g., carbon tetrachloride), or heavy metals (e.g., chromium, lead, mercury) are mixed with radionuclides are more difficult to clean up than sites where radionuclides are the sole contaminant, because of the complex nature of the mixed waste coupled with the heterogeneous nature of the subsurface soils and sediments in which the wastes reside.

## Waste Inventories and Environmental Discharges

Since the end of World War II, large amounts of nuclear wastes have accumulated in temporary waste repositories around the world. **Table 1** shows the volumes of HLW, TRU, and LLW that have accumulated in the past and are projected to accumulate in the USA by the year 2030. Uranium mill tailings are also shown for comparison. In terms of shear volume, uranium mill tailings contribute by far the largest volume of waste (over 157 million cubic meters), with LLW contributing approximately one-fifth of the total volume (or nearly 38 million cubic meters of LLW). Nearly double the USA volumes have been generated in the former Soviet Union (FSU), with lesser volumes in other countries. Decay of radionuclides reduces the massive amounts of nuclear waste but does not eliminate the hazard, because long-lived radionuclides dominate in later years. As shown, the hazard of radioactive waste remains in perpetuity, so treatment or indefinite containment strategies are needed.

In addition to stored wastes, considerable volumes of wastes have been previously discharged directly to the environment. These environmental insults range from approximately $1.4 \times 10^{19}$ Bq (390 million curies [where 1 Ci $= 3.7 \times 10^{10}$ radioactive disintegrations per second]) in the USA (primarily from liquid discharges during weapons production) to more than $6.3 \times 10^{19}$ Bq (1700 million curies) in the FSU. At nuclear facilities, the discharges have been directly to the soils and adjacent waterways, burdening them with radioactive decay products whose half-lives range from a few years to millions of years. Long-term environmental consequences from these waste discharges focus primarily on inventories of carbon-14 (half-life of nearly 6000 years), technetium-99 (Tc-99) (half-life of 213 000 years) and iodine-129 (I-129) (half-life of nearly 16 million years). The chemical forms in which Tc-99 and I-129 are present are not strongly sorbed in most sediments so they tend to move to groundwater. Concentrations in excess of $7400\,\mathrm{Bq\,l^{-1}}$ of Tc-99 (more than 100 times permissible limits) have been found in the deep groundwater under waste tanks at the US Department of Energy (DOE) Hanford Site in Richland, WA, USA.

**Table 1** Current and projected volumes of various types of nuclear waste in the USA

| | *Volume of waste ($\times 10^3\,m^3$)* | | |
|---|---|---|---|
| *Types of waste* | *Year 2000* | *Year 2010* | *Year 2030* |
| *High-level waste* | | | |
| Commercial-liquid wastes | 0.24 | 0.24 | 0.24 |
| Commercial-spent fuel | 17 | 24 | 35 |
| Defense (total) | 395 | 335 | 346 |
| *Transuranic* | | | |
| Defense/environmental cleanup | 275 | 297 | 191 |
| *Low-level waste* | | | |
| Commercial | 1722 | 2055 | 2508 |
| Defense/decommissioning/ cleanup | 4707 | 22 777 | 36 524 |
| *Uranium mill tailings* | | | |
| Commercial and defense wastes | 152 400 | 155 400 | 157 400 |

## Radiation Exposures

Figure 3 shows the major pathways through which radioactivity from nuclear sources can affect living organisms. While the impact on the environment has been significant, immediate health effects to the general population have been small. Exposure to materials from the nuclear fuel cycle (including nuclear waste) is less than 0.2% of the total annual radiation dose to the general population. In contrast, dental and medical diagnosis and therapy account for more than 90% of all doses from nonnatural sources. Some occupations and geographic settings expose an individual to a greater-than-average radiation dose. A person living in Denver, CO, USA (1.5 km above sea level) is exposed to nearly twice as much cosmic radiation as a person living at sea level in San Diego, CA, USA. Individuals living in Spokane, WA, USA are exposed to higher concentrations of radon from the soil than those living in Midland, TX, USA. The question for many is that, while radiation doses to the general public are currently small, as nuclear waste discharges continue and dispersal to the environment proceeds, what future risks to the human population can be expected? What are the long-term risks? What can be done to ensure the risks are minimal and that, in the immediate and foreseeable future, the human race is protected against undue and unnecessary health risk from nuclear waste?

## Waste Storage and Disposition

Most HLW, including spent fuel rods, and reprocessing wastes are in temporary storage (**Table 2**). No permanent repository for HLW has been licenced anywhere in the world to date (2004). Spent fuel rods are largely contained at sites near nuclear reactors. These

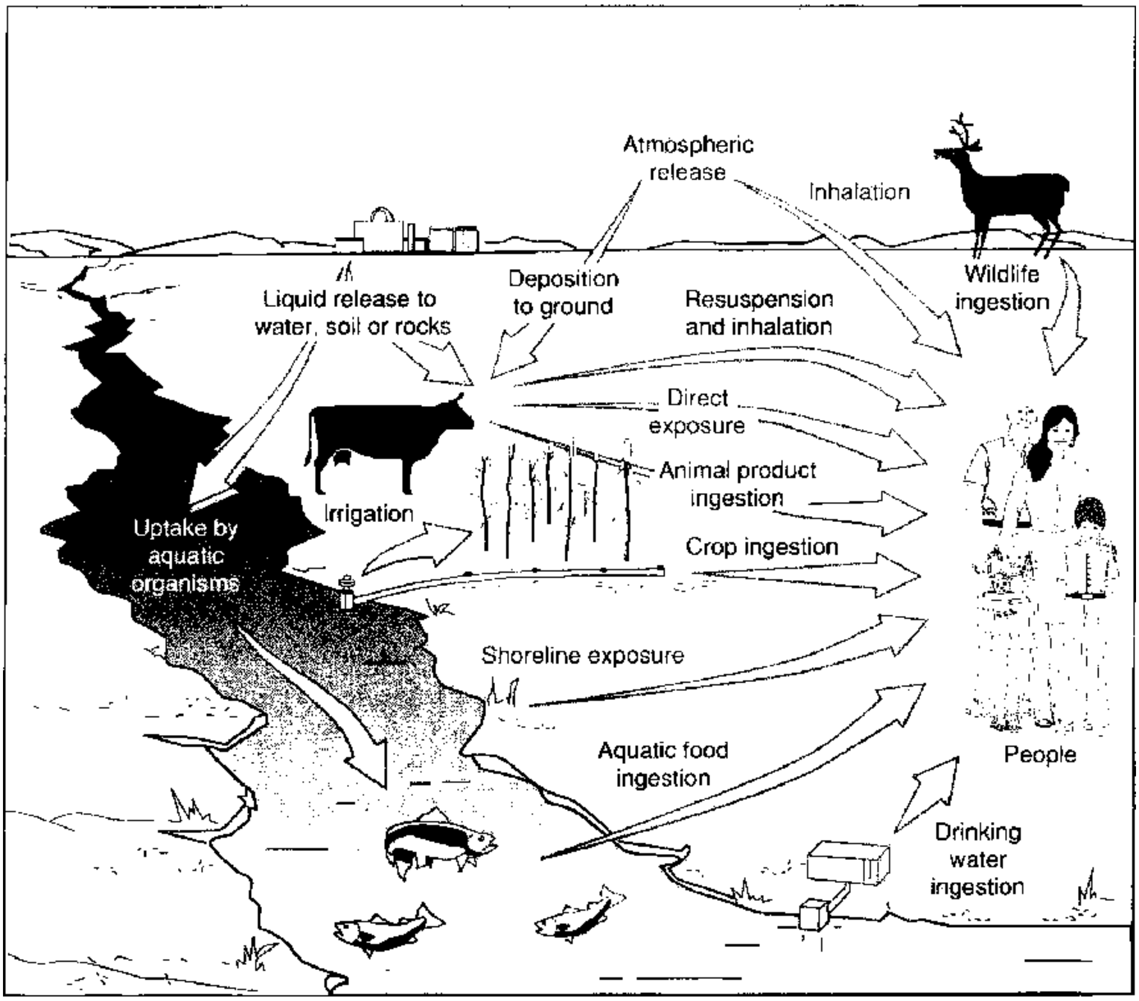


**Figure 3** Potential radionuclide pathways from nuclear wastes.

**Table 2** High-level waste burial programs in various countries

| *Country* | *Earliest planned year* | *State of program* |
|---|---|---|
| Belgium | 2020 | Underground disposal in clay at Mol, Belgium |
| Canada | 2025 | Bury irradiated fuel in granite at yet-to-be identified site |
| China | Undetermined | Fuel to be reprocessed. Burial planned for Gobi Desert |
| Finland | 2020 | Field studies conducted – final site selection not made |
| France | 2010 | Two sites selected and studied with final selection in 2006 |
| Germany | 2008 | Gorleben salt dome site only site studied |
| India | 2010 | Fuel to be reprocessed – granite repository to be selected |
| Italy | 2040 | Fuel to be reprocessed – storage for 50 years before burial in clay or granite |
| Japan | 2020 | Cooperative study at a Chinese research facility |
| Netherlands | 2040 | Interim storage for 50–100 years before burial sub-seabed or exterior |
| Russia | Undetermined | Uncertain |
| Spain | 2020 | Burial in clay, granite, or salt formation |
| Sweden | 2020 | Aspo granite site selected near Oskarshamn nuclear facility |
| Switzerland | 2020 | Burial in granite or sedimentary formation to be selected |
| United States of America | 2010 | Yucca Mountain – deep vadose zone, federally recommended site |
| United Kingdom (Great Britain) | 2030 | Interim storage approved through 2022. Options include sub-seabed burial |

fuel rods pose a threat to humanity because they can be reprocessed for plutonium, a key component of nuclear weapons, so permanent disposal is highly desirable. Many billions of dollars have been spent by the US government on characterizing Yucca Mountain as a potential repository for more than 70 000 $m^3$ of nuclear wastes, including spent fuel rods. The US Department of Energy is currently proceeding with a license application to store nuclear wastes at the Yucca Mountain site. Should a license be approved, no nuclear waste will be stored there until well after 2010. Issues of transportation, monitoring, and final waste configurations remain to be resolved. In the interim, most of the spent fuel remains in relatively vulnerable locations, often near large urban centers where nuclear terrorism may be difficult to prevent and where consequences of such acts of terrorism could be deadly and have long-range consequences. These issues suggest that action be taken soon to finalize waste repositories at locations that effectively control access to waste, particularly HLW that can be converted into weapons.

After reprocessing wastes are removed from underground storage tanks, some residual contaminants will remain. Several options are being considered for containing the waste and leaving the tanks, contaminated soil, piping, and support structures in place. The following options are being considered:

- Stabilization: tanks would be stabilized to prevent dome collapse by filling with inert material such as bentonite, sand, gravel, cement, or a combination of these materials;
- Immobilization: subsurface injections of either reactive chemicals, to immobilize contaminants and prevent them from moving, or physical barriers (low-permeability grouts, including bentonite and mineral wax), to retard the migration of the contaminants to the water table;
- Decontamination: chemical or mechanical decontamination by using high-pressure water or frozen carbon dioxide blasting, or in-place solvent washes and extractions;
- Flushing: a combination of a subsurface barrier and a soil-flushing treatment where chemicals such as carbonate solutions would extract contaminants. Subsurface barriers made of grout or polymers would minimize the chance of flushing the contaminants and mixing with groundwater;
- *In-situ* vitrification: this process uses heated electrodes to melt the waste in place. The waste and surrounding soil are vitrified (turned into glass) at temperatures exceeding 1500°C and when cooled resemble natural glass obsidian. Volatile organics would be destroyed and heavy-metal contaminants and radionuclides would be physically bound in the glass. This process has been successfully demonstrated for LLW and mixed wastes at a number of locations throughout the world;
- Surface barriers: after the empty tanks are treated in place, aboveground barriers could be placed over the tanks to protect the area against intrusion, wind and water erosion, and net water infiltration. If successful, such efforts would ensure that any residual contaminants not fully treated would tend to remain in the subsurface for thousands of years. Long-term surface barriers (intended to have a lifetime of at least 1000 years) consisting of specially designed soil layers have been tested and show promise for deployment at the Hanford Site in the USA. **Figure 4** shows a schematic picture of a backfilled storage tank covered by a 1000-year soil barrier.

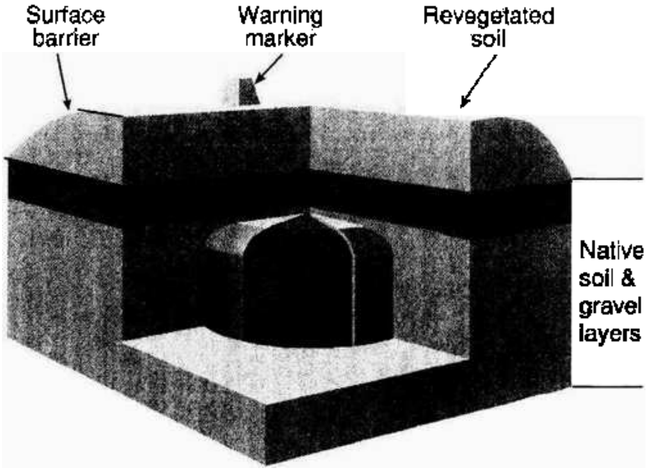


**Figure 4** Schematic of a surface barrier consisting of natural earthen materials placed over an evacuated and stabilized high-level waste tank. The surface barrier is designed to last at least 1000 years and protect against plant and animal intrusion, wind and water erosion, and water infiltration.

LLW disposal actions remain site-specific. Under present regulatory guidance, there will be minimal waste treatment, with site closures depending on cover systems that can provide control of erosion, intrusion, and minimization of water invasion into the waste. Longevity for such cover systems remains largely unproven, since most conventional landfill designs are built to last decades not millennia. Efforts to develop innovative and scientifically sound, long-term performance tests are still in their infancy at the beginning of the twenty-first century.

## Environmental Uncertainties

One of the characteristics of nuclear wastes is their potential to cause adverse health and environmental effects for many thousands of years. For nuclear waste disposal facilities and sites contaminated with radioactive materials, there is consequently a need to assess safety far into the future. In the USA, regulatory limits for the disposal of HLW and TRU are required to be met over a 10 000-year time period; a 1000-year period is used in analyses for decommissioning nuclear materials sites. Decisions that are intended to apply that far into the future rely, in large part, on models that predict radionuclide transport in the environment. Because such predictions are inherently uncertain, a variety of strategies have been adopted to provide confidence in decisions about nuclear waste disposal:

- Reliance on natural systems to limit transport (e.g., stable rock formations, arid sites);
- Redundant design of engineered components (to prevent catastrophic behavior if one component fails);
- Site-specific characterization of processes potentially affecting transport (to provide understanding of site behavior and to develop and parameterize models);
- Natural analog studies (e.g., of uranium transport);
- Accelerated degradation studies (e.g., of glass waste forms);
- Peer review;
- Monitoring of site behavior and comparison with model predictions;
- Sensitivity and uncertainty analyses of predictive models.

The extent of analyses conducted to support decisions on nuclear waste disposal varies significantly. In the USA, the disposal of HLW at Yucca Mountain, a remote site in southern Nevada, USA, and the disposal of TRU at the Waste Isolation Pilot Plant have been supported by large budgets, detailed site-characterization studies, complex modeling analyses, and extensive site monitoring. In contrast, decommissioning of nuclear materials sites may rely on very limited, site-specific data and relatively simplified modeling analyses. Regardless of the level of effort, however, predictive modeling of radionuclide transport in soils and sediments plays an important role. Assessing the uncertainty inherent in model predictions supports decision-making by increasing the credibility of the modeling analyses, improving the predictability of consequences, avoiding worst-case assumptions, and focusing data collection on variables most valuable in improving predictions and reducing uncertainty.

The sources of uncertainty that potentially have an impact on predictions of radionuclide transport in the environment are many. In a general sense they arise from:

- Measurement or sampling errors in characterizing the system;
- Spatial and temporal variability in the system's properties;
- Disparity among the scale of the system's features, events, and processes and the scales at which sampling and simulation occur;
- Randomness in the system's stresses, particularly transient external stresses often occurring over a short period of time (e.g., floods, earthquakes);
- Incomplete knowledge of the system being analyzed (i.e., ignorance of significant features, events, or processes).

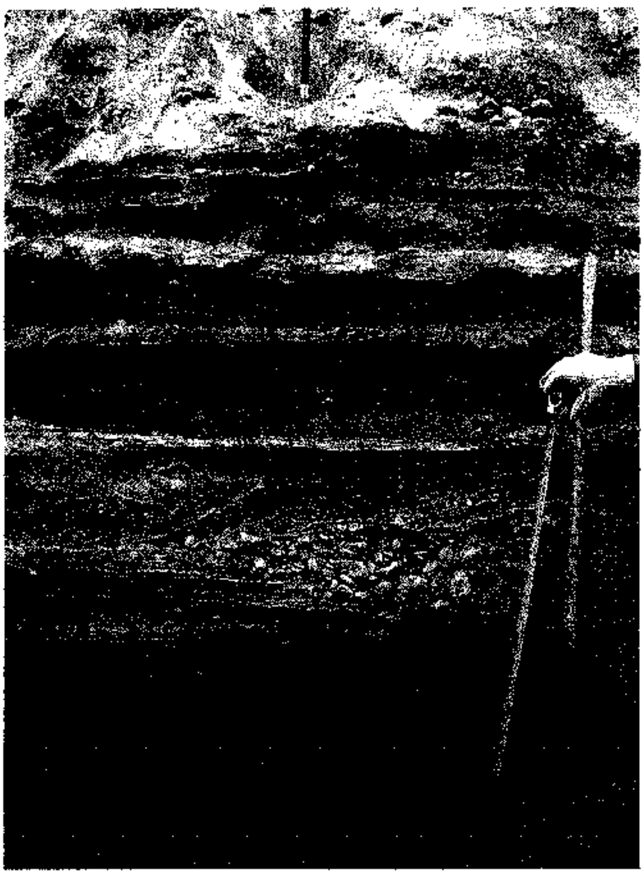


**Figure 5** Exposed trench face of layered, heterogeneous sediments in which nuclear wastes are stored at the Hanford Site, Washington state, USA. (Photo courtesy of John Selker, Oregon State University.)

The issues of variability and scale are demonstrated in **Figure 5**, which shows the spatial variability of sediments on a trench face at a location on the Hanford Site (Washington, USA). Fine, silty layers only a few centimeters thick lie between coarse, sand layers with varying amounts of gravel. Parameters describing flow and transport vary over a wide range at this site (by several orders of magnitude in the case of hydraulic conductivity). A small-scale measurement with a sample size of 30 cm or so (such as a neutron-probe measurement of water content) would probably sample over distinctly different materials. In a model, it would not be unlikely that the area of the entire photograph would be represented by a single parameter value. Uncertainty is thus introduced both in the interpretation of measured quantities and in the assignment of parameter values.

Transport of radioactive contaminants through the vadose zone (i.e., the zone between the ground surface and the water table) depends on the magnitude and variability of geochemical and hydrologic parameters. These parameters include particle-size distribution, specific surface area, cation-exchange capacity, chemical retardation factors, water retention, and unsaturated hydraulic conductivity, all of which control fate and transport of chemicals and radionuclides in the vadose zone, and all have significant levels of uncertainty associated with them at a given waste site. For example, understanding how leaks from HLW tanks move through soils presents a formidable challenge. Wastes are often hot, highly saline, and chemically reactive. Waste plume movement is both spatially and temporally affected by geochemical interactions as well as hydrologic controls. Concentrated water flow from water-line leaks and meteorologic sources (winter rains and snowmelt) can cause contaminants to move in preferential flow paths that may accelerate downward movement. On the other hand, horizontal layering of sediments can cause plumes to spread laterally, delaying their arrival at the groundwater table.

In modeling the performance of nuclear waste disposal sites, it is convenient to classify uncertainties into three basic types: scenario uncertainty; model uncertainty; and parameter uncertainty. Scenario uncertainty addresses the possibility that the future state of the system may be significantly different from what is assumed as the initial case. This may be the result of direct human activity (e.g., irrigation or groundwater extraction) or a more natural event (e.g., increased precipitation due to climate change).

Model uncertainty arises when a complex system is conceptualized for the purpose of modeling and when that concept is implemented as a mathematical model for the purpose of prediction. It is frequently the case that multiple conceptual models can be proposed for a single site to explain observed behavior and that site data are insufficient to determine which model best represents reality. An example is the case of flow and transport in the unsaturated zone at the Yucca Mountain site. Understanding of the relative importance of fractures versus the rock matrix in contributing to seepage and radionuclide transport, and the best way to represent the fracture and/or matrix flow and transport processes, has evolved over time in response to improved analyses and additional site data.

Parameter uncertainty addresses our inability to specify precisely the parameters of the models developed to describe flow and transport of radionuclides in the environment. Even if the models are accurate representations of the actual system processes, it is impossible to measure exhaustively the required model parameters due to their spatial andtemporal variability and apparent randomness. Worse yet, the quantities that are often measured are only indirectly related to model parameters and may be on a scale different (either larger or smaller) from the scale of the model parameters. Model calibration (minimizing the difference between observed system behavior and model-predicted behavior by systematically varying parameter values) is an important method to reduce and quantify parameter uncertainty.

Parameter uncertainty is most amenable to characterization using formal probabilistic methods. In soils, this generally means using measured data and/or expert judgment to specify probability distributions for flow and transport characteristics. Because measurement of these characteristics is frequently expensive and time-consuming, more easily measured soil characteristics are often used to infer distributions of the desired parameters. Examples include soil texture measurements, to infer unsaturated hydraulic properties, and soil mineralogy (along with other quantities), to infer chemical effects on radionuclide transport (i.e., adsorption). Model uncertainty is generally quantified by comparing flow and transport results from alternative models, potentially weighting these results by estimates of the relative probability of the models. Scenario uncertainty can be similarly quantified, although in this case it is possible that absolute probabilities of alternative scenarios can be derived (e.g., the probability of a flood occurring in a given time-period).

In relatively extensive studies of uncertainty (such as for Yucca Mountain and the Waste Isolation Pilot Plant), parameter, model, and scenario uncertainties have all been found to contribute significantly to the overall system uncertainty (e.g., uncertainty in the performance of the nuclear waste disposal facilities). In addition, results of the uncertainty analyses have been used to identify model components and parameters that have the greatest impact on system performance. This information has been used to target data collection to provide the greatest reduction in system uncertainty.

## Summary

Nuclear waste disposal is a global concern. Discharges to the environment have been significant, with radioactive production and subsequent dispersal of billions of curies. To date these waste inventories have been largely confined to locations near nuclear-production facilities primarily in the USA and FSU. Dispersal from global fallout and from nuclear accidents have produced measurable radiation doses to the global population, but compared with natural sources have contributed only a small fraction of total annual dose and, as of 2004, do not present a significant health risk to the general public. However, the public needs to be informed of hazards that remain in those areas where soils and waters have been exposed to nuclear waste discharges and remain highly contaminated.

Safe disposal of nuclear waste is a priority among most nations and will continue to be a major expense for those nations involved in past nuclear weapons production or with past and current nuclear-power generation. Treatment, immobilization, and containment are methods that are being proposed to handle nuclear wastes until they decay to harmless levels. Vitrification is an expensive but viable technology for immobilizing high-level radioactive waste, and deep, geologic storage is a scientifically sound method for ensuring long-term public protection. Low-level radioactive waste disposal in shallow landfills, covered by soil covers, is allowed (in 2004) and should be acceptable if longevity of these disposal facilities and their isolation capabilities can be demonstrated. This remains a scientific challenge.

*See also:* **Isotopes in Soil and Plant Investigations**; **Remediation of Polluted Soils**; **Waste Disposal on Land**: Liquid; Municipal

## Further Reading

Bradley DJ (1997) *Behind the Nuclear Curtain: Radioactive Waste Management in the Former Soviet Union.* Columbus, OH: Battelle Press.

Crowley KD and Ahearne JF (2002) Managing the environmental legacy of US nuclear weapons production. *American Scientist* 90(6): 514–523.

Gee GW, Ward AL, and Wittreich CD (2002) *The Hanford Site 1000-Year Cap Design Test.* PNNL-14143. Richland, WA: Pacific Northwest National Laboratory.

Gephart RE (2003) *Hanford, A Conversation about Nuclear Waste and Cleanup.* Columbus, OH: Battelle Press.

Gephart RE and Lundgren RE (1998) *Hanford Tank Cleanup: A Guide to Understanding the Technical Issues.* Columbus, OH: Battelle Press.

Gerber MS (2002) *On the Home Front, the Cold War Legacy of the Hanford Nuclear Site*, 2nd edn. Lincoln, NE: University of Nebraska Press.

Joint Norwegian–Russian Expert Group (1997) *Sources Contributing to Radioactive Contamination of the Techa River and Areas Surrounding the "Mayak" Production Association, Urals, Russia.* Joint Norwegian–Russian Expert Group for Investigation of Radioactive Contamination in the Northern Areas. Osteras, Norway: Gjovik Trykkeri.

League of Women Voters Education Fund (1993) *The Nuclear Waste Primer.* New York: Lyons and Burford.

Stenhouse MJ and Kirko VI (2001) *Defense Nuclear Waste Disposal in Russia: International Perspective.* Dordrecht, the Netherlands: Kluwer Academic Press.

USDOE (2000) *Pioneering Nuclear Waste Disposal (The Waste Isolation Pilot Plant).* DOE/CAO-00-3124. Carlsbad, New Mexico: US Department of Energy.

USDOE (2002) *Yucca Mountain Science and Engineering Report Review 1.* DOE/RW-0539-1. Las Vegas, NV: US Department of Energy.

Ward AL, Gee GW, and White MD (1997) *A Comprehensive Analysis of Contaminant Transport in the Vadose Zone beneath Tank SX-109.* PNNL-11463. Richland, WA: Pacific Northwest National Laboratory.

# NUTRIENT AVAILABILITY

**N K Fageria**, EMBRAPA, Santo Antônio de Goiás, GO, Brazil
**V C Baligar**, USDA Agricultural Research Service, Beltsville, MD, USA

## Introduction

Agricultural production and productivity are directly linked with nutrient availability. For sustained high crop yields, the application of nutrients is required. The term 'nutrient availability' has been used and defined in many ways. Hence, any use of the term needs to be accompanied by an appropriate definition. The Soil Science Society of America has defined available nutrients: (1) the amounts of soil nutrients in chemical forms accessible to plant roots or compounds likely to be convertible to such forms during the growing season, and (2) the contents of legally designated available nutrients in fertilizers determined by specified laboratory procedures which in most states constitute the legal basis for guarantees.

Soil productivity and nutrient availability are interrelated. Soil productivity is the capacity of a soil to produce a certain yield of crops or other plants with a specified system of management. Management of nutrients is an important aspect of maintaining soil productivity. Quantity of nutrient present in the soil is defined in terms of soil fertility, and, generally, soil analysis is used as a criterion to make fertilizer recommendations for field crops. Nutrient availability is evaluated by observations and tests which are used to predict the response of plants and management of nutrients. The presence of sufficient total quantities of essential nutrients in a soil does not guarantee the availability of these nutrients to growing plants, because other factors such as soil-moisture content, soil temperature, pH, soil physical conditions, presence of toxic elements and/or salts may be limiting. Hence, adequate levels of nutrients alone do not guarantee soil productivity. A productive soil is one that has optimal total environmental conditions for plant growth. Since soil is a continuum, it is a matrix in constant change. It is very difficult under practical conditions to have all crop production factors at an optimal level; hence, most of the crop production factors are usually at a suboptimal level.

In modern agriculture, maximizing and sustaining crop yields are the main objectives. One of the major problems constraining the development of an economically successful agriculture is nutrient deficiency for crop production. As much as 50% of the increase in crop yields worldwide during the twentieth century was due to the adoption of chemical fertilizers. Soil infertility (natural element deficiencies or unavailability) is probably the single most important factor limiting crop yields worldwide.

## Essential Nutrients for Plant Growth and Their Functions

The supply and absorption of chemical elements needed for growth and metabolism may be defined as nutrition, and the chemical elements required by an organism are termed 'nutrients.' Sixteen nutrients are essential for plant growth and development. On the basis of quantity required by plants, nutrients are divided into macro and micro groups. The essential nutrients such as carbon (C), hydrogen (H), oxygen (O), nitrogen (N), phosphorus (P), potassium (K), calcium (Ca), magnesium (Mg), and sulfur (S) are known as macronutrients; whereas iron (Fe), manganese (Mn), zinc (Zn), copper (Cu), boron (B), molybdenum (Mo), and chlorine (Cl) are termed micronutrients. Further, the essentiality of silicon (Si), sodium (Na), vanadium (V), and cobalt (Co) has been considered but is not yet proven. Plants obtain C, H, and O from air and water, and the remaining nutrients are absorbed from soil solution as inorganic ions or oxides by growing plant roots. Nutrient absorption by plants is usually referred to as 'ion uptake' or 'ion absorption,' because it is the ionic form in which nutrients are absorbed. Cations and anions may be absorbed independently and may not be absorbed in equal quantities; however, electroneutrality must be maintained in the plant and in the growth medium; therefore, ionic relationship achieves major importance in plant nutrition. The majority of cations in plant tissues are in the inorganic form, predominantly $K^+$, $Ca^{2+}$, and $Mg^{2+}$, and the majority of anions are in the organic form. The organic ions are synthesized within the tissue, while inorganic ions are absorbed from the growth medium. Generally, the monovalent cations are absorbed rapidly, whereas the divalent cations, especially $Ca^{2+}$, are absorbed more slowly. Similarly, the monovalent anions are generally absorbed more rapidly than polyvalent anions. The so-called physiological acidity or alkalinity of a salt depends upon which ion of the salt, the cation or the anion, is absorbed most rapidly. Thus a salt such as $K_2SO_4$ would be physiologically acid, since the K would enter the roots more rapidly than the $SO_4^{2-}$. By the same token, $CaCl_2$ should be physiologically alkaline, since the $Ca^{2+}$ enters slowly and the $Cl^-$ rapidly.

Mineral nutrients can also be absorbed by higher plants when applied as foliar sprays in appropriate concentrations. However, in modern high-yielding cultivars, nutritional requirements (macronutrients) are rarely met with foliar applications. Furthermore, foliar application of macronutrients requires several sprays and can be washed off by rain. The plant needs sufficient leaf area for absorption, and leaf damage by high nutrient concentrations is a serious practical problem. Despite these drawbacks, under some situations foliar application is the most effective method to correct a nutritional disorder. For example, iron deficiency in calcareous soils can be corrected by foliar application of ferrous sulfate solution more efficiently than the soil application of iron sources.

The macro- and micronutrient classification is simply based on the amount required: macronutrients are needed in relatively larger quantities than micronutrients. Each of these essential nutrients performs a specific biophysical or biochemical function within plant cells. Hence all nutrients are equally important for production of higher crop yields. If any one of these nutrients is deficient in the growth medium, plant growth is reduced. The low requirements of plants for micronutrients can be explained by the participation of these elements in enzymatic reactions and as constituents of growth hormones rather than as components of major plant products such as structural and protoplasmic tissue. To obtain higher yields, essential nutrients should be present in adequate concentrations and in available form, as well as in appropriate balance. An imbalance results when one or more nutrients are present in either deficit or excess supply. **Table 1** shows forms of absorption and functions of essential plant nutrients.

**Table 1** Forms of absorption and functions of essential nutrients in plants

| *Nutrient* | *Forms taken up by plants* | *Functions* |
|---|---|---|
| Carbon | $CO_2$ | Basic molecular component of carbohydrates, proteins, lipids, and nucleic acids |
| Hydrogen | $H_2O$ | Hydrogen plays a central role in plant metabolism. Important in ionic balance and as main reducing agent; plays a key role in energy relations of cells |
| Oxygen | $H_2O$, $O_2$ | Oxygen is like carbon in that it occurs in virtually all organic compounds of living organisms |
| Nitrogen | $NH_4^+$, $NO_3^-$ | Nitrogen is a component of many important organic compounds ranging from proteins to nucleic acids |
| Phosphorus | $H_2PO_4^-$, $HPO_4^{2-}$ | Central role in plants in energy transfer and protein metabolism |
| Potassium | $K^+$ | Helps in osmotic and ionic regulation. Functions as a cofactor or activator for many enzymes of carbohydrate and protein metabolism |
| Calcium | $Ca^{2+}$ | Calcium is involved in cell division and plays a major role in the maintenance of membrane integrity |
| Magnesium | $Mg^{2+}$ | Component of chlorophyll and a cofactor for many enzymatic reactions |
| Sulfur | $SO_4^{2-}$ | Sulfur is like phosphorus in that it is involved in plant cell energetic reactions |
| Iron | $Fe^{2+}$, $Fe^{3+}$ | An essential component of many heme and nonheme Fe enzymes and carriers, including the cytochromes (respiratory electron carriers) and the ferredoxins. The latter are involved in key metabolic functions such as N fixation, photosynthesis, and electron transfer |
| Zinc | $Zn^{2+}$ | Essential component of several dehydrogenases, proteinases, and peptidases, including carbonic anhydrase, alcohol dehydrogenase, glutamic dehydrogenase, and malic dehydrogenase |
| Manganese | $Mn^{2+}$ | Involved in the $O_2$-evolving system of photosynthesis; is a component of the enzymes arginase and phosphotransferase |
| Copper | $Cu^{2+}$ | Constituent of a number of important oxidase enzymes, including cytochrome oxidase, ascorbic acid oxidase, and lactase, and important in photosynthesis, and protein and carbohydrate metabolism |
| Boron | $B(OH)_3^0$ | Activates certain dehydrogenase enzymes, involved in carbohydrate metabolism, synthesis of cell wall components; is essential for cell division and development |
| Molybdenum | Mo | An essential component of nitrate reductase and $N_2$-fixation enzymes; is required for normal assimilation of N |
| Chlorine | $Cl^-$ | Essential for photosynthesis and as an activator of enzymes, involved in splitting water, and functions in osmoregulation of plants growing on saline soils |

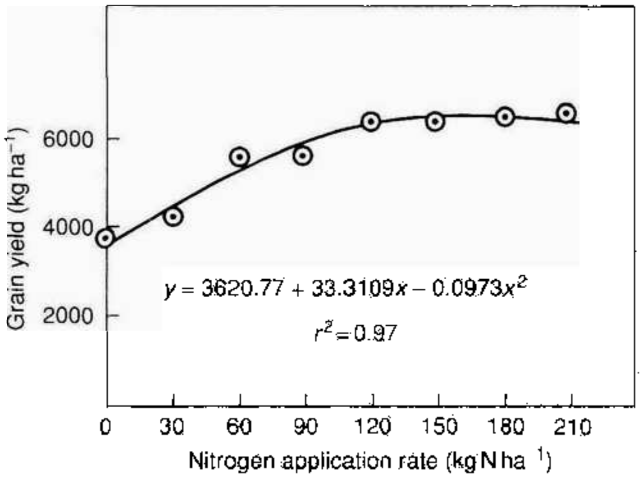


**Figure 1** Response of lowland rice to nitrogen fertilization grown on an Inceptisol of central Brazil.

Nutritional disorders (deficiency and/or toxicity) limit crop yields in all types of soils around the world. For example, in Brazilian Oxisols and Inceptisols, deficiency of N, P, K, and Zn limits the yield of almost all field crops. **Figure 1** shows the importance of N application in increasing yield of lowland rice grown on a Brazilian Inceptisol. It is clear from Figure 1 that yield of this important field crop increases with increasing levels of N until an adequate level in the soil is achieved. To obtain good yields of crops, nutrient disorders must be corrected. The first step in this direction is to identify the nutritional disorder.

## Diagnostic Techniques for Nutrient Availability

Diagnostic techniques for nutrient availability refer to the methods for identifying nutrient deficiencies, toxicities, or imbalances in the soil–plant system. Nutritional deficiency can occur when there is insufficient nutrient in the medium or when it cannot be absorbed and utilized by plants as the result of unfavorable environmental conditions. **Table 2** shows soil conditions associated with nutrient deficiencies. Nutritional deficiencies are very common in almost all field crops worldwide. The magnitude varies from crop to crop and region to region. Even some cultivars are more susceptible to nutritional deficiencies than others within a crop species. Four methods assess nutrient availability or sufficiency of mineral nutrients for plant growth: (1) visual symptoms, (2) soil testing, (3) plant analysis, and (4) crop growth response. These four approaches are becoming

**Table 2** Soil conditions inducing nutrient deficiencies for crop plants

| *Nutrient* | *Conditions inducing deficiency* |
|---|---|
| Nitrogen | Excess leaching with heavy rainfall; low organic matter content; burning the crop residue |
| Phosphorus | Acidic, organic, leached, and calcareous soils; high rate of liming |
| Potassium | Sandy, organic, leached, and eroded soils; high liming application, intensive cropping system |
| Calcium | Acidic, alkaline, or sodic soils |
| Magnesium | Acidic, alkaline, or sodic soils |
| Sulfur | Low organic matter content of soils; use of N and P fertilizers containing no sulfur; burning the crop residue |
| Iron | Calcareous soils; soils high in P, Mn, Cu, or Zn; high rate of liming |
| Zinc | Highly leached acidic soils; calcareous soils; high levels of Ca, Mg, and P |
| Manganese | High organic matter content; calcareous soils |
| Copper | Sandy soils; high liming rate in acid soils |
| Boron | Sandy soils; naturally acidic leached soils; alkaline soils with free lime |
| Molybdenum | High podzolized soils; well-drained calcareous soils |

**Table 3** Nutrient-deficiency symptoms in field crops

| *Nutrient* | *Symptoms* |
|---|---|
| N | Chlorosis starts in old leaves; in cereals tillering is reduced; under field conditions, if deficiency is severe, whole crop appears yellowish and growth is stunted |
| P | Growth is stunted; purple–orange color of older leaves; new leaves dark green; in cereals tillering is drastically reduced |
| K | Older leaves may show spots or marginal burn starting from tips; increased susceptibility to diseases, drought, and cold injury |
| Ca | New leaves become white; growing points die and curl |
| Mg | Marginal or interveinal chlorosis with pinkish color of older leaves; sometimes leaf-rolling, like drought effect; plants susceptible to winter injury |
| S | Chlorosis of younger leaves; under severe deficiency whole plant becomes chlorotic and similar to appearance in N deficiency |
| Zn | Rusting in strip of older leaves with chlorosis in fully matured leaves; leaf size is reduced |
| Fe | Interveinal chlorosis of younger leaves; under severe deficiency whole leaf becomes first yellow and finally white |
| Mn | Similar to iron deficiency; at advanced stage necrosis develops instead of white color |
| Cu | Chlorosis of young leaves, rolling, and dieback |
| Mo | Mottled pale appearance in young leaves; bleaching and withering of leaves |
| B | Pale green tips of blades, bronze tints; death of growing points |

widely used separately or collectively as nutrient-availability, -deficiency or -sufficiency diagnostic aids. They are extremely helpful, yet are not without limitations.

### Visual Symptoms

When the supply of a particular nutrient is at an inadequate level in the soil or when plant roots are not able to absorb required amounts, owing to unfavorable conditions in the rhizosphere, plants show certain growth disorders. These disorders may be expressed in reduced height, reduced tillering in cereals, leaf discoloration, reduced root growth, and reduced growth of newly emerging parts of the plant. Table 3 summarizes the symptoms of nutritional deficiency in crop plants. Using visual symptoms is the cheapest method of diagnosis for nutritional disorders. However, this technique needs a lot of experience on the part of the observer, because deficiency symptoms can be confused with problems caused by drought, insect and disease infestation, herbicide damage, soil salinity, soil acidity and inadequate drainage. Sometimes, a plant may be on the borderline with respect to deficiency and adequacy of a given nutrient. In this situation there are no visual symptoms, but the plant is not producing at its capacity. This condition is frequently called 'hidden hunger.'

Deficiency symptoms normally occur over an area and not on an individual plant. If a symptom is found on a single plant, it may be due to disease or insect injury or a genetic variation. Also, the earlier symptoms are often more useful than late, mature symptoms. Some nutrients are relatively immobile in the plant, while others are more mobile. In general, deficiency symptoms caused by immobile nutrients first appear on the younger or upper leaves. The older leaves don't show any symptoms because immobile nutrients do not move or translocate from older to newer leaves. Immobile nutrients are Ca, Zn, B, Cu, Fe, Mn, and Mo.

In contrast, when there is a deficiency of a mobile nutrient, the symptoms first appear on the older leaves of the plant. This is because the mobile nutrients move out of the older leaves to the younger part of the plant. The mobile nutrients are N, P, K, and Mg. S may behave as a mobile or an immobile nutrient. However, in rice plants S deficiency first appears in younger leaves.

The use of visible symptoms has the advantage of direct field application without the need of costly equipment or laboratory support services, as is the case with soil and plant analysis. A disadvantage is that sometimes it is too late to correct deficiency of a given nutrient, because for some disorders considerable yield loss may have already occurred by the time visible symptoms appear.

### Soil Testing

In a broad sense, soil testing is any chemical or physical measurement that is made on a soil. The main objective of soil testing is to measure soil nutrient status and lime requirements in order to make fertilizer and lime recommendations for profitable farming. Soil testing is an important tool in high-yield farming but produces best results only when used in conjunction with other good farming practices:

> There is good evidence that the competent use of soil tests can make a valuable contribution to the more intelligent management of the soil

This statement was made by the US National Soil Test Workgroup in its 1951 report and is still applicable today.

Soil testing involves collecting soil samples, preparation for analysis, chemical or physical analysis, interpretation of analysis results, and finally making fertilizer and lime recommendations for the crops.

The use of soil analysis as a fertilizer recommendation method is based on the existence of a functional relationship between the amount of nutrient extracted from the soil by chemical methods and crop yield. When a soil analysis test shows low levels of a particular nutrient in a given soil, application of that nutrient is expected to increase crop yield. Generally, nutrient analysis is arbitrarily classified as very low, low, adequate, high, and excessive. At very low nutrient levels, relative crop yield is expected to be less than 70% and a larger application of fertilizer for soil-building purposes is required. After the application of the nutrient, growth response is expected to be dramatic and profitable. At the low fertility level, relative yield is expected to be 70–90%. Under these conditions, annual application of fertilizer is necessary to produce maximum response and increase soil fertility. The increased yield justifies the cost of fertilization. When soil analysis testing shows adequate levels, relative crop yield under this situation is expected to be 90–100%. Normal annual applications to produce maximum yields are recommended. In this case more fertilizer may increase yields slightly, but the added yield would not pay back the expense of the additional fertilizer. At high levels of nutrient, there is no increase in yield, but small applications of fertilizer maintain soil level. The recommended amount may be doubled and applied in alternative years. When soil testing shows very high or excessive nutrient levels, yield may be reduced due to toxicity or imbalances of nutrients. In this situation there is no need to apply nutrient until the level drops back into the low range. To get such a nutrient level and yield relationship, it is necessary to conduct fertilizer yield trials in several locations in a given agroecologic region for different crops. Some specific recommendations for soil analysis are summarized here:

1. Soil samples must be representative of the land area in question. The recommendation is to take a minimum of one composite sample per 12–15 ha for lime and fertilizer evaluations. A representative soil sample is composed of 15–20 subsamples from a uniform field with no major variation in slope, drainage, or past fertilizer history. Any of these listed factors, if changed, will have an effect on the number of samples and unit area from which the sample is obtained;
2. Depth of sampling for mobile nutrients such as N should be 60 cm and, for immobile nutrients such as P, K, Ca, and Mg, 15–20 cm. For pasture crops, a sampling depth of 10 cm is normally sufficient to evaluate nutrient status and make liming and fertilizer recommendations;
3. The appropriate extractor must be selected. Three extracting solutions, $0.05\,\mathrm{mol\,l^{-1}}$ HCl – $0.025\,\mathrm{mol\,l^{-1}}$ $H_2SO_4$ (Mehlich 1), $0.03\,\mathrm{mol\,l^{-1}}$ $NH_4F + 0.025\,\mathrm{mol\,l^{-1}}$ HCl (Bray-P1), and $0.5\,\mathrm{mol\,l^{-1}}$ $NaHCO_3$ at pH 8.5 (Olsen) are the most commonly used extractants for P at the present time and are generally adequate to cover a broad range of soils. Commonly used extractants for K, Ca, and Mg are double acid (Mehlich 1), $1\,\mathrm{mol\,l^{-1}}$ $NH_4Ac$ at pH 7, and NaOAc at pH 4.8. Multielement extracting reagents are replacing the more familiar, single-element extractants. After mixing with an appropriate aliquot of soil, the obtained extract is assayed by an inductively coupled plasma emission spectrometer (ICP-AES). A flow-injection analyzer (FIA) is another multielement analyzer capable of assaying these soil extracts;
4. Optimum soil test values for macro- and micronutrients, which vary from soil to soil, crop to crop, and extraction to extraction, are required. They are usually approximately greater than $10\,\mathrm{mg\,P\,kg^{-1}}$, greater than $50\,\mathrm{mg\,K\,kg^{-1}}$, greater than $600\,\mathrm{mg\,Ca\,kg^{-1}}$, greater than $120\,\mathrm{mg\,Mg\,kg^{-1}}$, and greater than $12\,\mathrm{mg\,S\,kg^{-1}}$ for producing satisfactory results for most soils and crops. For micronutrients, the critical values reported are: Fe, $2.5\text{–}5\,\mathrm{mg\,kg^{-1}}$; Mn, $4\text{–}8\,\mathrm{mg\,kg^{-1}}$; Zn, $0.8\text{–}3\,\mathrm{mg\,kg^{-1}}$; B, $0.1\text{–}2\,\mathrm{mg\,kg^{-1}}$, Cu, $0.5\text{–}2\,\mathrm{mg\,kg^{-1}}$, and Mo, $0.2\text{–}0.5\,\mathrm{mg\,kg^{-1}}$, respectively;
5. pH values should be ascertained; the pH of agricultural soils is in the range of 4–9. It is difficult to define optimum pH values of different plant species. Most food crops grow well in acid soils of pH approximately 6. Lime is called the foundation of crop production or 'workhorse' in acid soils;
6. Fertilizer field trials establish nutrient requirement and are the most important part of the overall

diagnostic process of a nutrient disorder, whereby the full interplay of soil, plant, climate, and management factors can occur.

## Plant Analysis

Plant analysis is based on the concept that the concentration of an essential element in a plant or part of a plant indicates the soil's ability to supply that nutrient. This means it is directly related to the quantity in the soil that is available to the plant. For annual crops, the primary objective of plant analysis is to identify nutritional problems or to determine or monitor the nutrient status during the growing season. If deficiency is identified early in the growth stage of a crop, it is possible to correct it during the current season. Otherwise, steps should be taken to correct it in the next crop. Plant analysis can be useful for the prediction of nutrient needs in perennial crops, usually for the year following the time of sampling and analysis. Like soil analysis, plant analysis also involves plant sampling, plant tissue preparation, analysis, and interpretation of analytical results. All these steps are important for a meaningful plant analysis program.

Many factors such as soil, climate, plant species, and their interactions affect absorption of nutrients by growing plants. However, the concentrations of the essential nutrients are maintained within rather narrow limits in plant tissues. Such consistency is thought to arise from the operation of delicate feedback systems, which enable plants to respond in a homeostatic fashion to environmental fluctuations. **Table 4** shows average adequate concentrations for essential nutrients in important field crops. No doubt these values vary with soil, climate, crops, and management practice, and it is very difficult or not logical to make generalizations. Generalized values (**Table 4**) do give the reader some idea about what adequate levels of nutrients are in crop plants.

For the interpretation of plant analysis results, a critical nutrient concentration concept has been developed. This concept is widely used in the interpretation of plant analysis results for diagnosing nutritional disorders. Critical nutrient concentration is usually designated as a single point within the bend of the curve when crop yield is plotted against nutrient concentration where the plant nutrient status shifts from deficient to adequate. The critical nutrient concentration has been defined in several ways: (1) the concentration that is just deficient for maximum growth; (2) the point where growth is 10% less than the maximum; (3) the concentration where plant growth begins to decrease; and (4) the lowest amount of element in the plant accompanying the highest yield. The above definitions are similar, but not identical. It is well known that there is a good relationship between concentration of a nutrient and yield of a given crop. This is shown in **Figures 2** and **3**, which show a highly significant correlation ($R^2 = 0.94$ and $R^2 = 0.78$) between N in lowland rice plants (*Oryza sativa* L.) and average yields of lowland rice from three field experiments in the central part of Brazil. Similarly, N uptake (N concentration × dry weight) in shoots as well as in grain has a significant quadratic relationship with shoots as well as grain yield in lowland rice (**Figures 4 and 5**). In addition to the critical nutrient-concentration approach, tissue

**Table 4** Average concentration of essential nutrients in dry matter sufficient for adequate growth of field crops[a]

| | *Concentration* | |
|---|---|---|
| *Nutrient* | *g kg$^{-1}$ or mg kg$^{-1}$* | *μmol g$^{-1}$* |
| Carbon | 450 | 37 500 |
| Oxygen | 450 | 28 125 |
| Hydrogen | 60 | 60 000 |
| Nitrogen | 14 | 1 000 |
| Phosphorus | 1.9 | 60 |
| Potassium | 9.8 | 250 |
| Calcium | 5.0 | 125 |
| Magnesium | 1.9 | 80 |
| Sulfur | 1.0 | 30 |
| Iron | 112 | 2 |
| Manganese | 55 | 1 |
| Zinc | 20 | 0.3 |
| Copper | 6 | 0.1 |
| Boron | 22 | 2 |
| Molybdenum | 0.10 | 0.001 |
| Chlorine | 106 | 3 |

[a] Values of macronutrients are in grams per kilogram and values of micronutrients in milligrams per kilogram.

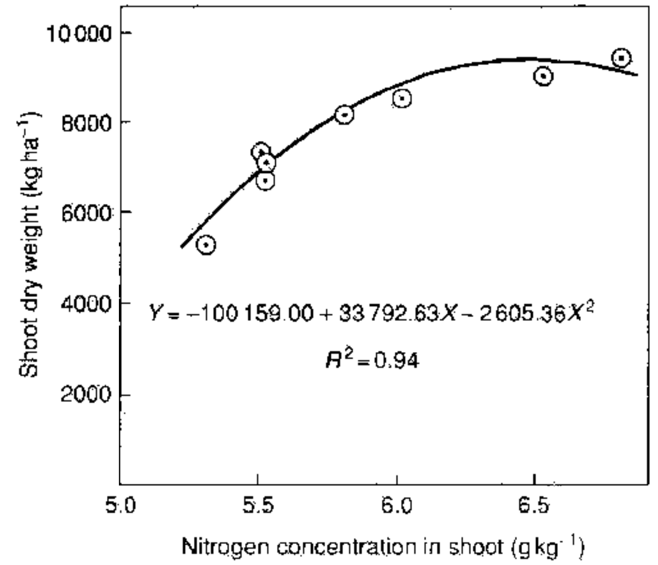


**Figure 2** Relationship between nitrogen concentration in shoots and shoot dry weight of lowland rice.

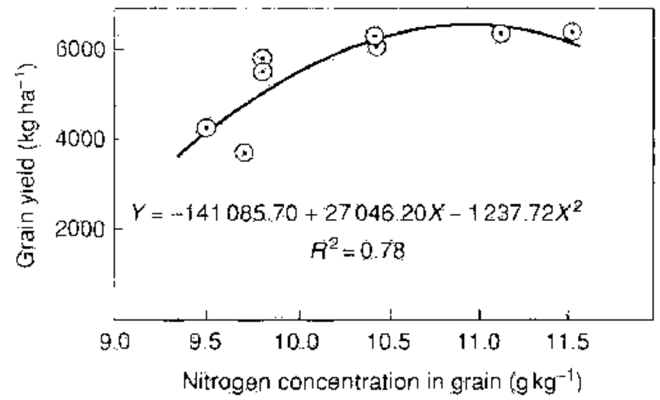


**Figure 3** Relationship between nitrogen concentration in grain and grain yield of lowland rice.

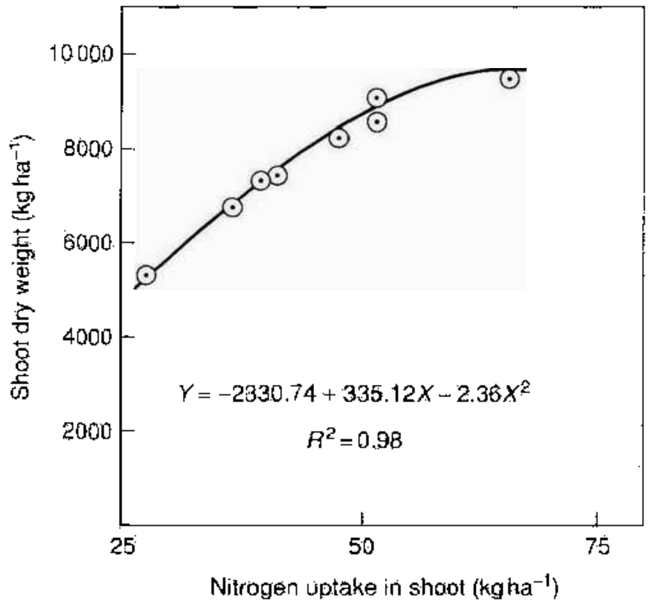


**Figure 4** Relationship between nitrogen uptake in shoots and shoot dry weight of lowland rice.

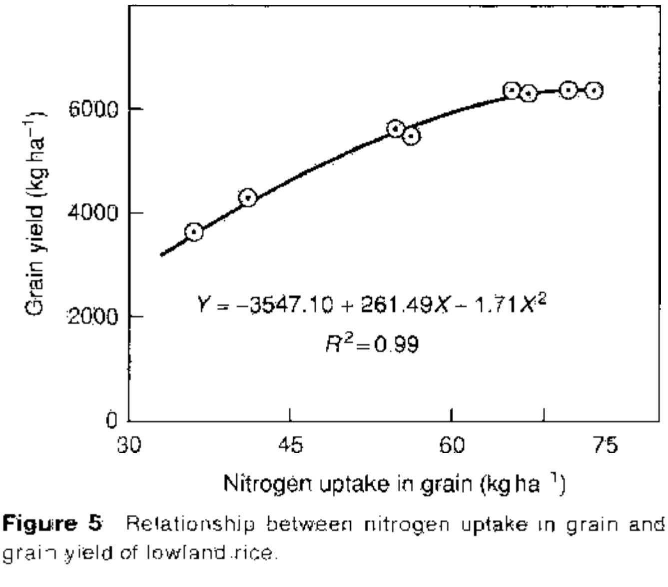


**Figure 5** Relationship between nitrogen uptake in grain and grain yield of lowland rice.

**Table 5** Approximate nutrient accumulation by important field crops

| *Crop* | *Yield (t ha$^{-1}$)* | *N (kg ha$^{-1}$)* | *P (kg ha$^{-1}$)* | *K (kg ha$^{-1}$)* | *Mg (kg ha$^{-1}$)* | *S (kg ha$^{-1}$)* |
|---|---|---|---|---|---|---|
| Alfalfa (hay) | 9 | 240 | 28 | 141 | 24 | 25 |
| Cotton (lint) | 1 | 120 | 20 | 75 | 24 | 20 |
| Peanut | 2 | 170 | 13 | 91 | 12 | 15 |
| Rice | 6 | 100 | 22 | 133 | 12 | 10 |
| Oat | 4 | 130 | 20 | 133 | 9 | 20 |
| Potato | 40 | 175 | 35 | 257 | 24 | 20 |
| Sweet potato | 40 | 190 | 33 | 282 | 39 | – |
| Sugar beet | 45 | 200 | 39 | 249 | 54 | 35 |
| Sugar cane | 100 | 130 | 39 | 282 | 48 | 60 |
| Rapeseed | 3 | 165 | 31 | 183 | 18 | 65 |
| Barley | 5 | 150 | 24 | 125 | 15 | 20 |
| Pea | 2 | 125 | 15 | 86 | 9 | – |
| Dry bean | 2 | 125 | 15 | 66 | 9 | – |
| Sunflower | 3 | 120 | 26 | 199 | 33 | 15 |
| Cassava | 40 | 150 | 31 | 291 | 24 | 20 |
| Corn | 6 | 120 | 22 | 100 | 24 | 25 |
| Soybean | 3 | 220 | 18 | 141 | 24 | 20 |
| Sorghum | 4 | 120 | 18 | 83 | 18 | 15 |
| Wheat | 6 | 170 | 33 | 145 | 18 | 30 |

Reproduced from Potash and Phosphate Institute of Canada (1988) *Potash: Its Need and Use in Modern Agriculture*. Saskatchewan, Canada: Potash and Phosphate Institute of Canada.

nutrient status can also be diagnosed by the diagnosis and recommendation integrated system (DRIS) and composition nutrient diagnosis (CND).

Nutrient concentrations in plant are generally expressed in milligrams per kilogram or grams per kilogram. Sometimes there is confusion between concentration and accumulation. Concentration is content per unit dry weight of plant tissue, whereas accumulation is concentration multiplied by dry matter. Nutrient accumulation is generally expressed in kilograms per hectare. **Table 5** shows nutrient accumulation by important field crops. Nutrient accumulation is related to crop yield and crop species. Nitrogen and potassium are accumulated in higher quantities as compared to P, Mg, and S. Uptake and absorption are also very common in expressing plant analysis results. Uptake or absorption normally refers to the process of elemental movement into and within plants and not to quantitative-based terms such as concentration or total accumulation.

Data related to the quantity of nutrients accumulated during a cropping season give an idea of fertility depletion and help in managing soil fertility for the succeeding crops. The best time to determine nutrient accumulation is at flowering or at harvest, when

**Table 6** Methods of correcting nutrient deficiency and Al and Mn toxicity

| *Nutrient/element* | *Corrective measures* |
|---|---|
| Nitrogen | Addition of organic matter to the soil; application of N fertilizers, including legumes in the crop rotation; use of foliar spray of 0.25–0.5% solution of urea |
| Phosphorus | Adjustment of extreme pH; application of phosphorus fertilizers |
| Potassium | Application of potassium fertilizers, incorporation of crop residues |
| Calcium | Liming of acid soils; addition of gypsum or other soluble calcium source where lime is not needed; foliar spray in acute cases with 0.75–1% calcium nitrate solution |
| Magnesium | Application of dolomitic limestone; foliar application of 2% magnesium sulfate solution |
| Sulfur | Use of fertilizer salt containing sulfur such as ammonium sulfate and single superphosphate; application of gypsum or elemental sulfur |
| Zinc | Addition of zinc sulfate to soil; foliar spray of 0.1–0.5% solution of zinc sulfate |
| Iron | Foliar spray of 2% iron sulfate or 0.02–0.05% solution of iron chelate; use of efficient cultivars |
| Copper | Soil application of copper source of fertilizer or foliar spray of 0.1–0.2% solution of copper sulfate |
| Boron | Soil application of boron source or foliar spray of 0.1–0.25% solution of borox |
| Molybdenum | Liming of acid soils, soil application of sodium or ammonium molybdate; foliar spray of 0.07–0.1% solution of ammonium molybdate |
| Manganese | Foliar application of 0.1% solution of manganese sulfate |
| Aluminum/ manganese | Application of lime; use of tolerant species or cultivars |

accumulation is expected to be maximum. Grains as well as straw should be analyzed and their weights per unit area determined to calculate total accumulation.

### Crop Growth Response

Soil and plant analyses are the common practices for identifying nutritional deficiencies in crop production. The best criterion, however, for diagnosing nutritional deficiencies in annual crops is through evaluating crop responses to applied nutrients. If a given crop responds to an applied nutrient in a given soil, this means that the nutrient is deficient for that crop. Relative decrease in yield in the absence of a nutrient, as compared to an adequate soil fertility level, can give an idea of the magnitude of nutrient deficiency. For example, research conducted on Oxisols of central Brazil provides evidence of which major nutrient is most limiting for upland rice, common bean, corn, soybean, and wheat. The conclusion was that P deficiency was the primary yield-limiting nutrient for annual crop production. Methods for correcting nutrient deficiencies and AL and Mn toxicities are summarized in **Table 6**.

## Summary

Soil is our most precious nonrenewable resource in food production. In the long run we must make sure that the productivity of the soil is maintained or improved. Maintaining adequate level of nutrients is one way to achieve this objective. Historically, the study of soil fertility has focused on managing soil nutrient status to create optimal conditions for plant growth. The amount and availability of soil nutrients are determined by physical, chemical, and biological processes, which are often managed in an attempt to increase or optimize uptake of nutrients and yields. There are 16 essential nutrients for plant growth and their presence in the soil at an adequate level and proportion is fundamental for obtaining higher yields. The important criteria in evaluating nutrient availability or deficiency or sufficiency are visual symptoms, soil testing, plant tissue testing, and crop response to applied nutrients. Soil testing is generally recognized as the best available method for diagnosing soil nutrient limitations before a crop is planted so that correction can be made in that year by appropriate fertilization.

Soil fertility is measured in terms of quantity of available nutrient in a given soil. However, plant nutrition involves transport of nutrients from soil to plant roots, absorption by roots, translocation from absorbing site to various parts of the plants, and finally utilization in the growth and development processes of plants. This means it is a very dynamic process and influenced by soil, climate, and plant factors. A thorough knowledge of these factors is essential to understand properly the subject of mineral nutrition of plants.

Integrated nutrient management systems should be used to improve the overall nutrient availability and use efficiency by crops. Successful application of many practices which are part of the integrated nutrient management can increase the profitability and minimize the environmental impacts of nutrient use, the fundamental goal of supplying adequate amounts of nutrients in appropriate proportions for higher yields.

*See also:* **Fertigation; Fertility; Fertilizers and Fertilization; Foliar Applications of Nutrients; Macronutrients; Micronutrients; Nutrient Management**

## Further Reading

Baligar VC, Fageria NK, and He ZL (2001) Nutrient use efficiency in plants. *Communication in Soil Science and Plant Analysis* 32: 921–950.

Barber SA (1995) *Soil Nutrient Bioavailability: A Mechanistic Approach*, 2nd edn. New York: John Wiley.

Bennett WF (1993) *Nutrient Deficiencies and Toxicities in Crop Plants*. St. Paul, MN: American Phytopathological Society.

Black AM (2000) Bioavailability of major essential nutrients. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 3–37. Boca Raton, FL: CRC Press.

Fageria NK and Baligar VC (1997) Integrated plant nutrient management for sustainable crop production: an overview. *International Journal of Tropical Agriculture* 15: 1–18.

Fageria NK and Baligar VC (2001) Lowland rice response to nitrogen fertilization. *Communication in Soil Science and Plant Analysis* 32: 1405–1429.

Fageria NK and Baligar VC (2001) Improving nutrient use efficiency of annual crops in Brazilian acid soils for sustainable crop production. *Communication in Soil Science and Plant Analysis* 32: 1303–1319.

Fageria NK and Gheyi HR (1999) *Efficient Crop Production*. Campina Grande, Brazil: Federal University of Paraiba.

Fageria NK, Baligar VC, and Jones CA (1997) *Growth and Mineral Nutrition of Field Crops*, 2nd edn. New York: Marcel Dekker.

Foth HD and Ellis BG (1997) *Soil Fertility*, 2nd edn. Boca Raton, FL: CRC Press.

Grundon NJ (1987) *Hungry Crops: A Guide to Nutrient Deficiencies in Field Crops*. Brisbane, Australia: Department of Primary Industries.

Marschner H (1995) *Mineral Nutrition of Higher Plants*, 2nd edn. New York: Academic Press.

Mengel K, Kirkby EA, Kosegarten H, and Appel T (2001) *Principles of Plant Nutrition*, 5th edn. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Mortvedt JJ (2000) Bioavailability of micronutrients. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 71–88. Boca Raton, FL: CRC Press.

Tisdale SL, Nelson WL, Beaton JD, and Havlin JL (1993) *Soil Fertility and Fertilizers*, 5th edn. New York: Macmillan Publishers.

---

# NUTRIENT MANAGEMENT

**G D Binford**, University of Delaware, Newark, DE, USA

## Introduction

The management of nutrients in soils is important for both economic and environmental reasons. Plants require certain amounts of each essential nutrient to maintain optimum health. During the production of crops, it is important to have enough of each nutrient present in the soil to optimize productivity and maximize profitability. These nutrients come primarily from native soil minerals, the application of fertilizers, nitrogen fixation, or the application of animal manures. Although these nutrients are essential for optimum plant growth and crop production, it is important from an environmental standpoint that these nutrients are not applied in excess of what is required by the plant. When excessive amounts of nutrients, primarily nitrogen (N) and phosphorus (P), are applied to the soil, these nutrients can leak from the soil and move into nearby water supplies. As surface waters increase in concentrations of N and P, these increases can result in accelerated growth of water plants. An overabundance of plants growing in the water can result in a process known as eutrophication, which can be detrimental to the overall health of the water ecosystem. Excess N, as nitrate, can also leak from soils and get into groundwater supplies, which can be a concern for both human and animal health.

The potential for losses of nutrients is especially great in areas with intensive concentrations of animals, because the animal wastes from these facilities contain significant quantities of nutrients. As a result, proper nutrient management is an especially major concern in areas with large animal production facilities. These animal wastes are usually land applied to provide nutrients for the crops to be grown on the land. These wastes, however, often contain a higher P:N ratio than the ratio needed by the crop. In addition, because N is often needed in relatively high amounts for crops such as corn, when these manures are applied to meet the N demands of the crop, the amount of P applied is much greater than that required by the crop. As a result, after decades of similar applications, P concentrations in soils become quite excessive. Research has shown that, when P concentrations of soils reach a certain level, the risk of P loss from the soil to nearby water supplies greatly increases.

Losses of nutrients are not only confined to operations with intensive livestock operations. In fact, nutrients can be lost from any soil when nutrients are present at levels greater than growing plants can utilize and/or the soil can hold. This can occur on farms, golf courses, lawns, or any other place nutrients are applied to soils. Also, poor application techniques can lead to nutrient losses from soils. For example, nutrients that are applied accidentally to impervious surfaces such as roads or sidewalks will most likely be lost in runoff and could impact surface-water quality.

In recent years, the concept of nutrient-management planning has been promoted as a practice that will help prevent overapplication and improper application of nutrients to the land. The primary goals of nutrient-management planning are to maximize profitability of crop production and to prevent a buildup of excess nutrients on the land, thereby reducing the potential for nutrient losses from land to water.

## Nutrient-Management Planning

There is a specific process that is typically followed in the development of a sound nutrient-management plan. It is important that a nutrient-management plan be developed by someone who understands both soil and crop sciences, so that both an environmentally sound and an economically feasible nutrient-management plan can be prepared. The process of developing a nutrient-management plan can be broken down into several steps. These steps include:

1. Development of a map for the operation;
2. Determination of crop or plant nutrient requirements;
3. Inventory of nutrient production on the operation;
4. Determination of nutrient balance on the operation;
5. Emergency action plan;
6. Record keeping.

### Step 1: Creating Maps

Detailed maps of the operation are developed and these maps include:

1. Identification of property boundaries;
2. Identification of field boundaries with field identification codes;
3. Location of any areas sensitive to nutrient loading, such as surface waters or wellheads;
4. Direction of flow for any streams on the property;
5. Identification of all soil types where nutrients are applied;
6. If present, location of animal production facilities;
7. Location of nutrient storage (fertilizer and/or manure);
8. Location of roads, neighboring homes, or public facilities.

These detailed maps are important for understanding the layout of the operation; they can be used to guide nutrient applications near areas that are critical to nutrient loading such as surface waters. If necessary, these maps can also be used as a guide to limit applications of animal wastes near public roads, neighboring homes, or areas of public gatherings. The maps with field identification codes are also helpful to the person implementing the plan so that the appropriate nutrient recommendation can be applied to the corresponding area of the field. Maps can be drawn by hand, with computer software, or created from aerial photographs.

### Step 2: Nutrients Required by Crops or Plants

Identifying which nutrients are required includes:

1. Determining the crop rotation or plants to be grown on the facility;
2. Determining crop/plant nutrient requirements;
3. Soil testing and/or plant analysis;
4. Nutrient credits for animal waste applications, if applicable;
5. Nitrogen credits for previous legume crops, if applicable;
6. Nitrogen credits for nitrate in irrigation waters, if applicable.

The first step in determining plant nutrient requirements is deciding on what type of crop or plants will be grown and the rotation of crops or plants for the period of the nutrient-management plan. After this decision has been made, a determination of the required nutrients is needed. The primary procedure for making this determination in most situations is by soil testing.

Soil testing is the best method available for determining how much nutrient is needed for optimum production of most crops. For some crops, the use of plant analysis may be the preferred method of determining crop nutrient needs. Making nutrient-application decisions without soil testing or plant analysis can be a risky endeavor because this will probably result in reduced profitability from either under- or overapplying nutrients; in addition, overapplication of nutrients can result in reduced water quality. It is important to follow recommended procedures for taking soil and plant samples.

When making N recommendations for many crops, the determining factor on how much N to apply is based on the expected yield of the crop. The best method for determining expected yield is from previous yield records. It is important to use an expected yield that is realistic, because if an expected yield is used that is much greater than can be achieved realistically then more nutrient will be applied than the crop can utilize. Applying excess nutrient increases the risk of nutrients getting into nearby water supplies and jeopardizing water quality. At the same time, using an expected yield that is too low can limit yields due to nutrient deficiencies, which reduces profitability of crop production.

Once the recommended rate of fertilizer has been determined, it is important to make any adjustments in the recommended rate for nutrient credits from animal wastes, previous legume crops, and N that is present in irrigation water. These credits can be determined for animal wastes by having a nutrient analysis done by a laboratory. Quantification of nitrates in irrigation water can also be determined by a laboratory. Credits for previous legume crops are usually standard values that have been determined by crop-management and soil-fertility research.

Because a nutrient-management plan is typically written to make fertilizer recommendations for all crops throughout the rotation, it is important to estimate the amount of nutrients that will be removed when a crop is harvested. Estimating the amounts of nutrient removed from the soil by a crop will usually be necessary for making appropriate fertilizer recommendations for subsequent crops in the rotation, unless soil samples are taken for each crop. Typically, soil samples for soil testing are taken every 2–3 years depending on the rotation and soil type. In some situations, a crop grower will apply enough nutrients (i.e., nonmobile nutrients) to meet the nutrient requirements of the current crop and the subsequent crop, which reduces application costs. An example of this practice would be a corn–soybean rotation, where a grower may often apply enough P and potassium (K) to the corn crop so that this application not only meets the requirements of the corn crop but also meets the P and K requirements of the following soybean crop.

### Step 3: Inventory of Available Nutrients on the Operation

It is only necessary to make an inventory of available nutrients for operations that have animals and are producing animal waste; if no animals are present, then this step will not be needed. If an operation has no animals, such as a golf course or a grain production farm, then there are no nutrients (i.e., animal waste) being generated on the operation, so all nutrient needs on an operation of this type are fulfilled by importing nutrients to the operation. If animals are present on the operation, then it is important to determine quantities of nutrients (i.e., waste) being generated on the operation so that these nutrients can be utilized most appropriately so that potential losses of these nutrients to the environment are minimized and the economic value of these nutrients is maximized.

The first phase in determining nutrient inventory is to determine how much animal waste is being produced on the operation. If there are no records from past animal waste production from the operation, it is necessary to use average animal waste production values for the development of the nutrient-management plan. These average production values have been generated from previous research and can be found through most University/Cooperative Extension programs. To improve the accuracy of determining the amount of animal waste being generated on an operation, records can be kept on the amount of manure that is removed from the operation. These amounts can be determined by weighing each load of animal waste as it leaves the animal production facility or by determining the volume of animal waste as it leaves the facility and then converting this volume to an estimated weight. Actual measurements from the animal operation would be the most accurate and reliable method for determining nutrient inventory on an animal operation, as compared to just using established book (i.e., average) values.

In addition to determining the quantity of animal waste being produced, the nutrient content of the animal waste should be determined. This can be done by sending a representative sample of the animal waste to a laboratory for analysis. It is important to make sure a representative sample is taken and also that the sample is handled properly before sending to the laboratory. The laboratory should be able to provide details on the proper procedure for taking a representative sample of animal waste; procedures vary depending on the type of animal waste being sampled. It is also possible to obtain average book values for the nutrient content of animal wastes; however, laboratory analyses provide a better determination for each specific animal operation than an average book value.

It is important to remember when calculating the amount of available nutrient present on the operation that the availability of N in animal manures will be dependent on how the manure is handled before and during application. For example, significantly less N will probably be available for future crops when manure is applied on the soil surface and not incorporated compared with manure that is applied and incorporated immediately; this is because of potential volatilization of N from the nonincorporated manure. General guidelines have been developed for determining the amount of N that is potentially available to crops after the manure has been field-applied. It is important to remember that these are just guidelines and that weather conditions have a major impact on the amount of N that is potentially available to subsequent crops. Because of this potential variability in the amount of plant-available N that is present from field-applied manures, it is advisable to use diagnostic tools (e.g., the pre-sidedress soil nitrate test or PSNT, plant tissue analyses, and chlorophyll meters) to monitor the N status of the growing crop when deciding how much additional N to apply as fertilizer to the growing crop. The use of such tools helps to reduce excess applications of N, which reduce profitability and increase potential environmental harm.

### Step 4: Nutrient Balance

As in the previous step, balancing of nutrients is only needed if animals are present on the operation. If no animals are present, then this step is also unnecessary. After determining the quantity of nutrients needed by the planned crop rotation through proper soil testing and the quantity of nutrients that are being produced from the animal production facility, the next step is to evaluate the nutrient balance. The main question that has to be answered before the nutrient balance can be calculated is to determine whether the nutrient management plan will be an N- or P-based plan. In some situations, it may be appropriate to have a metals-based plan. In other words, the animal waste is applied based on the metal content of the waste. In the development of the plan, it is necessary to assess each management area (i.e., area from which the soil sample is taken) for its nutrient needs and whether that area will require an N-based or P-based plan.

Historically on many operations, animal waste has been applied at rates to meet the N demands of the crop or at even higher rates in situations where the waste has been considered a disposable by-product of the animal production facility. In recent years, the concept of P-based planning has become popular. This popularity in P-based planning has resulted from areas where animal wastes have been applied at P rates much greater than the amounts of P removed in the harvested portion of the crop. As a result, after decades of this practice, soil-test P levels have increased to levels that are considered a potential threat for off-site movement of P into nearby water supplies.

Before deciding whether a P-based plan should be used, the criteria to be used to define when the risk of P loss warrants the use of P-based planning instead of N-based planning must be determined. Simply defining a soil-test level that is considered too high is one method; however, a method usually referred to as a P site index probably makes more sense and would be more representative of P-loss potential from a soil. There is currently a great deal of research being conducted on the development of an appropriate P site index. It is important that a P site index be adapted and developed for local conditions. A P site index evaluates all aspects of potential P transport from soil to water and not just the current level of P in the soil (i.e., soil test P). A P site index provides a relative risk of P loss from a soil and takes into account such things as soil erosion potential, distance from water, P loading rates, soil-test P, and subsurface drainage.

A farm balance of nutrients can be determined after a decision has been made on which areas will need a P-based plan and which areas will need an N-based plan. The total nutrient requirements of the crops to be grown can be calculated by summing the total nutrient needs of each area that was soil-sampled. The difference between the total supplies of nutrients produced by the animal operation can then be compared with the total nutrient needs of the crops to be grown. If an excess of nutrients (i.e., animal waste) is available on the operation, an alternative plan should be developed for transporting the excess animal waste from the operation. It is also important to consider when the animal waste will be removed from the animal production facility and when it will be applied to the land or removed from the operation so that appropriate storage can be planned.

It is likely that significant future improvements in nutrient farm balances can be achieved through diet modifications of livestock. Currently, many livestock operations are adding P to animal diets at rates greater than needed for optimal animal health. As a result, much of this added P ends up in the manure and has to be land-applied. New technologies in diet formulations and animal genetics also offer opportunities for reducing the P content of animal manures. For example, recent work in Delaware and Maryland, USA, has shown that significant reductions can occur in the P content of poultry manure by utilizing phytase enzyme and low-phytate corn in the feed ration of the birds. In fact, it is possible that future use of these new technologies and with improved feed management the ratio of P:N in animal manures could be much more in balance with the ratio of P:N that is needed by crops. This would allow manures to be applied at rates equal to the amount of nutrient that will be removed by the crop, which would eliminate nutrient buildup in soils, thereby, reducing the potential for nutrient losses from land to water.

### Step 5: Emergency Action Plan

As a part of the overall nutrient-management plan, it is important to have a planned course of action in the event of a nutrient spill during transport or storage. This action plan should be in place for either manure or fertilizer that is collected, stored, transferred, or applied on the operation. This emergency plan should include telephone numbers for reporting a spill and for seeking assistance. The plan should also show anticipated flow paths in the event of a spill or discharge of nutrients. To be most helpful as a part of this action plan, the anticipated flow path should be shown on the overall site map of the operation.

### Step 6: Record Keeping

Maintaining detailed records is an important part of having a useful nutrient-management plan. Development of future nutrient-management plans is improved and less time-consuming if detailed records are maintained concerning the operation. Detailed records also provide documentation on procedures used on the operation, which could be helpful should the need arise to prove activities used on the operation. Many of the needed records for determining animal waste production may not be available the first time a nutrient-management plan is written, which will require the use of average book values. Future nutrient-management planning, however, will be made easier and more site-specific if on-site records are available from the operation.

The end result of nutrient-management planning is the same result of any planning process. It does not matter if we are talking about a business plan, a financial plan, or a nutrient-management plan: the more time and effort put into the planning process, the better the results. A well-written and detailed nutrient-management plan should result in an improved efficiency of nutrient use. This improved efficiency should result in an operation that has increased potential profitability and decreased potential damage to the environment.

## Summary

Nutrient management planning offers a systematic process for determining amounts of nutrients produced (i.e., animal wastes) and amounts of nutrients needed for optimal plant growth. Any operation that applies nutrients to land should have a nutrient-management plan prepared by an individual with a sound understanding of plant and soil sciences. There are six basic steps to follow when developing a nutrient management plan; however, the details involved with each step will vay depending on the type of operation. It is important to remember that, for normal growth and development, plants must acquire nutrients from soils through root uptake. Therefore, if soil testing indicates a need for nutrients, then nutrients must be added to the soil to support optimal plant growth. Nonetheless, if too much nutrient is applied to land, these excess nutrients can leak from soils and result in water-quality concerns. Therefore, the overall goal of sound nutrient management is to optimize plant health, while minimizing environmental harm.

*See also:* **Fertility; Fertilizers and Fertilization; Fertigation; Foliar Applications of Nutrients; Macronutrients; Micronutrients; Nutrient Availability**

## Further Reading

Beegle DB, Carton OT, and Bailey JS (2000) Nutrient management planning: justification, theory, practice. *Journal of Environmental Quality* 29: 72–79.

Hoeft RG (2002) Developing a Nutrient Management Plan. In: *Illinois Agronomy Handbook*, pp. 132–141. Urbana, IL: University of Illinois Cooperative Extension.

Koelsch R and Shapiro C (1997) *Determining Crop Available Nutrients From Manure*. NebGuide G97-1335A. Lincoln, NE: University of Nebraska–Lincoln.

Leytem AB and Sims JT (1999) *Nutrient Management for Delaware Agriculture – Developing a Comprehensive Nutrient Management Plan*. Fact Sheet NM-02. College of Agriculture, University of Delaware. Newark, DE: University of Delaware Press.

Midwest Farm Service (2002) *Livestock and Poultry Environmental Stewardship (LPES) Curriculum*. Ames, IA: Midwest Farm Service.

Natural Resources Conservation Service (2000) *Comprehensive Nutrient Management Planning Technical Guidance*. National Planning Procedures Handbook, Subpart E, Parts 600.50-600.54; Subpart F, Part 600.75. NRCS, US Department of Agriculture. Washington, DC: US Government Printing Office.

*Nutrient Management Planning with Animal Feeding Operations. Handbook for Nutrient Management Planners* (1999) Auburn, AL: Alabama Cooperative Extension System/Auburn University Press.

Sims JT and Moore PA Jr (1998) Nutrient management planning: phosphorus or nitrogen-based? In: National Poultry Waste Management Symposium. Auburn, AL: National Poultry Waste Management Symposium Committee.

# O

## ORGANIC FARMING

**C A Francis**, University of Nebraska, Lincoln, NE, USA

### Introduction

One of the most rapidly growing sectors of the food industry is organic products, in both the USA and Europe. Market share in 2002 was 2% in the USA and up to 5% in other countries. In part this growth is due to increasing concern in the general consumer population about the amounts of chemical pesticides that may be carried over from field production on the surface of fruits and vegetables and eventually find their way into the human food chain. Another concern comes from a recent series of outbreaks of foot and mouth disease, as well as bovine spongiform encephalitis in domestic animals in northern Europe, events that have caused many to begin buying meat from animals produced using organic methods. There are ambitious national goals in Europe to make 10–20% of all agriculture organic and/or ecologic before 2012.

Organic farming is defined as production of crop, animal, and other products without the use of synthetic chemical fertilizers and pesticides, transgenic species, or antibiotics and growth-enhancing steroids, or other chemicals. This is called 'ecological farming' in some northern European countries. It involves a philosophy of improving soil fertility, protecting crop plants and animals through natural methods and system design, maintaining the rural environment as a desirable place to live, and showing concern for the distribution of benefits from agriculture and the food system. In a sense, organic farming is an ancient art that evolved with the first farmers and dominated food production up until approximately 100 years ago. Before that, fossil fuel-derived pesticides or fertilizers were not available, and food was produced essentially by organic methods in most of the world. The landmark book *Farmers of Forty Centuries*, by F.H. King of the US Department of Agriculture, describes the systems used in China, Korea, and Japan, and these would qualify today as organic production systems. What we today call 'organic farming' has its roots in these ancient and highly sustainable systems in Asia.

Organic farming practices have depended on limited scientific research and extensive practical experiences by farmers and proponents of this approach. There are important precedents in Rudolph Steiner's lectures on biodynamic farming in Germany, as well as Sir Albert Howard's extensive observations and experiments in India on composting and improving soil humus. Developments in organic farming over recent decades have included increased attention to building soil fertility, fine-tuning methods of pest management, and emergence of certification organizations that provide a specific label to guarantee the authenticity of food produced by organic methods. Interest in ecologic agriculture in the Nordic region and throughout Europe has been especially intense in recent years. Useful sources that document this history are shown in **Table 1** and in the Further Reading section. These books describe the evolution of agricultural practices as well as the growing popularity of organic products over the last century.

Building on a brief history, this article continues with discussions of how organic farmers maintain and build soil fertility, and how they design systems to protect against pest problems, especially weed management. Important to development of organic farming are the contributions of the biodynamic movement and integrated systems research. Although organic crop yields generally are lower per unit area, many people speculate that this is due to less research and development invested in these systems. In fact, the best organic growers often have yields equal to or better than country or regional averages for crops, and most farmers benefit from lower production costs and premium prices. Labeling and marketing are important issues in the commercialization of organic food, and these must include discussion of claims for nutritional benefits from the products of organic farming practices. Finally, there is intense discussion about the industrialization of organic farming and food products, and this will continue as a contentious issue into the future.

**Table 1** Key background information on organic farming and gardening, from historical antecedents to current reviews

| *Key reference* | *Description of relevance to organic farming* |
|---|---|
| King, 1911[a] | Travelogue and descriptions of observations of farming systems by USDA's Dr. King during travels in China, Korea, and Japan in the early 1900s; a primer on classic, sustainable agriculture that has endured for more than 4000 years |
| Steiner, 1958[b] | Original eight lectures by Rudolph Steiner from Germany, founder of anthroposophy and father of biodynamic agriculture; these provide a source of many unique ideas on management and uniting the spiritual with the biological |
| Howard, 1943[c] | Detailed instructions on compost management and development of humus in the soil, based on experiences of Alfred Howard in India, plus ideas on integrated farming |
| Rodale, 1955[d] | Visionary author and publisher who demonstrated organic farming and gardening techniques on his research farm; founder of Rodale Press, he popularized the concept in the USA |
| Edwards *et al.*, 1990[e] | Comprehensive review of sustainable agriculture in a series of chapters, including Historical Roots written by Richard Harwood; contains many references and updates on organic farming |
| Lampkin, 1990[f] | Highly prolific writer and scientist who has chronicled the growth of the organic industry and published detailed lists of practices and practical economic guidelines (since 1995) in the UK |
| Sooby, 2003[g] | Survey results of organic farming research in the US landgrant system, showing that there is proportionately little research compared to conventional farming and market demand |
| Lotter, 2002[h] | An unusually complete review of the literature on organic farming; focuses on cultural practices as well as conversion process, economics, consumer demand, and certification activities in USA |

[a] King FH (1911) *Farmers of Forty Centuries, or Permanent Agriculture in China, Korea, and Japan.* Emmaus, PA: Rodale Press.
[b] Steiner R (1958) *Agriculture: A Course in Eight Lectures.* London, UK: Biodynamic Agriculture Association.
[c] Howard A (1943) *An Agricultural Testament.* London, UK: Oxford University Press.
[d] Rodale JI (1955) *Organic Gardening: How to Grow Healthy Vegetables, Fruits, and Flowers Using Nature's Own Methods.* Garden City, NJ: Hanover House.
[e] Edwards CA, Lal R, Madden P, Miller RH, and House G (eds) (1990) *Sustainable Agricultural Systems.* Ankeney, IA: Soil and Water Conservation Society.
[f] Lampkin N (1990) *Organic Farming.* Ipswich, UK: Farming Press.
[g] Sooby J (2003) *State of the States: Organic Systems Research at Land Grant Institutions, 2001–2003.* Santa Cruz, CA: Organic Farming Research Foundation.
[h] Lotter D (2002) Organic farming. *Journal of Sustainable Agriculture* 21(4): 59–128.

## Building Soil Fertility

Organic farmers build and maintain soil fertility through a number of practices that mimic the natural ecosystem, rather than supplying crop needs through application of chemical, fossil fuel-derived commercial fertilizers. One characteristic of organic farmers is their strong belief in the soil as a complex place that is home to a living ecosystem of billions of organisms that work together to provide cycling of nutrients, and not merely a physical medium that holds plants upright and provides a place to store chemical nutrients until they are taken up by plants. The care of this living system and the cycling of its nutrients in field crops, vegetables, pastures, and domestic animals are central to the philosophy of organic farming and a foundation on which farmers build sustainable systems.

Crop diversity, careful soil management, legumes, and animals in the system are key strategies for the design of efficient and sustainable nutrient relationships in each field. These diverse components can enhance the process of nutrient cycling within the system and contribute to minimizing the losses of nutrients. There is no special formula for creating more nutrients within the organic farming system, but rather a logical planning to minimize extraction or loss of nutrients from the field and farm, and to replace those nutrients using internal resources on the farm (see Table 2). The crop sequence and mixture are designed to maximize nitrogen fixation and preservation of organic matter, efficient cycling of nutrients, and retention of nutrients where they are needed on the farm. Some organic farmers maintain that it is impossible to design an efficient system without including ruminant animals, while others increase the proportion of legumes in the system and use this strategy to accomplish the same goals. The bottom line is to incorporate continuously as many nutrients as the amount extracted in harvest and removal from the farm. A goal is to increase organic matter and nutrients to a more healthy level for soil and plants when these factors are deficient, but not to a level that creates a risk for erosion or leaching of soluble nutrients. The concept of balance is paramount, and crop rotation is essential. (*See* Crop Rotations.)

Soil organic matter is central to nutrient storage and cycling in the soil, and to providing nutrients to crop and pasture plants at the right time in their growth cycles. Along with the clay mineral fraction of the soil, organic matter provides negatively charged sites for cations (positively charged nutrient ions such as potassium, calcium, and magnesium) to be adsorbed, thus contributing to the cation exchange

**Table 2** Nutrient management practices and processes used by organic farmers to maintain soil fertility adequate for efficient and economical plant growth

| *Practice or process* | *Description and impact on soil quality and crop production* |
|---|---|
| Legumes in rotation | Nitrogen fixation, contributions to soil organic matter, fibrous root systems contribute to loosening soil and quality |
| Manure application | Provides nitrogen and other nutrients, increases level of soil organic matter, improves soil tilth and quality |
| On-farm cycling | Use of primarily green manure crops and animal manure produced by feeding grain or forage on same farm helps balance |
| Cover crops | Provide cover over the ground surface during much of the year to prevent erosion, trap nutrients, contribute organic matter |
| Timely seeding | Synchronize crop planting and growth cycle with availability of nutrients from soil reserve |
| Reduced tillage | Maintains crop residues on top of soil, where they break down more slowly and provide nutrients when needed by later crops |
| Crop–Ley rotation | Rotation of crops with perennial pastures and grazing animals provides manure, fixed nitrogen, and green biomass to soil |
| Contour planting | Seeding crops on contours or terraces reduces soil erosion and thus reduces nutrient losses with eroding soil |
| Crop–animal systems | Integration of a diverse array of animal enterprises provides an internal market for forages and helps cycle nutrients |

capacity of the soil. Organic matter is increased by incorporating green manure crops and growing forage crops or pastures, as long as they are fed to ruminants and manure recycled in the field, and these crops are not harvested and sold off the farm. The important organic matter is the active fraction that is being broken down by soil organisms and chemical reactions in the soil solution, thus releasing nutrients available to crops.

Algae are the only soil organisms to capture energy directly from sunlight, while all the others (bacteria, fungi, actinomycetes, and a host of macroorganisms) are dependent on breakdown of organic matter to provide their energy and nutrients, which they release and make available to crop and pasture plants. Soils that are too acid or too basic have lower levels of microbial and macroorganism activity, and thus break down soil organic matter more slowly. Most soil organisms such as earthworms are sensitive to soil acidity, and most have optimum activity at near-neutral soil conditions ($\mathrm{pH} = 7.0$). Organic systems are designed to achieve this balance and build soil organic matter. 'Humus' is a term used for a combination of decomposed organic matter and high-molecular-weight compounds that are generally more resistant to biological decomposition than the active organic matter. Preserving organic matter by preventing erosion, minimizing extraction from the field (such as hay or silage harvest), and increasing the proportion of crop residues or green-manure crops in the sequence can enhance the long-term soil organic matter levels and overall soil quality. Higher levels of soil organic matter also help break the force of rain drops, especially when there is residue on top of the soil surface, and contribute to water storage by acting like a large sponge above and within the total soil mass. Thus soil organic matter contributes nutrients and increases water-holding capacity.

Organic systems depend on legumes as a key to maintaining soil nitrogen and other nutrients at levels adequate for successful crop and pasture-plant growth. The atmosphere is approximately 78% elemental nitrogen, and this can be fixed in plants by symbiotic bacteria, which live on the roots of legumes, and made available to subsequent crops when nitrates are excreted into the soil solution or when the legumes or parts of them are incorporated into the soil as green manure or as crop residues. In conventional farming, nitrogen is supplied primarily by application of chemical forms of nitrate, and higher levels applied in this way actually suppress the nitrogen fixation by legumes. It is possible to fix more than 100–150 kg ha$^{-1}$ nitrogen in 1 year from an appropriate legume such as alfalfa or clover. It is also possible to lose 25–150 kg ha$^{-1}$ of nitrogen in 1 year if this is applied as a soluble chemical, especially if there are high-rainfall events and low-organic-matter levels in the soil. In organic farming systems, the key is to fix and store adequate nitrogen for crop or pasture needs, while not having an excessive amount from legumes that exceed crop needs. The soluble nitrate ions are susceptible to leaching during rainfall events that exceed the ability of soil to hold this moisture.

Elements such as phosphorus, potassium, sulfur, calcium, magnesium, and others are also important for crop and pasture growth, but are less soluble and thus more tightly stored in the soil profile. Potassium is adequate in most agricultural soils and not easily leached out, and in pastures most is returned to the soil by grazing ruminants. Phosphorus is highly insoluble, and its availability is enhanced by the presence of fungal species called 'mycorrhizae' that infect plant roots and help in absorption of phosphate. Since there are often adequate supplies of phosphorus deep in the soil profile, deep-rooted crops with a strong tap root can 'mine' this element from lower soil strata. These elements and others can be supplied from organic sources, especially straw or feed that is cycled back to the field as manure from animals, and again the balance in the system is the important goal. A smaller amount of these other elements is needed for plant growth compared with nitrogen, but their management is no less important due to serious effects of deficiencies in some soils. A number of soil-management practices, such as contour planting and reduced tillage, are useful to preserve nutrients in the field, in both organic and conventional production.

## Protecting Plants against Pests

Similar to the strategies used for maintaining soil fertility, there are comparable lessons from natural ecosystems that help inform our design of agroecosystems to tolerate debilitating or completely destructive pest populations. Many of these principles relate to cropping diversity in each field and on the farm. Genetic resistance to insects and pathogens are commonly used in both organic and conventional farming systems. Well-adapted and healthy crop, pasture, and domestic animal species with tolerance to prevalent pest problems promote the least-expensive control methods, since this alternative may require no further management costs after the crop is planted or the basic animal enterprise is established. Other practices can be incorporated into the management strategy that will control or minimize losses to pests (see **Table 3**).

**Table 3** Pest-management practices and processes used by organic farmers to maintain crops, pastures, and domestic animals in a cost-efficient and environmentally sound manner

| *Practice or process* | *Description and impact on pest populations and production* |
|---|---|
| Crop rotation | Breaks the reproductive cycle of many pest species, provides natural controls through diverse crops in most available niches |
| Cultivation | Reduces weed populations by mechanical means and reduces their potential for reproduction and spread |
| Competitive species | Choice of highly competitive crop and pasture species and varieties that will use resources quickly and outgrow weeds |
| Resistant crops | Genetically resist insects and pathogens that would normally attack susceptible species or crop varieties |
| Planting dates | Adjustment can allow additional cultivation prior to planting (weeds) or avoidance of problems (insects) |
| Strip-cropping | Adds spatial diversity to field, habitats for some predators and parasites to aid control of unwanted species |
| Intercropping | Diversity in the field provides more competition for weeds, also reduces pest outbreaks due to resistance of many individual plants |
| Allelopathy | Planting certain crops such as winter rye, sorghum, and cucurbits that suppress weeds through chemicals excreted by roots or leaves |
| Summer–winter crop | Rotation of crops that grow in different seasons disrupts weed reproductive cycles and allows mechanical control |
| Annual/perennial | Rotation of crops with different life cycles helps control weeds that become more abundants in either species grown continuously |
| Genetic resistance | Crop or animal resistance or tolerance to pathogens or insect problems is an effective and inexpensive means of control |
| Floral fencerows | Field borders that are biologically diverse and provide a range of flowering species can harbor predators that reduce crop pests |

Crop rotations are desirable in most production systems, and these are both prescribed in organic rules and widely used to enhance soil fertility, reduce weed populations, and protect against many insects and plant pathogens. Rotations of unalike crop species will favor different weed species in each season or year and thus break the reproductive cycles of these weed species or aid in their control by mechanical means. Different summer crops such as maize and soybean have different growth habits and timing of competition with weeds, their root systems, and patterns of nutrient uptake, thus favoring a somewhat distinctive weed complex. Rotations of summer with winter crops further exploit this principle, since winter annual weeds that develop with the winter crop can be easily controlled in a subsequent summer crop, and vice versa. Rotating annual crops with perennial crops also provides a means of weed management, since the perennial weeds that develop with perennial crops can often be controlled mechanically in the next annual crop. Lastly, allelopathic species such as winter rye can suppress many weed species that will not be a problem in the succeeding summer annual crop. Cucurbits such as squash grown with maize and beans in the well-known, historical three-sisters crop combination in Central America control many weeds both by shading and by exudates from the leaves and roots. Crop varieties that are rapidly growing early in their cycle can quickly shade weeds and provide a measure of control. These are all methods used by organic farmers. (*See* **Crop Rotations**.)

Spatially diverse cropping patterns such as strip-cropping and intercropping are used to diversify the field planting pattern, confuse some insects that enter from the air, and disrupt their movement once they have entered the field. These patterns in organic agriculture may interrupt or discourage pest reproduction and reduce damage to crops. Pathogens likewise may have more difficulty in dispersion when they occur in a diverse system, compared with their rapid dispersion in monoculture. Diverse agroecosystems may include the added diversity potential of flowering wild species in fencerows, windbreaks, and field borders, providing habitats for predators and parasites that also help control unwanted pests in the adjacent production fields.

These combinations of genetic resistance, spatial and temporal diversity, and different crops in the field can be combined with adequate nutrient management to produce healthy organic crops. When crops and pasture species are growing rapidly, they often outgrow competition from small weed populations and are more resistant to insect pests and plant pathogens. Likewise, a healthy herd or flock of domestic animals is less susceptible to disease or insect problems. When these animal species in organic systems are allowed free range in pastures and crop residues, they also have fewer problems than the same species raised in confinement. Organic production practices favor this type of outdoor management whenever possible for domestic animals, and appropriate regulations assure the humane treatment of all livestock.

## Conversion to Organic Production and Certification

The transition or change from conventional production systems to an organic system is called 'the conversion process,' and this requires 2 years in the UK and 3 years in the USA to complete before a farmer can apply for certification. Most advisors who have worked with farmers say that the most important and often difficult step in the conversion is a psychological change, an embracing of the concept that nonchemical farming is possible, achievable, practical, and economically viable. Armed with this confidence, gained by talking with current organic farmers, reading, attending workshops, or finding motivation from some other source, a farmer is ready to undertake the often considerable changes in farming system design and attendant practices that are required for certification.

Often a number of innovations or major changes are needed in planning the system. It is necessary to plan a conversion to nonchemical methods of providing adequate nutrients for crop and pasture growth. Most often this requires introducing more legumes and green-manure crops, or putting fields to pasture. Farmers often embark upon conversion by planting a forage crop that has no additional nutrient needs or complicated protection against pests. Animal numbers need to be adjusted to fit the carrying capacity of the available area of the farm, since it is highly desirable to minimize the amount of purchased feed or fodder that increases costs. In European countries, there are rules about the amount of nonorganic grains and fodder that can be used for animals while maintaining organic certification. There are also rules in the USA about what is permitted in animal production, and, although these vary somewhat among the private certification groups, there is a basic standard in the new organic legislation that must be followed by all groups. In the overall organic system, there needs to be a balance of areas in crops, in pastures, and in green-manure species to provide a sustainable level of fertility for optimum crop and animal growth, and profitability.

This balance may be achieved by either converting the entire farm at one time, or by converting a fraction of the farm each year over a period of years. The former is desirable because it is possible to take advantage of organic premiums in the marketplace much sooner if all fields are certified after the minimum waiting period. However, it may be too difficult to convert the entire farm at once if this requires planting forage crops in every field and there is not an adequate market for this product. Also, if there is an expected reduction in production until the organic system is established and in balance, it may be too difficult financially to make such a change in all fields at the same time. A more rational approach may be to convert some proportion of the farm each year, perhaps 33% or 25%, so that the process can be completed in 3–4 years. This spreads the risk as well as the reduced income over a longer period of time, and allows for the learning that must take place in making necessary changes.

In addition to changes in crops, cropping sequence, and use of inputs, there are likely to be modifications needed in field equipment, additions of facilities for processing or packaging products, and the development of innovative marketing channels. These changes are not all necessary for every farmer who converts to organic production, and the amount of modification in practices and sales options depends on the number of new enterprises added, the distance to markets and complexity of marketing, and the decisions to add more value to products by some direct sales. If there is a modest change in the balance of crops already grown on the farm, it is possible that no new field equipment will be needed. If there is diversification of current enterprises it is likely that additional equipment for new and specialized uses will be essential. How much processing, packaging, and direct sale will take place on the farm will determine whether additional buildings or equipment are needed for these activities. There is always an option of leasing equipment, time-sharing with neighbors, or custom work on operations that would not justify purchase of equipment for a specialized job. Yet this reduces the opportunity to bring return to a farmer's labor, and most operations are probably done more economically by the owner or manager of each farm. Each of these steps may require new information, and this can be acquired by reading, from the Internet, or from workshops and classes (offered by Cooperative Extension in the USA), or through a certifying organization. People starting in organic farming may underestimate the need for additional information that will help them make the transition and avoid catastrophes in the field.

## Biodynamic Farming, Biological Farming, and Integrated Systems

Several other schools of thought and specific groups have contributed to current philosophy and practices in organic farming. Biodynamic farming grew from the lectures and philosophy of Rudolph Steiner in Germany, founder of the movement called 'anthroposophy.' His teachings embrace a holistic and spiritual interpretation of agricultural systems, and especially the central and special role of humans in those systems. In this movement, the farm is viewed as a self-contained and dynamic organism that can be managed using internal resources and special chemical preparations (called 'preparats') that are essential to the health and vitality of composts, the farm, and its surroundings. The practices reflect a belief in cosmic as well as terrestrial forces on the biological rhythms on the farm. Although most organic farmers do not accept the tenets of Steiner, they do embrace the holistic view of the farm, the rejection of chemical fertilizers and pesticides, and the sound management techniques that were fundamental to the original lectures.

Biological farming includes a range of other movements from the USA, Europe, and India that focus on efficient use of animal manures and compost, green-manure crops, and careful attention to nutrients and soil fertility. Concentration on the humus fraction of soil and its importance in supplying nutrients is central to most systems, and soil health is often related to livestock and human health for those practicing the systems. Such studies have led to more research on the composting of human wastes, a practice long accepted in Asia but not a part of most European or Western Hemisphere fertility strategies in recent years. These systems evolved to include integrated management, a focus that continues today. Integrated farming is currently a strategy that takes into account balance of livestock and crop production, organization of the whole farm into a system that cycles nutrients and provides efficient production of several enterprises, and embraces diversity in those plans. It is easy to see how many of these approaches and strategies for farming overlap and complement each other, although they have different names and are vigorously pursued as different strategies by their supporters.

## Marketing and Consumption of Organic Products

The organic food industry is growing rapidly in response to consumer demand and due to the new availability of organic crops and livestock. Countries with the largest areas in organic production include Australia (7.5 million ha), Argentina (3 million ha), Italy (1 million ha), and the USA (900 000 ha) in the year 2000. Those countries with the highest per capita consumption of organic products include Denmark, Switzerland, and Austria. Approximately two-thirds of organic foods marketed are fresh vegetables and fruits, with lower sales in processed and packaged foods. The majority of sales still take place from natural-food stores, although this is changing rapidly as the mainstream supermarkets explore this new and enticing market with their customers. There has been a lack of balance between supply and demand in organic food, for example in the USA there has been more demand than supply, and many farmers are increasing production to meet this demand by the public. In some European countries, such as Germany, there has been adequate production for domestic markets and enough available for organic cereal export to the Nordic region. National planners and farmers in this region are responding by moving toward incentives for producing ecologic grains that can substitute for these imports.

Premiums for organic products are highly variable, ranging from less than 10% where there is limited demand to more than 100% when a product is scarce. A review by Donald Lotter illustrates this spread in great detail. Fresh products in the US market for an average of 60% higher price than comparable conventional foods, while processed organic products may cost 100–300% more. In the European Union (EU) these price spreads are more in the order of 10–50%, depending on the product, country, and nature of subsidies for production. These subsidies for both the conversion process and for continuing to grow organic products have helped to narrow the differences in prices between the two. However, the support programs differ among countries, although the EU is slowly making these more uniform, and are so complicated that a Swedish farmer was heard to say that the most important advisor in his organic farming business was the one who could help him fill out all the EU forms for subsidies. In this realm, both organic and conventional farmers are in the same situation, with a high dependence on external subsidies to maintain profitability.

Consumer demand for organic products has been studied in a number of countries. It is generally found that more affluent customers, those with higher levels of education and smaller families, are more apt to buy organic food. Consumers appear to be motivated by two major factors, that organic foods are more healthy for them and that the organic systems that produce them are more environmentally safe. A recent consumer survey in Hedmark County in Norway found that people who regularly purchased organic food were similar in most ways to those who purchased conventional food products, although they were more concerned about conditions under which the food was produced, the importance of local production, and animal welfare on the farm. There is a general concern about GMOs (transgenically modified organisms) among those who practice organic farming and those who purchase the products, and these new crops and performance-enhancing products for livestock are not allowed by any of the organic certification organizations, nor in the new national US legislation on organic labels.

There is great public interest and continuing debate about the safety and the health benefits of organic food products compared with those considered conventional. There is no question about the research, and surveys that have shown less pesticide residue (most often zero) on organic than on conventional produce. Thus organic food is a safer product than the general offering of vegetables and fruits in the marketplace. In terms of food value, there are numerous claims by organic proponents of better taste, freshness, and nutritional quality of organic foods. These qualities are difficult to substantiate, although it is obvious when produce comes from a local farmer and is purchased quickly and directly by the consumer that it will be fresher and probably have a better taste than a comparable product that has traveled 1000 km or 1500 miles (average distance that US food travels from production to consumer). Definitive clinical comparisons of nutritive value generally fail to show differences between organic and conventional food, results that are broadcast by the large food companies to support their claims of no differences and despaired of by organic food enthusiasts who are convinced of the nutritive value of their chosen products. This is still a valuable area for further research.

## Future Perspectives for Organic Farming and Foods

If consistent growth in markets over a period of nearly two decades is any indication, there is a large potential for the expansion of organic food production and sale. One of the best indications of the growth potential for this sector is the current interest of larger corporations and multinationals in acquiring smaller organic processing and distributing companies. In the International Federation of Organic Agriculture Movements (IFOAM) Congress held in Victoria, BC, Canada, in August of 2002, there was extensive debate about the entrance of large corporations into the production and marketing of organic foods. The concern in traditional organic circles is that large corporations (both large farms and large processors and marketers) may fulfill the legal requirements of no pesticides or chemical fertilizers, thus producing a 'clean' product, but they are not in step with the goals of the majority of smaller organic producers to create local food systems that will promote the local food economy and establish personal relationships between farmers and producers. This industrialization of organic foods is likely to be the center of long debates in the near future, as people weigh the importance of organic certification versus food produced locally that meets more of the social goals that are part of the original organic agriculture movement.

Other questions embedded in the organic farming debate include the value and desirability of subsidies for organic production, whether price premiums will disappear if organic foods are widely grown, and the potentials for organic agriculture to feed a growing world population. Proponents of the completely free market argue that consumers will sort out food values in the marketplace, and thus there is no need for subsidies. Opponents argue that there are additional costs and limited research to support their conversion to organic farming, and financial incentives are still needed for farmers to begin organic farming to meet the obvious demands. If there is widespread adoption of organic farming techniques and there are many more organic products on the market, it is likely that price differentials will be reduced, but it is unlikely that they will disappear completely. On the issue of global food production, the best organic farmers today produce yields equal to those on conventional farms, although average production is 10–15% lower on organic farms. Part of this is due to lack of research and adequate extension recommendations in organic systems, and part is due to a lower intensity of resource use in organic farming where there is dependence on internal resources and limited import of fossil fuel-derived materials from outside the farm. An organic farming strategy can be profitable if the production costs are lower and if price premiums are available. There is general agreement that the major problem contributing to hunger on a global scale is inadequate distribution of benefits from today's food system, i.e., that poverty is more to blame than lack of production. This will not be solved by producing more, argue some supporters of organic farming; and we can save the environment by reducing applications of pesticides and fertilizers that reduce water quality and quality of life for many rural residents.

The unique model of Cuba illustrates what can be done when fossil fuel supplies are severely limited in agriculture, and an entire country embraces organic farming and gardening as a major strategy for local food production. The majority of food consumed in major cities in Cuba – fruits, vegetables, small animals and their products – is produced within or adjacent to the city limits. Most are marketed directly from farmers to consumers. In 2002 Cuba enjoyed a healthy diet based on this food system, after some difficult years of adjustment or 'conversion' in their entire agricultural system. This is one model embedded in a unique political system, including diversity in land ownership, wide participation in food production, and a reliance on primarily internal resources to produce an adequate diet for the entire population. In conclusion, this is a model that should be studied as a specific case where organic farming is highly successful.

*See also:* **Crop Rotations**

## Further Reading

Altieri MA, Nicholls CI, and Itiel JI (1994) *Biodiversity and Pest Management in Agroecosystems.* Binghampton, NY: Food Product Press/Haworth Press.

Bender J (1994) *Future Harvest: Pesticide-Free Farming.* Lincoln, NE: University of Nebraska Press.

Carson R (1962) *Silent Spring.* New York: Houghton Mifflin.

Daily GC (ed.) (1997) *Nature's Services: Societal Dependence on Natural Ecosystems.* Washington, DC: Island Press.

Edwards CA, Lal R, Madden P, Miller RH, and House G (eds) (1990) *Sustainable Agricultural Systems.* Ankeney, IA: Soil & Water Conservation Society.

Francis CA, Flora CB, and King LD (eds) (1990) *Sustainable Agriculture in Temperate Zones.* New York: John Wiley.

Funes F, Garcia L, Bourque M, Perez N, and Rosset P (eds) (2002) *Sustainable Agriculture and Resistance.* San Francisco, CA: Food First Books.

Howard A (1943) *An Agricultural Testament.* London, UK: Oxford University Press.

King FH (1911) *Farmers of Forty Centuries, or Permanent Agriculture in China, Korea, and Japan.* Emmaus, PA: Rodale Press.

Lampkin N (1990) *Organic Farming.* Ipswich, UK: Farming Press.

Lampkin N and Measures M (eds) (2001) *2001 Farm Management Handbook,* 4th edn. Organic Farming Unit, Welsh Institute of Rural Studies. Aberystwyth, UK: University of Wales.

Lampkin N and Padel S (1996) *Economics of Organic Farming: An International Perspective.* Wallingford, UK: CAB International.

Lotter D (2002) Organic farming. *Journal of Sustainable Agriculture* 21(4): 59–128.

Rodale JI (1995) *Organic Gardening: How to Grow Healthy Vegetables, Fruits, and Flowers Using Nature's Own Methods.* Garden City, NJ: Hanover House.

Rodale R (1983) Breaking new ground: the search for a sustainable agriculture. *The Futurist* 1(1): 15–20.

Sooby J (2003) *State of the States: Organic Systems Research at Land Grant Institutions, 2001–2003.* Santa Cruz, CA: Organic Farming Research Foundation.

Soule JD and Piper JD (1993) *Farming in Nature's Image.* Covelo, CA: Island Press.

Steiner R (1958) *Agriculture: A Course in Eight Lectures.* London, UK: Biodynamic Agriculture Association.

# ORGANIC MATTER

## Principles and Processes

**M Schnitzer**, Agriculture and Agri-Food Canada, Ottawa, ON, Canada

### Introduction

The presence of organic matter (OM) distinguishes the soil from a mass of rock particles and allows it to become a living system. Soil organic matter (SOM) serves as a soil conditioner, nutrient reservoir, substrate for microbial activity, preserver of the environment, and major determinant for sustaining and increasing agricultural productivity. The OM content of soils ranges from less than 1% in desert soils to close to 100% in organic soils. A typical agricultural soil may contain between 1 and 5% OM in the top 15 cm.

The main emphasis of this article is on the analytical characteristics of SOM fractions and on studies on their chemical structure. Based on the results of extensive researches, a two-dimensional humic acid (HA) model structure was formulated which has been subsequently converted, with the aid of computational chemistry, to a three-dimensional HA model structure.

The term 'SOM,' as used in this article, refers to the sum total of all organic carbon-containing substances in soils. The major components of SOM are humic substances, especially HA; other significant components are carbohydrates, proteinaceous materials, and lipids.

### Extraction and Fractionation of Humic Substances

#### Extraction of Humic Substances from Soils

Because, in soils, humic substances and inorganic soil constituents are closely associated, it is necessary to separate the two before either can be examined in greater detail. The most efficient extractants for humic substances are either dilute base (0.1–0.5 mol l$^{-1}$ NaOH solution) or dilute neutral salt solutions such as aqueous 0.1 mol l$^{-1}$ $Na_4P_2O_7$.

#### Fractionation of Extracted Humic Substances

Most soil OM chemists partition the humic substances extract into the following three fractions: (1) HA, which is that fraction of the extract that coagulates when the extract is acidified to pH 2; (2) fulvic acid (FA), which is that fraction of the extract that remains in solution when the alkaline extract is acidified, that is, it is soluble in both alkali and acid; and (3) humin, which is that humic fraction that remains with inorganic soil constituents, that is, it is insoluble in both alkali and acid.

A number of objections have been raised to the use of alkaline extractants. The latter may extract silica from minerals, protoplasmic components from organic tissues, and bring about the auto-oxidation of some organic components, and the condensation of amino groups of amino acids with carbonyl groups of reducing sugars to form Maillard-type reaction products. Some of these changes can be minimized by doing the extractions in atmospheres of inert gases, but not all possible changes can be excluded.

### *In Situ* Analysis of SOM

Cross-polarization magic angle spinning $^{13}C$ nuclear magnetic resonance (CP-MAS $^{13}C$ NMR) and pyrolysis-field ionization mass spectrometry (Py-FIMS) are used for the *in situ* analysis of OM in soils. These methods identify the major chemical components of SOM without extraction and fractionation and yield valuable information on the main chemical structures of these materials. The major chemical components of SOM that have so far been identified include carbohydrates, phenols, lignin monomers and dimers, saturated and unsaturated fatty acids, alkanes and alkenes, alkyl monodi- and triesters, alkyl benzenes, methyl naphthalenes, methyl phenanthrenes, amino acids, and heterocyclic N-compounds. While the *in situ* analysis of OM in whole soils by the two methods referred to above does away with the extraction of OM from soils and its subsequent fractionation into HA, FA, and humin, it defines instead SOM in terms of its chemical makeup. In contrast to earlier workers, who thought that HA and FA were chemically pure substances that were well separated by the extraction procedure, we know now from chemical, $^{13}$C NMR, and mass spectrometric analyses that the assumptions of workers prior to the early years of the twentieth century are incorrect. HA, FA, as well as humin, are not distinct chemical substances. Each of these fractions consists of hundreds of compounds, which appear to be associated at molecular levels by mechanisms not yet well understood.

If soil chemists want to retain the extraction and fractionation approach, they need to define each fraction not on the basis of color and solubility or insolubility in base or acid, but by its $^{13}$C NMR spectrum, its infrared (IR) spectrum, its mass spectrum, its elemental and functional group analyses, and so on.

## Analytical Characteristics of HAs and FAs

### Elemental Composition and Functional Group Content

The elemental composition and oxygen-containing functional group analysis of a typical HA (extracted from the Ah horizon of a Haploboroll) and a FA (extracted from the Bh horizon of a Spodosol) are shown in Table 1.

A more detailed analysis shows that: (1) the HA contains approximately 10% more C, but 36% less O than the FA; (2) there are quantitatively smaller differences between the two materials in H, N, and S contents; (3) the total acidity and COOH content of the FA are significantly higher than those of the HA; (4) both materials contain per unit weight significant concentrations of phenolic OH, alcoholic OH, and ketonic and quinonoid C=O groups; (5) the HA is richer than the FA in quinonoid C=O groups, while both materials contain relatively few $OCH_3$ groups; (6) about 74% of the total O in the HA is accounted for in functional groups, but all of the O in the FA is similarly distributed; and (7) the $E_4/E_6$ ratio (ratio of absorbances at 465 and 665 mm) of the FA is almost twice as high as that of the FA, indicating that the FA has lower molecular weight than the HA.

**Table 1** Analytical characteristics of a Haploboroll humic acid and a Spodosol fulvic acid

| | *HA* | *FA* |
|---|---|---|
| *Element (g kg$^{-1}$)* | | |
| C | 564 | 509 |
| H | 55 | 33 |
| N | 41 | 7 |
| S | 11 | 3 |
| O | 329 | 448 |
| *Functional groups (cmol kg$^{-1}$)* | | |
| Total acidity | 660 | 1240 |
| COOH | 450 | 910 |
| Phenolic OH | 210 | 330 |
| Alcoholic OH | 280 | 360 |
| Quinonoid C=O | 250 | 60 |
| Ketonic C=O | 190 | 250 |
| $OCH_3$ | 30 | 10 |
| $E_4/E_6$ | 4.3 | 7.1 |

HA, humic acid; FA, fulvic acid.

Reproduced with permission from Schnitzer M (2000) A lifetime perspective on the chemistry of soil organic matter. *Advances in Agronomy* 68: 1–59, © Academic Press, San Diego.

### Infrared and Fourier Transform Infrared Spectrophotometry

IR and Fourier transform IR (FTIR) spectra of HAs and FAs show bands at 3400 cm$^{-1}$ (H-bonded OH) 2900 cm$^{-1}$ (aliphatic stretch), 1725 cm$^{-1}$(C=O of COOH, C=O stretch of ketonic C=O), 1630 cm$^{-1}$ (COO$^-$, C=O of carbonyl and quinone), 1450 cm$^{-1}$ (aliphatic C–H), 1400 cm$^{-1}$ (COO$^-$), 1200 cm$^{-1}$ (C–O stretch or OH deformation of COOH), and 1050 cm$^{-1}$ (Si–O of silicates). The bands are usually broad because of extensive overlapping of individual absorbances. IR and FTIR spectra of HAs and FAs reflect the preponderance of oxygen-containing functional groups, i.e., COOH, OH, and C=O in these materials. Some FTIR spectra show the presence of C=C and polysaccharides. While IR and FTIR spectra provide useful information on the functional groups and their participation in metal–humic as well as pesticide–humic interactions, they tell us little about the chemical structure of these materials.

## Chemical Structure of SOM

### $^{13}$C NMR Spectroscopy

One of the most important methods currently available for elucidating the chemical structure of organic substances is $^{13}$C NMR. For the analysis of SOM and fractions derived from it such as HA, FA, and humin as well for the direct *in situ* analysis of OM in whole soils, CP-MAS or solid-state $^{13}$C NMR is especially useful. CP-MAS $^{13}$C NMR spectra of a Mollisol HA and a Spodosol FA are shown in **Figure 1**. Both the HA and FA show distinct resonances in the aliphatic (0–105 ppm), aromatic (106–150 ppm), phenolic (155–169 ppm), and carboxyl (170–190 ppm) regions. The signals in the HA spectrum at 17, 21, 25, 27, and 31 ppm are probably due to alkyl C. The resonance at 17 ppm is characteristic of terminal $CH_3$ groups and that at 31 ppm of $(CH_2)_n$ in straight paraffinic chains. The peak at 40 ppm could include

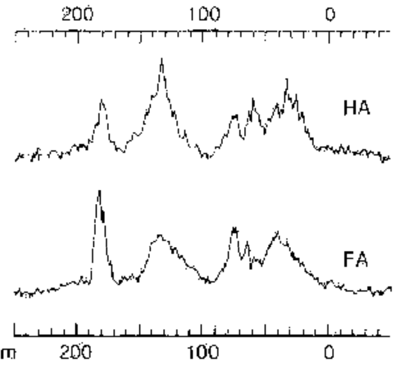


**Figure 1** Cross-polarization magic angle spinning $^{13}C$ nuclear magnetic resonance (CP-MAS $^{13}C$ NMR) spectra of HA (extracted from the Ah horizon of a Haploboroll and FA (extracted from the Bh horizon of a Spodosol). (Reproduced with permission from Schnitzer M (2000) A lifetime perspective on the chemistry of soil organic matter. *Advances in Agronomy* 68: 1–58. © Academic Press, San Diego.)

**Table 2** Distribution of C (%) in a Haploboroll HA and a Spodosol FA as determined by $^{13}C$ NMR

| | *% of C* | |
|---|---|---|
| *Chemical shift range (ppm)* | *HA* | *FA* |
| 0–40 | 24.0 | 15.6 |
| 41–60 | 12.5 | 12.8 |
| 61–105 | 13.5 | 19.3 |
| 106–150 | 35.0 | 30.3 |
| 151–170 | 4.5 | 3.7 |
| 171–190 | 10.5 | 18.3 |
| Aliphatic C (0–105 ppm) | 50.0 | 47.7 |
| Aromatic C (106–150 ppm) | 35.0 | 30.3 |
| Phenolic C (151–170 ppm) | 4.5 | 3.7 |
| Aromaticity[a] | 44.1 | 41.6 |

[a] ((Aromatic C + phenolic C)/(aromatic C + phenolic C + aliphatic C)) × 100.
Reproduced with permission from Schnitzer M (2000) A lifetime perspective on the chemistry of soil organic matter. *Advances in Agronomy* 68: 1–58. © Academic Press, San Diego.

contributions from both alkyl C and amino acid C. The broad signal at 53 ppm and the sharper one at 59 ppm appear to be due to $OCH_3$. Amino acid C may also contribute to this region. Carbohydrates in the HA would be expected to produce signals in the 60–65, 70–80, and 90–105 ppm regions; also other types of aliphatic C bonded to O could do so. The aromatic region contains a relatively sharp peak at 130 ppm due to alkyl aromatics. The signal at 155 ppm shows the presence of O- and N-substituted aromatic C (phenolic OH and/or $NH_2$ bonded to an aromatic C). The broad signal near 180 ppm is due to C in COOH groups, although amide C and C in esters could also contribute to this resonance.

The CP-MAS $^{13}C$ NMR spectrum of the FA (**Figure 1**) consists of a number of aliphatic resonances in the 20- to 50-ppm region, followed by signals due to C in $OCH_3$ groups, and C in amino acids. Signals due to carbohydrate C are present between 50 and 85 ppm. Broad signals between 130 and 133 ppm indicate the presence of C in alkyl aromatics. The strong signal between 170 and 180 ppm shows the presence of C in COOH groups. In general, fewer sharp signals are observed in the spectrum of the FA than that of the HA, possibly because of more H-bonding in the FA.

The CP-MAS $^{13}C$ NMR data for the HA and FA are summarized in **Table 2** in terms of the distribution of C in the different spectral regions. An examination of the data shows similar C distributions in the two humic materials. The HA is slightly more aromatic than the FA, but the FA is richer in COOH groups, which appears to be the main difference between the two substances. Other differences are that the HA is richer in paraffinic C but poorer in carbohydrate C than the FA. However, on the whole, the main structural features, as well as the aromaticity and aliphaticity, are similar so that the HA and FA have similar chemical structures.

As far as the CP-MAS $^{13}C$ NMR spectrum of humin, that portion of SOM that stays behind after extraction of the soil with dilute alkali or neutral salt solution, is concerned, the spectrum, after extensive leaching, is very similar to that of HA. This indicates that humin is essentially HA-bound strongly to soil minerals.

## Electron-Spin Resonance Spectroscopy

Humic substances contain free radicals that can participate in a wide variety of organic–organic and organic–inorganic interactions. The electron-spin resonance (ESR) spectrum of a typical HA consists of a single line devoid of hyperfine splitting. From the magnitudes of the *g*-values (the spectroscopic splitting constant), the prominent free radicals in humic materials appear to be semiquinones or substituted semiquinones. The latter can be produced either by the reduction of quinones or the oxidation of phenols. Except for indicating the presence of phenols, semiquinones, and quinones in these materials, ESR spectroscopy has so far contributed little to our understanding of the structural chemistry of these substances. The main reason for this is that it has been difficult to split the signal.

It is also known that there are two types of free radicals in humic materials: (1) permanent free radicals with long lifetimes, and (2) transient free radicals with relatively short lives (several hours). Transient free radicals in humic substances can be generated in relatively high concentrations by chemical reduction, irradiation, or increases in pH.

### Oxidative Degradation

The oxidative degradation of HAs, FAs, humins, and whole soils under a wide variety of experimental conditions produces aliphatic carboxylic, phenolic, and benzene carboxylic acids. Among aliphatic oxidation products are mono-, di-, tri-, and tetracarboxylic acids. Major phenolic acids include aromatic rings substituted by one to three OH groups and by one to five COOH groups, while major benzene carboxylic acids are the di-, tri-, tetra-, penta-, and hexa-forms. From the oxidation products identified and from CP-MAS $^{13}C$ NMR spectra of humic substances, it appears that the aromatic rings are cross-linked by paraffinic chains. Oxidation converts $CH_3$ and $CH_2$ groups closest to the aromatic rings to COOH groups, whereas $CH_2$ groups in alkyl chains are oxidized to mono- and dicarboxylic acids and/or to $CO_2$. The formation of significant concentrations of $CO_2$ from the oxidation of side chains may explain the low oxidation yields of aliphatic acids compared with phenolic and benzene carboxylic acids.

The model structure shown in **Figure 2** has an aromaticity of 50% if functional groups are excluded. Of special interest are the following: (1) isolated aromatic rings are important structural components of all humic substances; (2) aliphatic chains link aromatic rings to form an alkyl aromatic network; and (3) the structure contains voids of various dimensions that can trap organic and inorganic soil constituents. These characteristics are typical of soil humic substances.

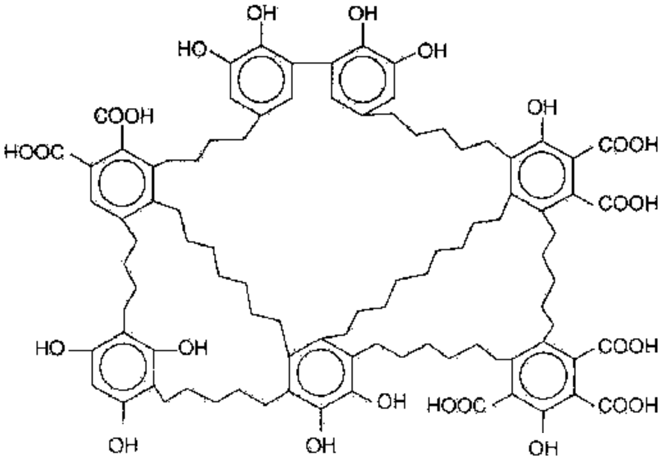


**Figure 2** Chemical structure of humic substances based on oxidation products. (Reproduced with permission from Schnitzer M (2000) A lifetime perspective on the chemistry of soil organic matter. *Advances in Agronomy* 68: 1–58. © Academic Press, San Diego.)

### Reductive Degradation

Relatively mild reduction of humic substances with Na amalgam produces phenols and phenolic acids. But the harsher Zn-dust distillation and Zn-dust fusion yield methyl-substituted naphthalene, anthracene, phenanthrene, pyrene, and perylene. The methyl groups on the polycyclics are probably the remains of longer alkyl chains linking the polycyclics in HA, FA, and humin structures.

### Pyrolysis-Field Ionization Mass Spectrometry of HAs, FAs, and Humins

Compounds identified in the Py-FI mass spectra of a whole soil, an HA, an FA, and a humin are summarized in Table 3. The most abundant compounds identified in all three fractions are carbohydrates, phenols, lignin monomers, lignin dimers, *n*-fatty acids, *n*-alkyl diesters, and *n*-alkyl benzenes. Minor components include *n*-alkyl monoesters, methylnaphthalenes, methylphenanthrenes, and N-compounds. HA tends to be enriched in *n*-fatty acids and the whole soil and the humin in *n*-alkyl diesters and *n*-alkyl benzenes.

### Curie-Point Pyrolysis–Gas Chromatography/Mass Spectrometry of HAs and FAs

Curie-point pyrolysis–gas chromatography/mass spectrometry (Cp Py-GC/MS) of HAs and FAs indicates the presence of relatively high concentrations of aromatic rings substituted by alkyl chains. Of special significance is the identification of a series of $C_1$–$C_{22}$ *n*-alkyl benzenes. In addition, ethylmethyl benzene, methylpropyl benzene, methylheptyl benzene, methyloctyl benzene, and methylundecyl benzene are also detected. Other compounds identified

**Table 3** Compounds identified by Py-GC/MS in the initial Armadale soil and in HA, FA, and humin fraction isolated from it

| *Compound identified* | *Soil* | *HA* | *FA* | *Humin* |
|---|---|---|---|---|
| Carbohydrates | ++[a] | ++ | ++ | ++ |
| Lingin monomers | ++ | ++ | ++ | ++ |
| Lingin dimers | ++ | ++ | ++ | ++ |
| *n*-Fatty acids | + | +++ | + | ++ |
| *n*-Alkyl monoesters | + | + | + | + |
| *n*-Alkyl diesters | ++ | + | + | ++ |
| *n*-Alkyl benzenes | + |  | + | ++ |
| Methylnaphthalenes | + |  |  | + |
| Methylphenanthrenes | + |  |  | + |
| N compounds | + | + |  | + |
| *n*-Alkanes | + |  |  |  |

[a] +, weak (relative intensity <20%); ++, intense (relative intensity 20–60%); +++, very intense (relative intensity >60%).

Reproduced with permission from Schnitzer M and Schulten H-R (1995) Analysis of organic matter in soil extracts and whole soils by pyrolysis-mass spectrometry. *Advances in Agronomy* 55: 168–217. © Academic Press, San Diego.

are trimethyl- and tetramethyl benzenes, alkyl naphthalenes, and alkyl phenanthrenes. The alkyl substitution on the naphthalenes ranges from one to five methyls, whereas that on the phenanthrene rings ranges from one to four methyls.

### Other Characteristics

From surface pressure, surface tension, and viscosity measurements at different pHs and varying concentrations of HA and neutral electrolytes, it appears that, under conditions normally prevailing in most agricultural soils, HA behaves like a flexible, linear polyelectrolyte. Additional support for this view comes from transmission electron micrographs of HA which show that in dilute aqueous solution HA forms flat, elongated, multibranched filaments, 20–100 nm in width. With increasing pH or HA concentration, a finely woven network of elongated fibers is formed which then coalesces into a sheet-like structure, perforated by voids of varying dimensions. The overall impression is that of a relatively flexible and open structure.

## Model Structures of HA

### Two- and Three-Dimensional Model Structures of HA

The two-dimensional HA model structure of HA shown in Figure 3 is based on long-term chemical, $^{13}$C NMR spectroscopic, degradative, electron-microscopic, X-ray, Py-FIMS, and Cp Py–GC/MS studies. In this structure, *n*-alkyl aromatics play a significant role. Oxygen is present in the form of COOH, phenolic and alcoholic OH, ester, ethers and ketones, whereas nitrogen occurs in nitriles and heterocyclic structures. The carbon skeleton contains voids of various dimensions which can trap and bind other organic and inorganic soil constituents as well as water. The molecular formula of the HA is $C_{308}H_{328}O_{90}N_5$, its molecular weight is 5540 Da, and its elemental analysis is 66.8% C, 6.0% H, 26.0% O, and 1.3% N. It is assumed that carbohydrates and proteinaceous materials are adsorbed on external HA surfaces and in internal voids and that hydrogen bonds play an important role in their

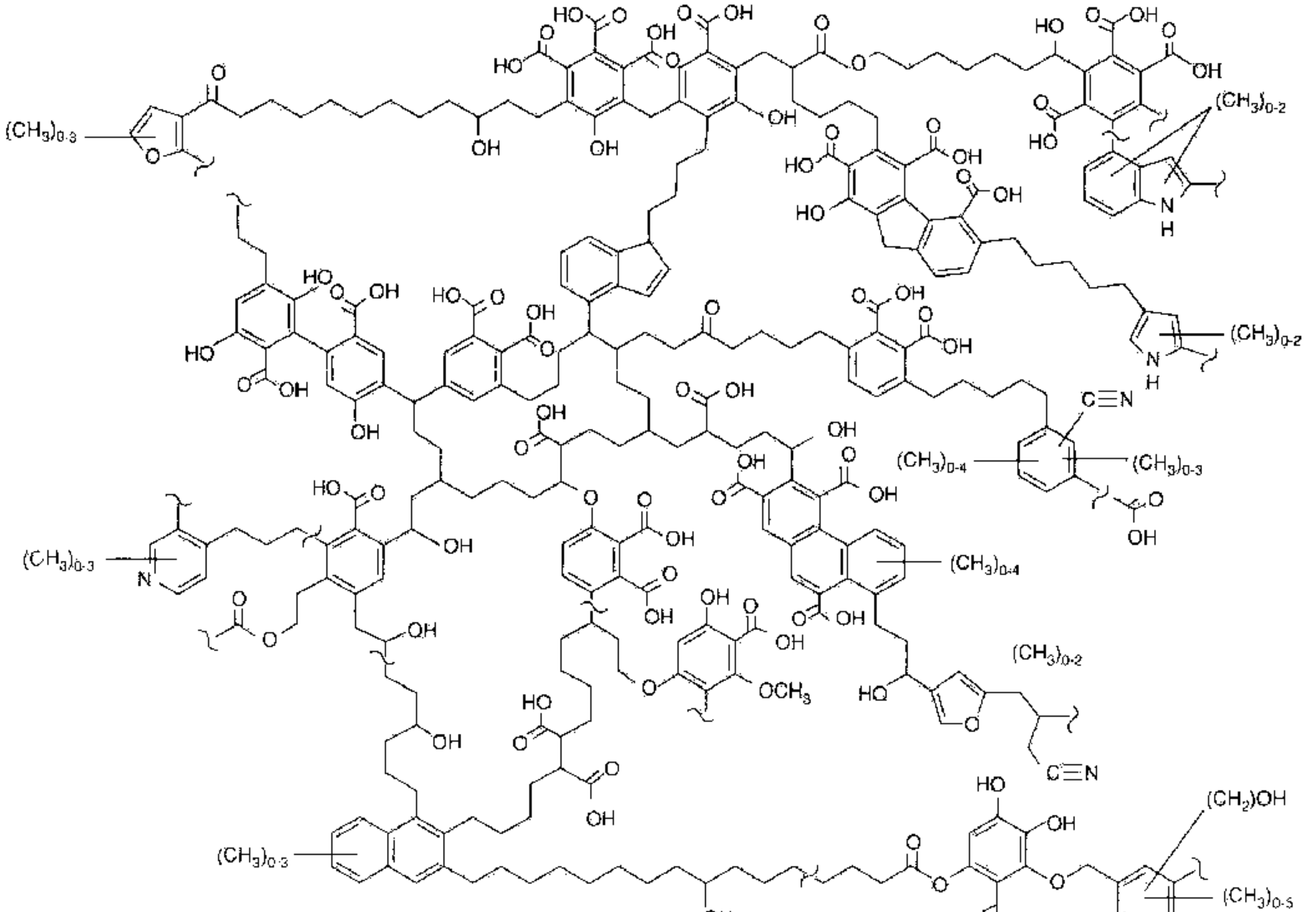


**Figure 3** Two-dimensional HA model structure. (Reproduced with permission from Schulten H-R and Schnitzer M (1993) A state of the art structural concept for humic substances. *Naturwissenschaften* 80: 29–30, Fig. 1. © Springer-Verlag, Heidelberg.)

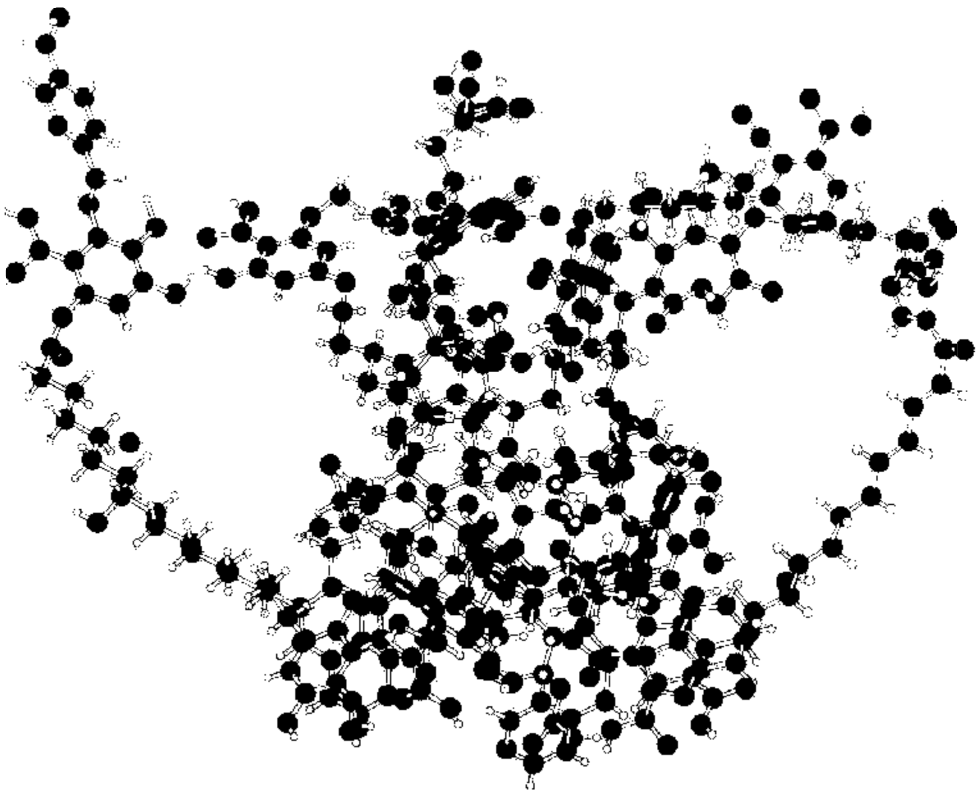


Figure 4 Three-dimensional humic acid model structure. Element colors are: carbon (light gray), hydrogen (white), oxygen (black), and nitrogen (dark gray). (Reproduced from Schulten H-R and Schnitzer M (1997) Chemical model structures for soil organic matter and whole soils. *Soil Science* 162: 115–130, © Lippincott, Williams & Wilkins, Baltimore.)

immobilization. Aside from carbohydrates and proteinaceous materials, the voids can also trap and bind lipids and biocides, as well as inorganics such as clay minerals and hydrous oxides.

The conversion, with the aid of HyperChem software, of the two-dimensional HA structure (**Figure 3**) to a three-dimensional HA structure is shown in **Figure 4**. The three-dimensional HA model structure has a molecular weight of 5547 and an elemental analysis of 66.7% C, 6.1% H, 26.0% O, and 1.3% N.

It is likely that applications of computational chemistry to the development of SOM structural models will increase in the future and lead to a better understanding of the spatial arrangements of the molecular constituents of SOM. A more comprehensive knowledge of the chemistry and reactions of SOM will certainly be beneficial to a sustainable agriculture and help to protect the environment.

## Reactions of SOM with Metals and Organics

### Formation of Water-Soluble Metal–Organic Complexes

Reactions in water near pH 7 between di- and trivalent metal ions and HAs and FAs are likely to proceed via one or more of the mechanism shown in **Figure 5**, taking a divalent metal ion $M^{2+}$ as an example. According to eqn [1] in **Figure 5**, one COOH group on the HA or FA reacts with one metal ion to form an organic salt or monodentate complex. Equation [2] in **Figure 5** describes a reaction in which one COOH and one adjacent OH group react simultaneously with the metal ion to form a bidentate complex or chelate. According to eqn [3] in **Figure 5**, two adjacent COOH groups interact simultaneously with the metal ion to form a bidentate chelate. Equation [4] shows the metal ion $M^{2+}$ linked to the HA or FA not only by electrostatic bonding, but also through a water molecule (water bridge) in its primary hydration shell to a C=O group of the ligand. The latter type of interaction is especially important when the cation has a high solvation energy and so retains its primary hydration shell. Equations [2] and [3] in **Figure 5** describe the formation of strong, inner-sphere complexes, whereas eqn [4] refers to the formation of a weaker, outer-sphere complex. While a number of workers have published stability constants of water-soluble metal–HA and metal–FA complexes, problems have been encountered with the analysis and interpretation of the data. Probably the most serious obstacle to progress in this area is our lack of knowledge of the chemical structures of HAs and FAs, the ligands.

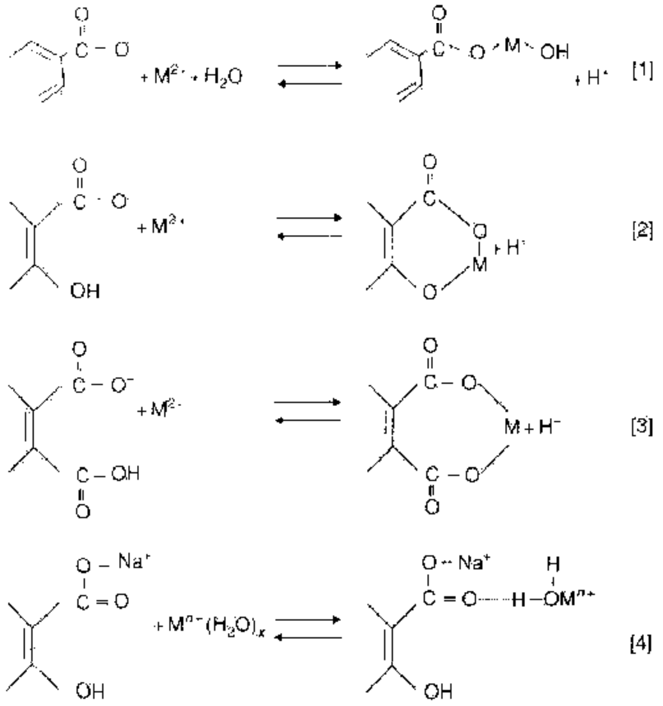


**Figure 5** Major metal–HA and –FA reaction mechanisms. (Reproduced with permission from Schnitzer M (1986) Binding of humic substances by soil colloids. In: Huang PM and Schnitzer M (eds) *Interactions of Soil Minerals with Natural Organics and Microbes*, pp. 77–101. Special Publication No. 17. Madison, WI: Soil Science Society of America.)

### Other Types of Metal SOM Interactions

The formation of mixed-ligand complexes of the type $M^{2+}$–FA – secondary ligand (Y) is known to occur in soils. Secondary ligands (Y) can be citrate tartrate, salicylate, phosphate, bicarbonate, sulfate, aspartate, glycinate, etc. In neutral to weakly acid solutions, these complexes are more stable than simple complexes. The formation of mixed ligand complexes prevents the precipitation of metal ions by hydrolysis at elevated pHs.

Due to their ability to complex di-, tri-, and tetravalent metal ions, dilute aqueous FA solutions and aqueous HA solutions at pH >6.5 can attack and degrade minerals to form water-soluble and water-insoluble metal complexes. Thus, the weathering of minerals in soils and sediments is often enhanced by the action of naturally occurring organic substances, especially FA. Because of its abundance in soils, its solubility in water, and its ability to complex with metal ions and interact with silica, the latter may increase the concentrations of these soil constituents in aqueous solutions to levels that far exceed their normal solubilities. In this manner, aqueous FA solutions may not only bring about the dissolution and degradation of existing minerals, but also lead to the synthesis of new minerals by permitting the complexed and dissolved metals and silica to form new combinations. Conversely, active surfaces of inorganic soil constituents may catalyze either the degradation or synthesis of HAs and FAs.

The extent of adsorption of HA and FA on mineral surfaces depends on the characteristics of the surface, the pH of the system, and its water content. For example, sepiolite, which has a channel-like surface, readily adsorbs FA. In untreated sepiolite, these channels are filled by bound and/or zeolitic water, which can be displaced by undissociated FA in aqueous solution at pH 3.

The interlayer adsorption of FA by expanding clay minerals is pH-dependent, being highest at low pH but no longer occurring at pH >5. The FA cannot be displaced from the clay interlayers by leaching with 1 mol $l^{-1}$ NaCl; an inflection is observed in the adsorption–pH curve near the pH corresponding to the p$K$ of the acid species of FA so that the adsorption can be classified as a ligand-exchange reaction. In this type of reaction, the anion (FA) is thought to penetrate the coordination shell of the dominant cation in the clay and displace water coordinated to the dominant cation in the clay interlayer.

### Nitrogen in Soils and SOM

Almost 95% of total soil N is associated with SOM. This N occurs in the forms of amino acids, peptides, proteins, amino sugars, heterocyclic N compounds, and ammonia. The amino acid composition of soils and SOM from widely differing origins is remarkably similar and resembles that of bacteria. This appears to indicate that microbes play a major role in the synthesis of amino acids, peptides, and proteins in soils. The most prominent amino sugars in soils and SOM are D-glucosamine and D-galactosamine. Present in smaller amounts are muramic acid, D-mannosamine, *N*-acetylglucosamine, and D-fucosamine.

By Py-FIMS and Cp Py-GC/MS, more than 100 heterocyclic N compounds have been identified in soils and SOM. These N compounds include nonsubstituted and substituted pyrroles, pyrrolidines, imidazoles, pyrazoles, pyridines, pyrazines, nitriles, indoles, quinolines, isoquinolines, benzothiazoles, and pyrimidines. Low-mass N-compounds detected are hydrocyanic acid, dinitrogen, dinitrogen monooxide, isocyanomethane, acetamide, and hydrazoic acid. N-derivatives of benzene detected include benzamine, benzonitrile, and isocyanomethylbenzene. According to the latest estimates, the distribution of total N in soils and SOM is as follows: proteinaceous materials (amino acids, peptides, proteins), 40%; amino sugars, 5–6%; heterocyclic N (including purines and pyrimidines), 35%; and $NH_3$-N, 19%. Thus, proteinaceous materials and N heterocyclics are the major N components. While there are indications that the proteinaceous materials and amino sugar are strongly adsorbed on the HA and SOM surfaces and in the voids, the heterocyclic N components appear to be chemically bonded to HA and SOM components; that is, they are integral structural components of HAs and SOM.

### Phosphorus and Sulfur in Soils and SOM

According to current estimates, about 75% of the total P in soils occurs in organically bound forms. Principal organic P forms include inositol phosphates, phospholipids, nucleic acids, phosphoproteins, and metabolic phosphates. Other organic P compounds identified by $^{31}P$ NMR in soils and SOM are alkyl phosphonic esters, alkyl phosphonic acids, choline phosphate, and orthophosphate mono- and diesters.

It is estimated that more than 90% of the total soil S in most noncalcareous soils occurs in organic forms. These include phenolic sulfates, sulfated polysaccharides, choline sulfate, and sulfated lipids. Also present are S-containing amino acids (cystine and methionine), cystathionine, djenkolic acid, taurine, biotin, and thiamine. In poorly drained soils, the decomposition of organic S compounds produces volatile compounds such as carbon disulfide, carbonyl sulfide, methyl mercaptan, diethyl sulfide, dimethyl sulfide, and dimethyl disulfide.

### Interactions of SOM with Pesticides and Herbicides

The persistence, degradation, bioavailability, leachability, and volatility of pesticides and herbicides in soils are directly related to the concentration of SOM. The pesticides and herbicides may sorb on SOM and be retained by van der Waals forces, hydrophobic bonding, hydrogen bonding, charge transfer, ion exchange, and ligand exchange. Pesticides and herbicides may also be adsorbed unchanged on external SOM surfaces or in internal voids. SOM can promote the nonbiological degradation of pesticides and herbicides in addition to chemically binding residues arising from their biological and chemical degradation. These processes may play important roles in the detoxification, protection, and preservation of the environment.

## List of Technical Nomenclature

| | |
|---|---|
| Ah | humus-enriched A horizon |
| Bh | humus-enriched B horizon |
| $^{13}C$ | carbon-13 |
| C=O | carbonyl group |
| COOH | carboxyl group |
| $CO_2$ | carbon dioxide |
| CP-MAS $^{13}C$ NMR | cross-polarization magic angle spinning carbon-13 nuclear magnetic resonance spectroscopy |
| CP Py-GC/MS | Curie-point pyrolysis–gas chromatography/mass spectrometry |
| $cmol\,kg^{-1}$ | centimoles per kilogram |
| $E_4/E_6$ | ratio of absorbances at 465 nm and 665 nm |
| ESR | electron-spin resonance spectroscopy |
| FA | fulvic acid |
| FTIR | Fourier-transform infrared spectrophotometry |
| $g\,kg^{-1}$ | grams per kilogram |
| HA | humic acid |
| N | nitrogen |
| NaOH | sodium hydroxide |
| $Na_4P_2O_7$ | sodium pyrophosphate |
| $OCH_3$ | methoxyl group |
| OH | hydroxyl group |
| OM | organic matter |
| P | phosphorus |
| Py-FIMS | pyrolysis-field ionization mass spectrometry |
| ppm | parts per million |
| S | sulfur |
| SOM | soil organic matter |
| TII | total ion intensity |
| Zn | zinc |

*See also:* **Carbon Cycle in Soils:** Dynamics and Management; **Carbon Emissions and Sequestration**; **Cover Crops**; **Crop Rotations**; **Manure Management**; **Mulches**; **Organic Farming**

## Further Reading

Chen Y (1998) Electron microscopy of soil structure and soil components. In: Huang PM, Senesi N, and Buffle J (eds) *Structure and Surface Reactions of Soil Particles*, pp. 155–182. Chichester, UK: John Wiley.

Schnitzer M (1995) Organic–inorganic interactions in soils and their effect on soil quality. In: Huang PM, Berthelin J, Bollag J-M, McGill WB, and Page AL (eds) *Environmental Impacts of Soil Component Interactions*, pp. 3–19. Boca Raton, FL: Lewis Publishers.

Schnitzer M (2000) A lifetime perspective on the chemistry of soil organic matter. *Advances in Agronomy* 68: 1–58.

Schnitzer M (2001) The *in situ* analysis of organic matter in soils. *Canadian Journal of Soil Science* 81: 249–254.

Schnitzer M and Schulten H R (1995) Analysis of organic matter in soil extracts and whole soils by pyrolysis-mass spectrometry. *Advances in Agronomy* 55: 168–217.

Schulten H-R and Schnitzer M (1997) Chemical model structures for soil organic matter and whole soils. *Soil Science* 162: 115–130.

Schulten H-R and Schnitzer M (1998) The chemistry of soil nitrogen: a review. *Biology and Fertility of Soils* 26: 1–15.

Schulten H-R, Leinweber P, and Schnitzer M (1998) Analytical pyrolysis and computer modelling of humic and soil particles. In: Huang PM, Senesi N, and Buffle J (eds) *Structure and Surface Reactions of Soil Particles*, pp. 281–324. Chichester, UK: John Wiley.

Stevenson FJ (1994) *Humus Chemistry*. New York: John Wiley.

Wilson MA (1987) *NMR Techniques and Applications in Geochemistry and Soil Chemistry*. Oxford, UK: Pergamon Press.

# Genesis and Formation

**K M Haider**, University Halle, Halle, Germany

**G Guggenberger**, University of Bayreuth, Bayreuth, Germany

## Introduction

The primary sources of soil organic matter (SOM: all organic C-containing compounds in the soil) are dead plant materials in the form of leaves, straw, twigs, roots, and other plant litter materials. The global biomass of higher plants amounts to approximately 500–700 Gt C (1 Gt = $10^9$ t) and contributes approximately 100 Gt C in litter annually in the soil, either on the surface or deposited belowground as rhizodeposition. About the same amount of carbon is annually released from soil as $CO_2$. The dead plant material consists of lignocelluloses with an average composition of 15–60% cellulose, 10–30% hemicelluloses, 5–30% lignin, and 2–15% protein. Minor components are phenols, sugars, amino acids, and peptides, as well as numerous secondary metabolites. Most of the compounds are used as nutrient and energy sources for microbial growth.

SOM is a natural product resulting from microbial activity in the inorganic and/or organic soil environment. The amount and accumulation of SOM are controlled by the composition and amounts of the plant residues, by climatic conditions, and soil texture. Other important factors are microbial activity, soil redox conditions, and other soil chemical and physical properties.

Soils may contain several tons of SOM per hectare. Most of it can be only slowly degraded and metabolized by soil organisms as a consequence of both its physical and chemical stabilization. Plant debris, with structural relationship to its origin, may become resistant to microbial degradation by interaction with minerals, whereas its completely humified products are produced by random condensation of refractory plant and microbial products. Their diversity and lack of regular polymeric structures do not favor efficient enzymatic degradation and energy production. Therefore, chemical composition and structure of the humified residues and their physical protection by spatial distribution are equally important for accumulation of soil carbon, which amounts to approximately 1400–1600 Gt C on a global scale.

Combining the concepts of chemical stability and physical protection, it is assumed that there are two categories of natural organic matter: labile (readily degrading) and resistant (slowly degrading). Furthermore, there are three categories of microenvironments: free unprotected particles, particles occluded or entrapped in an inorganic or organic microenvironment, and particles attached or adsorbed to surfaces by physical or chemical mechanisms. This gives rise to six categories of SOM:

1. Free particulate, intrinsically labile;
2. Free particulate, intrinsically resistant;
3. Occluded particulate, potentially labile but currently protected as a result of inaccessibility;
4. Occluded particulate, intrinsically resistant and further protected by its inaccessibility;
5. Adsorbed, potentially labile, but currently protected by adsorption;
6. Adsorbed, intrinsically resistant and further protected by adsorption.

## Soil Fabric and its Impact on SOM Stabilization

### Transformation and Stabilization Processes of Free and Occluded Particulate Organic Matter

The location of organic matter within the soil structural units has been demonstrated to control SOM dynamics by physical or chemical mechanisms. Crushing of macroaggregates (greater than 250 $\mu$m) provides plant-derived material from roots or leaves colonized by microorganisms, which decompose

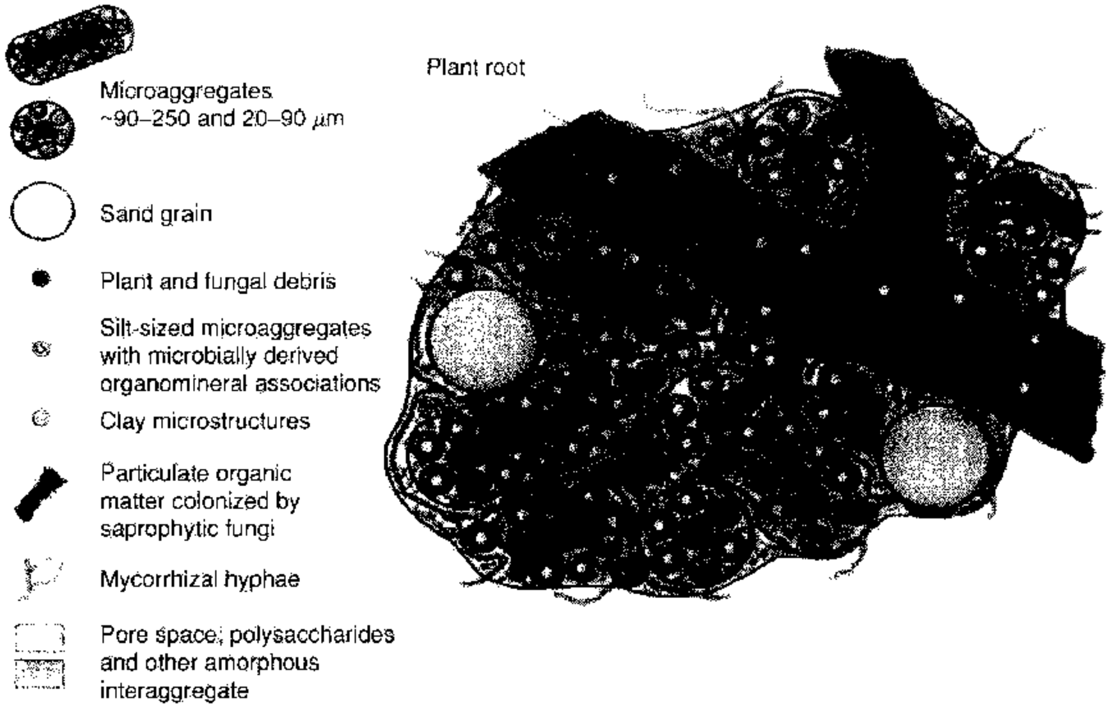


**Figure 1** Mineral and organic components within a macroaggregate. (Adapted from Jastrow JD and Miller RM (1998) Soil aggregate stabilization and carbon sequestration: feedbacks through organomineral associations. In: Lal R, Kimble JM, Follett RF, and Stewart BA (eds) *Soil Processes and the Carbon Cycle*, pp. 207–223. Boca Raton, FL: CRC Press.)

them and form humified material, microaggregates (less than 250 μm), and sand grains (Figure 1).

Microscopic and electron-microscopic evaluation gives direct access to the origin and the degree of decomposition of SOM within the different fractions. Mostly large, undecomposed root and plant fragments exist in the 500–2000-μm aggregate fractions; whereas, in the 10–100-μm aggregate fractions, more decomposed materials have been observed.

Proportions of free particulate organic matter vary widely and are affected by land use, vegetation and soil type, climate, and other factors. Its accumulation is favored in cold and dry climates and in continuously vegetated soils with a large input of plant litter, e.g., as in grass or forest vegetation areas. There, free particulate SOM can account for 15–40% of the total SOM in surface horizons, whereas in tilled soils it contributes less than 10% of the total SOM and reflects differences in cropping sequences and tillage. Free particulate SOM also includes organic particles which adhere loosely to surfaces of organomineral complexes. This type of particulate SOM can be recovered by minimal dispersion (e.g., density fractionation) of soil samples in which aggregates remain intact. It consists either of partly decomposed litter particles from recently deposited crop residues or of older, uncomplexed SOM previously occluded in aggregates but released by their destruction. These particles tend to be readily depleted when soils previously under a permanent native vegetation are brought into cultivation.

Plant-derived materials in soil are rapidly colonized by microorganisms. Microorganims and their mucilages are attached to clay particles and form microoaggregates which slow their biodegradation. With time the materials in macroaggregates become more fragile and mainly humified material is adsorbed on mineral surfaces in microaggregates, where it is hardly available to microorganisms and stabilized for a long time (Figure 2).

Many studies on different soil types have shown that macroaggregates contain younger C than microaggregates. Examples from investigations in Oxisols, Alfisols, and Inceptisols, where C3 vegetation was replaced by C4 vegetation, are shown in **Table 1**. The photosynthetic switch and the corresponding differences in $\delta\ ^{13}C$-values from C4 to C3 vegetation are frequently used to calculate average turnover times of aggregate-associated organic matter. The average turnover time for old C4-derived C is approximately 400 years for microaggregates compared with an average turnover time of approximately 140 years for macroaggregates.

## Stabilization of SOM by Sequestration in Microporous Structures or on Hydrophobic Surfaces

Physical factors confer biological stability to organic matter in soils through the constraints they place on the opportunities for reactions between substrates and enzymes or decomposer organisms. Microbial

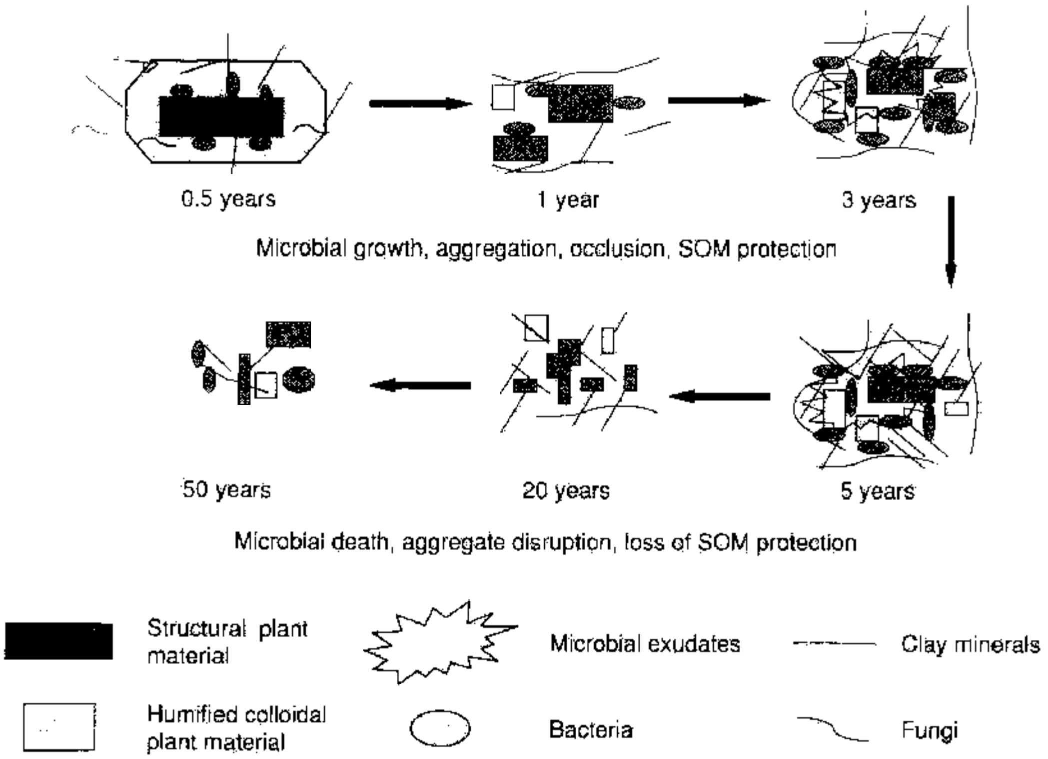


**Figure 2** Schematic presentation of the fate of soil organic matter (SOM) and soil aggregates. (Source: Balesdent J, Chenu C, and Balabane M (2000) Relationship of soil organic matter dynamics to physical protection and tillage. *Soil and Tillage Research* 53: 215–230.)

**Table 1** Percentage of young C (C4-derived C) in aggregates

| *Soil type (soil taxonomy)* | *Site* | *Experiment* | *Aggregate size (μm)* | *C4-C (%)* |
|---|---|---|---|---|
| Oxisol[a] | Australia | C4 grass for 35 years | 200–2000 | 49 |
| | | on former C3 rainforest | <200 | 41 |
| | | C4 grass for 83 years | 200–2000 | 75 |
| | | on former C3 rainforest | <200 | 67 |
| Alfisol[b] | France | Corn (C4) for 6 years | 200–6000 | 15 |
| | | on former C3 arable crops | <200 | 6 |
| | | Corn (C4) for 23 years | 200–6000 | 46 |
| | | on former C3 arable crops | <200 | 38 |
| Inceptisol[b] | Canada (Quebec) | Corn (C4) for 15 years | 2000–6000 | 21 |
| | | on former (C3) hay system | 1000–2000 | 18 |
| | | | 500–1000 | 11 |
| | | | 200–500 | 8 |
| | | | 50–200 | 3 |
| | | | <50 | 1 |

[a] Aggregates obtained by hand-grinding.

[b] Water-stable aggregates of slaked soil.

Adapted from Guggenberger G and Haider K (2002) Effect of mineral colloids on biochemical cycling of C, N, P and S in soil. In: Huang PM, Bollag J-M, and Senesi N (eds) *Interactions Between Soil Particles and Microorganisms and Their Impact on the Terrestrial Environment*. IUPAC Series on Analytical and Physical Chemistry of Environmental Systems, vol. 8, pp. 267–322. Chichester: John Wiley, with permission.

access to substrates when located in small pores or sorbed on solid surfaces may be limited, because they are inaccessible to microorganisms and their enzymes, and are thereby protected from rapid decomposition. Carbohydrates, lignin-degradation products, and nitrogenous substances can interact with soil colloids by sorption, polymerization, or entrapment in voids of inorganic or organic soil constituents.

Xenobiotics such as pesticides or polyaromatic hydrocarbons can be sorbed in the microporous structure of SOM or minerals. Similarly to xenobiotics, also natural SOM can be sorbed in micropores. Since such sorption sites are limited, the

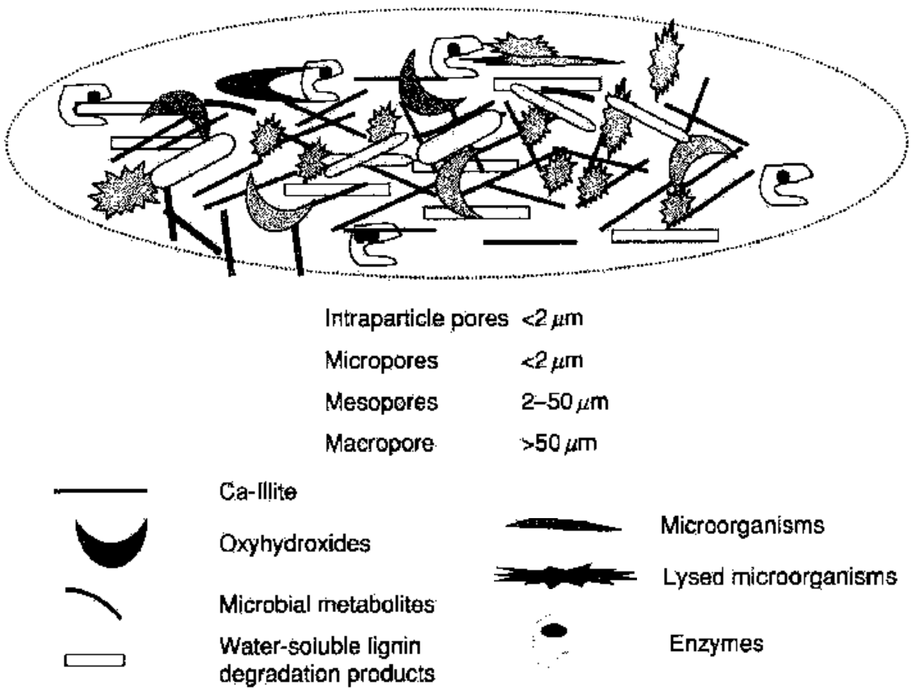


**Figure 3** Conceptual model of stabilization of soil organic matter within nanopores.

mineralization of phenols or phenolic acids, of amino acids, and of glucose or other carbohydrates can be more strongly retarded at low concentrations (10 ppm or less) than at higher concentrations. It can be presumed that the observed continuous decline in bioavailability (and extractability) is associated with a slow diffusion of the molecules to more remote sites. Intraparticle and intraorganic diffusion into nanopores is believed to account for these slow phases of sorption and desorption of nonionic molecules. Dissolved organic matter (DOM) can diffuse into nanopores of primary particles or sorb on active surfaces. Within pores less than 8 nm, DOM is inaccessible to exoenzymes. A conceptual model of the stabilization of sorbed SOM within nanopores is shown in **Figure 3**.

### Chemical Information About Organic Matter in Aggregates and Organomineral Complexes

Microscopic evidence suggests that SOM in the coarse sand fraction is mainly composed of little-decomposed plant residues, whereas that associated with fine sand consists of quite well-altered organic debris and fine particles of roots. This is corroborated by chemical analyses. Significantly lower C-to-N (C/N) ratios in fine fractions than in the coarser ones indicate a more pronounced microbial alteration in the fine-size separates (**Figure 4a**). This is also obvious from the pattern of monosaccharides released upon acid hydrolysis in the fractions. Individual sugar monomers released from SOM upon acid hydrolysis culminate in sand and clay separates and have their minimum in the silt-size fraction. There is a shift from primarily plant-derived monosaccharides in the coarse fractions to a dominating microbial contribution in the fine fractions (**Figure 4b**). Ratios of (galactose + mannose) to (arabinose + xylose) and of (rhamnose + fucose) to (arabinose + xylose) increase from sand to silt to clay, indicating primarily plant-derived polysaccharides in the sand-size separates and larger fractions of microbial polysaccharides in the clay. Amino acids, in particular diaminopimelic acid, which is confined to the peptidoglycans of cell wall of the prokaryotes, are recovered in elevated yields from clay. Because most of the bacterial biomass has been found in the vicinity of clay, high concentrations of bacterial-derived compounds (diaminopimelic acid, muramic acid) in the clay-size separates may be caused by a high production of microbial-derived substances.

As revealed by CuO-oxidation products, particle-size fractions of soils differ in the contribution of lignin to total organic carbon, but also in the degree of lignin alteration. Highest lignin contents obtained by the CuO-oxidation method are measured in the coarse fractions and decrease continuously with decreasing particle size (**Figure 4c**). A higher degree of oxidative alteration of lignin in finer-size separates compared with coarser fractions suggests that lignin in finer fractions had undergone more intensive microbial degradation.

Cross-polarization/magic angle spinning (CP/MAS) $^{13}$C nuclear magnetic resonance (NMR) spectroscopy

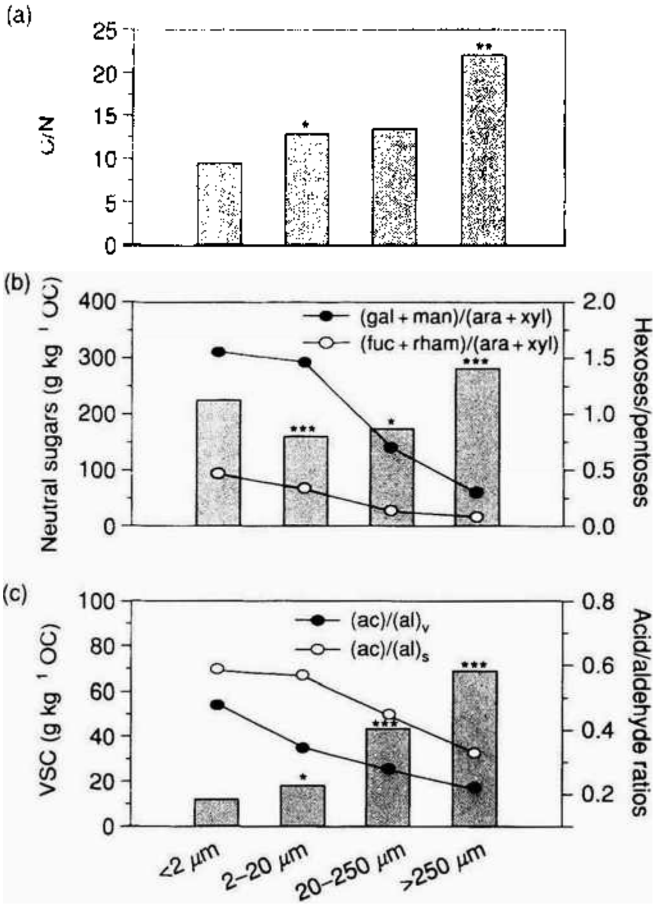


**Figure 4** Carbon-to-nitrogen ratio (C/N) in particle-size fractions of native topsoils (0–10 cm) (a); the concentration and the pattern of monosaccharides released upon acid hydrolysis (the pentoses arabinose (ara) and xylose (xyl) are considered primarily plant-derived, while the desoxysugars rhamnose (rham) and fucose (fuc) are considered synthesized mostly by microorganisms, the hexoses mannose (man) and galactose (gal) occur in both types of organisms) (b); and the concentrations of lignin-related phenols and phenolic acid/aldehyde ratios (filled circles, acid/aldehyde ratio vanillyl; open circles, acid/aldehyde ratio syringyl) after CuO oxidation of SOM in different particle sizes in native prairie soil (c). OC, organic carbon. *, **, *** Significant at the 0.05, 0.01 and 0.001 probability. (Adapted from Amelung W, Flach KW, and Zech W (1999) Lignin in particle-size fractions of native grassland as influenced by climate. *Soil Science Society of America Journal* 63: 1222–1228.)

is, at present, probably the most favored tool for detecting individual forms of C. Although this method is problematic with respect to quantification, it provides important insights into the chemical structure of SOM associated with different primary-particle fractions (Figure 5).

In general, O-alkyl resonances, mainly derived from carbohydrates or from ether bonds in degraded lignin predominate across the entire textural range, but they are minimized in the silt fraction. The strongest O-alkyl signals are always observed in the sand fraction, and spectra of the organic matter associated with sand strongly resemble those of fresh plant litter. Aromatic C, and in particular the aromatic C with other C substituents, is maximal in silt but is lowest in the clay fraction(s). Alkyl C steadily increases with decreasing particle size and represents the major or the second major C species in the clay(s). Long-chain aliphatic material derived from natural waxes is a major component of this alkyl C. Furthermore, microorganisms also synthesize long-chain aliphatic compounds.

Using the natural $^{13}$C-labeling technique where there is a shift in the photosynthetic pathway from C3-vegetation to C4-vegetation or vice versa (see also Table 1), it was shown that particle-size fractions represent SOM pools with different turnover times. Organic compounds in sand-size fractions are turned over rather rapidly within several years, whereas organic matter associated with silt and clay particles show markedly higher turnover times, thus being involved in mid- and long-term dynamics.

## Stabilization of SOM by Humification

As well as stabilization by spatial separation of organic matter particles with structural relationships to plant tissue material at different stages of microbial decomposition, the chemical stabilization of the completely humified products of the particles together with microbial compounds makes an important contribution to SOM. Accumulation of refractory humic compounds can be caused by preferential sorption of refractory fractions of organic matter in the dissolved stage (DOM). It results from leaching of dead plant residues, particularly those of the litter layer, with water or by solubilization of already humified compounds in the soil solution. Only small proportions of DOM can be identified in the form of organic acids, sugars, amino acids, and phenols. But most DOM is complex humified organic matter. The carbohydrate fraction of DOM contains a higher proportion of hexose and deoxysugars, which are typical for a microbial origin.

Adsorption of DOM on mineral phases results in a drastic decrease in its microbial degradability. Laboratory studies show that soil samples can adsorb DOM rapidly and there is evidence of a fast and strong adsorption of DOM to Al and Fe oxides and hydroxides. Sorption of the hydrophobic DOM fraction is strongly favored over the binding of the hydrophilic fractions in soils and on hydrous oxides. Furthermore, $^{13}$C NMR spectroscopy shows that the sorption of DOM to soil material leads to a preferential depletion of carbonyl and aromatic C, whereas alkyl C accumulates in the soluble portion. Also a preferential sorption of N compounds to oxides and clay minerals has been reported. Especially in the

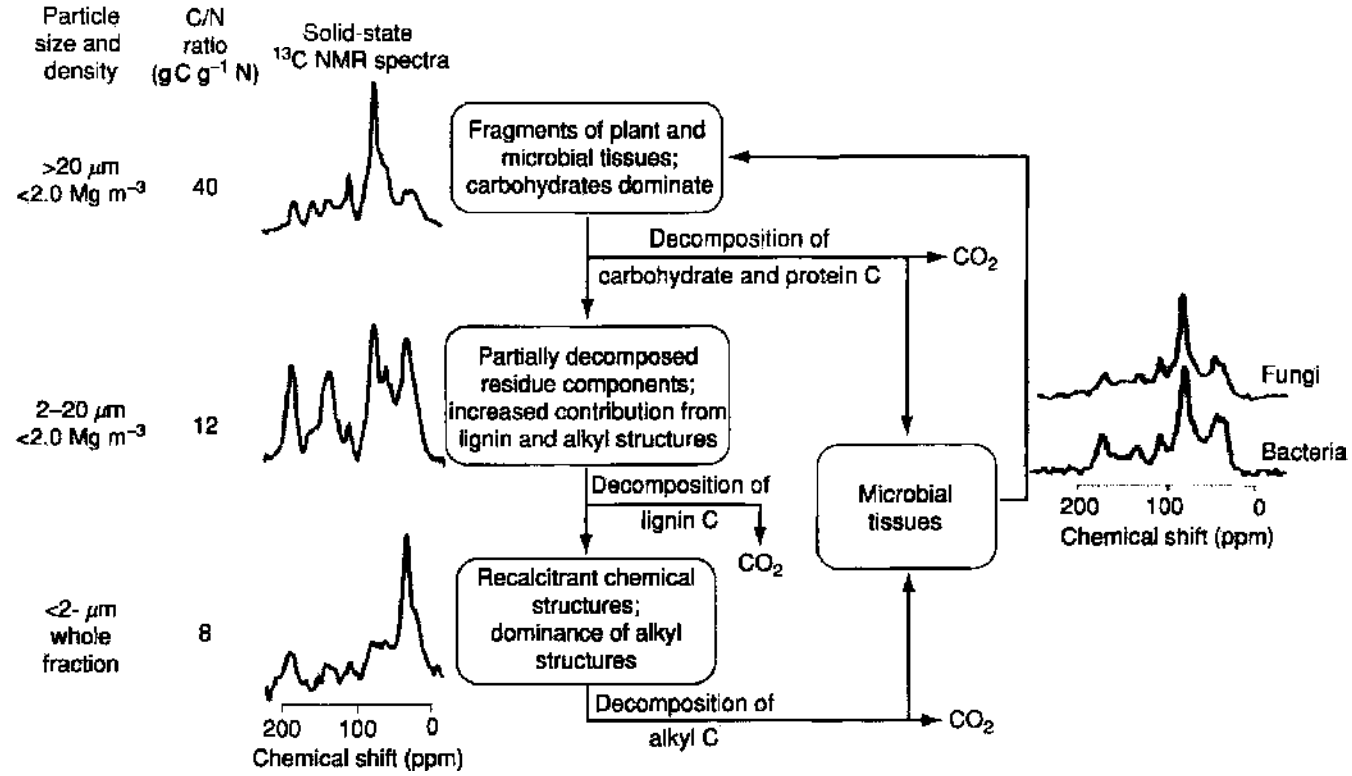


**Figure 5** Cross-polarization/magic angle spinning $^{13}$C-NMR spectra of the organic matter associated with different particle-size fractions. (Adapted from Baldock JA and Skjemstad JO (2000) Role of the soil matrix and minerals in protecting natural materials against biological attack. *Organic Geochemistry* 31: 697–710, with permission.)

deeper layers of the forest floor and in the A-horizon, low rates of DOM decomposition are found, which qualifies DOM as a source for the formation of stable humic substances. Sorption of DOM can create hydrophobic surfaces which limit microbial access by the reduction of wettability.

### Biochemical Stabilization Mechanisms of Main Components from Plant Residues in SOM

**Polysaccharide-related materials** With respect to quantity, polysaccharides account for 5–25% or more of soil organic carbon. Although most plant polysaccharides are readily decomposed, several microbially synthesized polysaccharides are relatively resistant against decomposition. Based on model experiments, mechanisms for their stabilization have been proposed, which may be caused by a close association with mineral colloids, an interaction with metal ions through uronic acid and mannose units, or by their interaction with tannins. Polysaccharides that are stabilized in such ways may represent a stable fraction of SOM. A further protection against rapid degradation arises from the special role of polysaccharides as 'bridges' between mineral particles, being directly involved in the formation of stable aggregates. There are indications that even more stable products of microbial resynthesis are also involved in the stabilization of SOM. One substance of importance is 'glomalin,' a glycoprotein with N-linked oligosaccharides, which is produced by hyphae of arbuscular mycorrhizal fungi and which is abundant and stable in soil.

**Lignin-related materials** Lignin, which is the second most abundant component of plant residues in terrestrial ecosystems, is relatively recalcitrant by itself, and degradation rates of lignocellulosic materials are negatively correlated to their lignin content or to their lignin-to-N ratio. Lignin as an aromatic biomolecule is degraded at a much slower rate than cellulosic and noncellulosic polysaccharides, and proteins. This recalcitrance is caused by the presence of nonhydrolyzable C–O–C and C–C bonds between the phenylpropanoid units, and the structural heterogeneity of stable ether and C–C bonds, which can only be cleaved by oxidative mechanisms through ligninolytic enzymes. These enzymes are directly or indirectly involved in the generation of lignin radicals by removing electrons from phenolic units. The resulting unstable radicals subsequently undergo a variety of spontaneous rearrangement reactions. During these reactions, water and oxygen molecules, and thereby additional hydroxyl, carboxyl, and C=O functions are introduced, which increase the oxygen content in the remnant lignin. Another peculiarity of

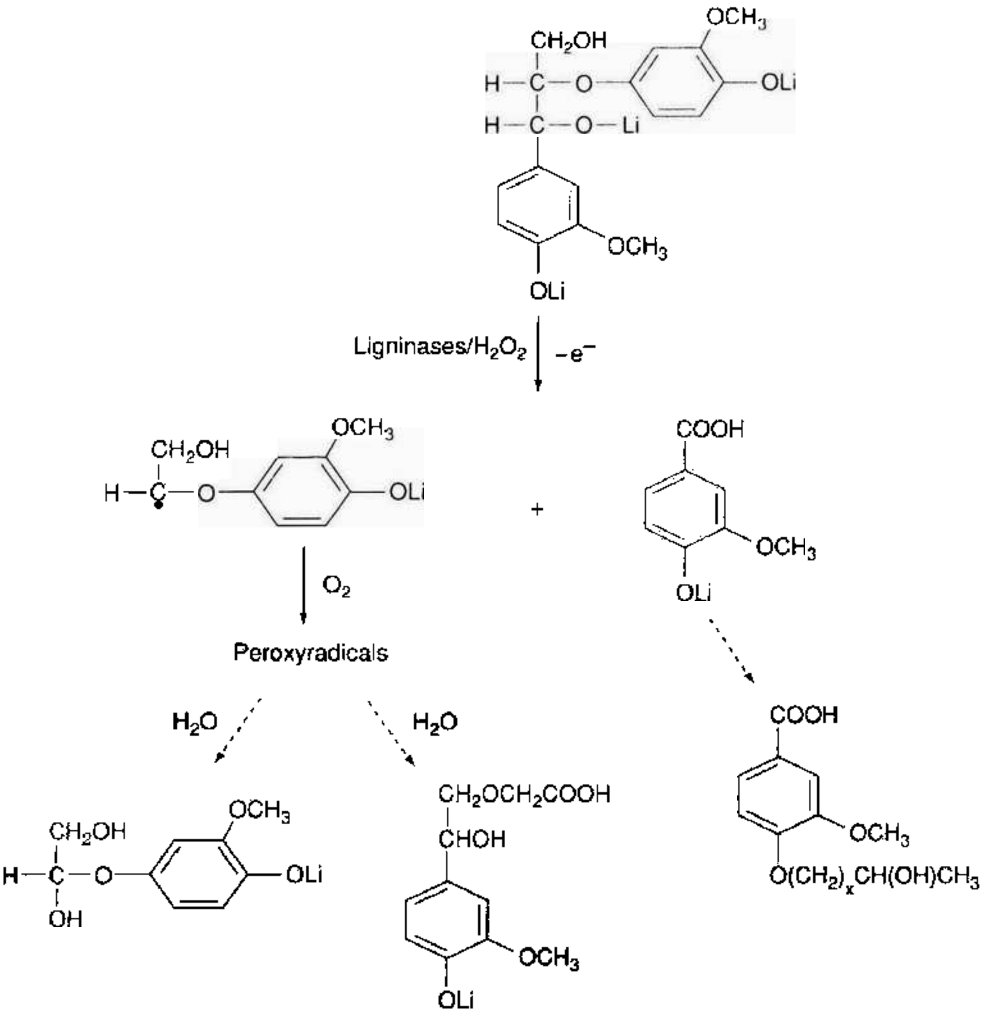


**Figure 6** Structural features of a microbially attacked lignin. Li, residues of lignin; •, radical (unpaired electron). (Adapted from Haider K (1992) Problems related to the humification processes in soils of temperate climates. In: Stotzky G and Bollag J-M (eds) *Soil Biochemistry*, vol. 7, pp. 55–94. New York: Marcel Dekker.)

lignin biodegradation, compared with other plant compounds, is the cometabolic character of the degradation process, by which it does not provide a source of energy or carbon for ligninolytic organisms. Thus, for an effective lignin degradation, a readily metabolizable cosubstrate such as carbohydrates is required.

Preservation of lignin constituents in humic substances has attracted attention for many years. Similarities and differences in structure and chemical reactivity between lignin and humic compounds have been reviewed recently in terms of existing humification hypotheses. $^{13}$C-NMR spectroscopy of microbially degraded lignin in soils, its oxidative CuO-splitting or its reductive HI-splitting products, pyrolysis in combination with previous chemolysis by tetramethylammonium hydroxide (TMAH) indicate that intact or only demethylated lignin does not appear to be stabilized in the long term in soils or in any of its fractions. However, the latter method reveals that phenolic structures characteristic for altered lignin structural units contribute significantly to the stable SOM pool in soils.

Based on investigations with ligninolytic organisms, enzymatic lignin degradation is a result of random peroxidative splitting of the lignin macromolecule in the $\alpha$-$\beta$ and other bonds of the side chain and in the aromatic rings, and by demethylation of the methoxyl groups (**Figure 6**). Thereby, the compact and hydrophobic structure of the lignin molecule becomes dispersed and, by introduction of additional carboxyl and hydroxyl groups, more hydrophilic. Additional hydroxyl and carboxyl groups in vicinal positions are apt to form complexes with metal ions in solution or on surfaces of clays or metal oxides.

### Impact and Function of the Soil Microflora in Relation to SOM Stability

In order to understand the role of microorganisms in the generation and turnover of refractory SOM, a better knowledge of population structure and the activity of the microbial communities in the soil habitat is necessary. Soils generally contain several tons of organic matter per hectare, but soil microorganisms can degrade most of this SOM pool only very slowly.

Models to describe the turnover of SOM assume the presence of easily and less-easily degradable pools in soil. These pools are, however, not clearly separated, but exist in a continuous sequence of overlapping residence times. Approaches to characterize SOM pools with different turnover rates distinguish between active, slow, and passive pools. The active or easily degradable pools with mean residence times of 0.1–4 years consist of free or only loosely adsorbed particulate organic matter (e.g., the light fraction, the microbial biomass). The slow-degrading SOM pools with mean residence times between 25 and 50 years consist of physically stabilized organic matter, i.e., occluded or sequestered in small aggregates. The passive or stable SOM pool contains the chemically stabilized humic materials, sorbed to fine silt and/or to clay, and has mean residence times of approximately 400–3000 years. On average, approximately 5–10% of the total C in SOM is involved in the active pools, and approximately 20–40% in the slow pools. To maintain a relatively constant nutrient level and release by mineralization, the pools have to be renewed continuously by fresh plant residues. The balance between decay and renewal processes controls nutrient availability and is sensitive to management practices. The slow pools are also affected by cultivation, as physical disturbance (caused, e.g., by ploughing) destroys macroaggregates and larger microaggregates and thus releases physically protected SOM. The passive pools contain the most stable SOM, which is protected by chemical and physical mechanisms. A strict distinction between physical and chemical effects of soil minerals on the stabilization of SOM can often not be made. The passive pool represents 50–70% of the total C in SOM and is seldom affected by management practices.

Whereas the compounds in the active and also some of the ingredients in the slow pools are used by microorganisms for metabolic and energy-producing processes, the great SOM portion of humic compounds in the passive pool is not readily available to microorganisms as a source of energy and products for microbial biosynthesis. On the other hand, small amendments of readily available substrates such as glucose or fructose, but even more in the form of cellulose, wheat straw, root exudates, or green manure, cause 2–5 times more C to be released from soil as $CO_2$ than was added with the substrate. This effect is termed the 'priming effect.' It arises either immediately or very shortly after the addition of available substrates and is caused by microbial activity, since similar effects have never been observed under sterile conditions. In the case of a 'positive priming effect,' the extra mineralization originates at least partly from the labile humus fractions or by an enhanced turnover of the native soil microbial biomass. In the case of a passive soil fraction, an increased induced enzyme production leading to a cometabolic decomposition has been found.

As for lignin, the completely humified products consist of randomly condensed intermediate compounds. In the case of a passive humus fraction, they represent refractory plant and microbial residues. Their diversity and lack of regular polymeric structures and the large number of enzymatic steps necessary to release $CO_2$ do not allow efficient enzymatic degradation and energy production. The slow decomposition of the passive SOM, however, becomes accelerated by small inputs of available substrates. Even when these inputs are relatively small compared with the large size of the passive fraction in most soils, they 'trigger' the synthesis of enzymes for the cometabolic degradation of a portion of the otherwise passive SOM fraction. This priming effect only lasts for a relatively short period after addition and disappears after metabolization of the added substrates. It can be hypothesized that this effect may limit the ultimate accumulation of passive SOM.

*See also:* **Carbon Cycle in Soils:** Dynamics and Management; **Carbon Emissions and Sequestration**; **Factors of Soil Formation:** Biota; **Organic Matter:** Principles and Processes; **Pedology:** Basic Principles

## Further Reading

Amelung W, Flach KW, and Zech W (1999) Lignin in particle-size fractions of native grassland as influenced by climate. *Soil Science Society of America Journal* 63: 1222–1228.

Baldock JA and Skjemstad JO (2000) Role of the soil matrix and minerals in protecting natural materials against biological attack. *Organic Geochemistry* 31: 697–710.

Balesdent J, Chenu C, and Balabane M (2000) Relationship of soil organic matter dynamics to physical protection and tillage. *Soil and Tillage Research* 53: 215–230.

Batjes N (1996) Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science* 47: 151–163.

Christensen BT (2001) Physical fractionation of soil and structural and functional complexity in organic matter turnover. *European Journal of Soil Science* 52: 345–353.

de Jonge H, de Jonge LW, and Mittelmeijer-Hazelegger MC (2000) The microporous structure of organic and mineral soil materials. *Soil Science* 165: 99–108.

Desjardins T, Andreux F, Volkoff B, and Cerri CC (1994) Organic carbon and $^{13}C$ contents in soils and soil size-fractions, and their changes due to deforestation and pasture installation in eastern Amazonia. *Geoderma* 61: 103–108.

Fontaine S, Mariotti A, and Abbadie L (2003) The priming effect of organic matter: a question of microbial competition? *Soil Biology and Biochemistry* 35: 837–843.

Guggenberger G and Haider K (2002) Effect of mineral colloids on biochemical cycling of C, N, P and S in soil. In: Huang PM, Bollag J-M, and Senesi N (eds) *Interactions Between Soil Particles and Microorganisms and Their Impact on the Terrestrial Environment*. IUPAC Series on Analytical and Physical Chemistry of Environmental Systems, vol. 8, pp. 267–322.

Haider K (1992) Problems related to the humification processes in soils of temperate climates. In: Stotzky G and Bollag J-M (eds) *Soil Biochemistry*, vol. 7, pp. 55–94. New York: Marcel Dekker.

Jastrow JD and Miller RM (1998) Soil aggregate stabilization and carbon sequestration: feedbacks through organomineral associations. In: Lal R, Kimble JM, Follett RF, and Stewart BA (eds) *Soil Processes and the Carbon Cycle*, pp. 207–223. Boca Raton, FL: CRC Press.

Kalbitz K, Solinger S, Park J-H, Michalzik K, and Matzner E (2000) Controls on the dynamics of dissolved organic matter in soils: a review. *Soil Science* 165: 277–304.

Kögel-Knabner I (1993) Biodegradation and humification processes in forest soils. In: Bollag J-M and Stotzky G (eds) *Soil Biochemistry*, vol. 8, pp. 101–137. New York: Marcel Dekker.

Ladd JN, Foster RC, Nannipieri P, and Oades JM (1995) Soil structure and biological activity. In: Stotzky G and Bollag J-M (eds) *Soil Biochemistry*, vol. 9, pp. 23–78. New York: Marcel Dekker.

Shevchenko SM and Bailey GW (1996) Life after death: lignin–humic relationship reexamined. *Critical Reviews of Environmental Science and Technology* 26: 95–153.

Zech W, Guggenberger G, Haumaier L *et al.* (1996) Organic matter dynamics in forest soils of temperate and tropical ecosystems. In: Piccolo A (ed.) *Humic Substances in Terrestrial Ecosystems*, pp. 101–170. Amsterdam, the Netherlands: Elsevier.

Zepp RG and Sonntag CH (eds) (1995) *The Role of Nonliving Organic Matter in the Earth's Carbon Cycle*. Chichester, UK: John Wiley.

# Interactions with Metals

**N Senesi and E Loffredo**, Università di Bari, Bari, Italy

## Introduction

Metal ions in soil are rarely found in the free state, but mostly occur in interactions with other soil organic, inorganic, and biological components, in the liquid and solid phases. The distribution and speciation of metal ions in soil, especially in trace concentrations, are extremely complex and are governed by a variety of reactions, among which complexation with soil organic matter (SOM) plays a key role.

Thus, understanding the molecular, mechanistic, and quantitative aspects of metal–SOM complexation is extremely important in the prediction of processes in which metals are involved in soil. These include: pedogenesis, mobility and transport, fixation and accumulation, chemical and biological reactivity, bioavailability, and toxicity to higher plants and microorganisms. The topic is also of considerable practical interest because of the continuous and increasing release of various metals in soil by numerous modern agricultural practices.

## Molecular and Mechanistic Aspects

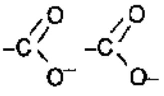
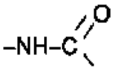
### Metal Reactivity, SOM Binding Sites, and Complexation Stoichiometry

Metal cations may be subdivided into three classes with respect to their reactivity with organic ligands: (1) 'hard' cations such as alkaline-earth metals $Ca^{2+}$ and $Mg^{2+}$ that are mainly involved in electrostatic interactions and may form rather weak, outer-sphere complexes only with hard oxygen ligands; (2) 'soft' cations such as $Cd^{2+}$, $Pb^{2+}$, and $Hg^{2+}$ that possess a strong affinity for and tend to form covalent bonds with intermediate (N) and soft (S) ligands; and (3) 'borderline' cations, including most transition metals such as $Fe^{3+}$, $Cu^{2+}$, $Zn^{2+}$, and $Mn^{2+}$ that have a character intermediate between hard and soft metals, and possess appreciable affinity for both hard and soft ligands.

The most important metal-binding functional groups present in SOM, classified according to their affinity for hard, borderline, and soft metals, are listed in Table 1. For hard cations the order of donor atom affinity is: $O > N > S$, whereas a reverse order is observed for soft metals. For bidentate ligand sites, affinity for a given hard metal decreases with the overall hardness of the donor atoms, in the order: $(O, O) > (O, N) > (N, N) > (N, S)$; whereas this order is reversed for soft metals. In general, the competitive reactions for a given ligand essentially involve hard and borderline metals for O sites, and borderline and soft metals for N and S sites, with competition between hard and soft metals being weak. However, the type, source, and concentration of SOM can affect metal-binding affinity, and the relative affinities are often dependent on the method used to measure metal bonding and pH.

Humic substances (HS), i.e., humic acids (HA) and fulvic acids (FA) are universally recognized to

**Table 1** Most important metal organic ligands of soil organic matter classified according to their preference for hard, borderline, and soft metals

| *Ligands preferred by hard cations (hard bases)* | *Ligands preferred by borderline cations* | *Ligands preferred by soft cations (soft bases)* |
|---|---|---|
|  (Ester)<br>(Carboxylate) | $-NH_2$, $=NH$, $\equiv N$<br>(Primary, secondary, tertiary amino groups) | $R^-$<br>(Alchil anion) |
| $-OH$ (Alcoholic and phenolic) | <br>(Amide) | $-SH$, $-S^-$<br>(Sulfhydryl, sulfide) |
| $C=O$, $-O-$<br>(Carbonyl) (Ether) | | $-S-S-$, $-S-$<br>(Disulfide) (Thioether) |

represent the most numerous (between 70 and 80%) and most metal-reactive SOM components. This is because they contain a large number of binding sites per molecule and thus behave as natural 'multiligand' systems.

Metal ion–HS interactions may yield mononuclear complexes, with the central group being either the HS macromolecule or the metal ion, and polynuclear complexes of various stoichiometries. Accumulated evidence shows that aromatic carboxyls and phenolic groups play a prominent role in the 1:1 binding of metal ions by HA and FA, by forming chelates that involve two COOH groups in a phthalic-type site (**Figure 1a**) and both COOH and phenolic groups in a salicylate-type site (**Figure 1b**). Nonaromatic carboxyl and hydroxyl sites, e.g., pyruvic and glycolic acid sites (**Figure 1c, d**), may also be involved in metal ion binding by HA and FA. The most stable complexes are believed to be formed with the more strongly acidic COOH groups, whereas the less stable complexes involve weakly acidic COOH and phenolic OH groups. Other possible combinations involve two phenolic OH, quinone, $NH_2$, and sulfhydryl groups, and conjugated ketonic structures (**Figure 1e**). Metal ions may also coordinate with ligands belonging to two (or more) HS molecules, forming 2:1 complexes (**Figure 1f, g**) and/or chelates (**Figure 1h, i**). Eventually, a chain structure (**Figure 1j**) may be produced that may result in the aggregation and precipitation as the chain grows at high metal-to-HS ratios.

Two main types of complexes may be formed between metal ions and HS, which are: (1) inner-sphere complexes that result in the formation of bonds with some covalent character between the ligand atom(s) and the metal ion, both completely or partially dehydrated; and (2) outer-sphere complexes that result in the electrostatic attraction between the ligand(s) and the metal ion that remains completely hydrated. For simplicity, **Figure 1a–j** shows formation of inner-sphere complexes, but they may represent outer-sphere HS complexes if the cation remains solvated (e.g., **Figure 1k**).

A number of molecular properties of HS and physical and chemical characteristics of the surrounding medium govern metal–HS binding processes. These include: polyfunctionality, polyelectrolyte character, hydrophilicity/hydrophobicity and extent of ionization of major acidic groups of HS, electronic and steric environment of the HS ligand site, and pH, ionic strength, and metal concentration of the medium.

## Experimental Evidence

Spectroscopic techniques such as infrared (IR), ultraviolet-visible (UV-VIS), fluorescence, electron-spin or electron paramagnetic resonance (ESR or EPR), X-ray absorption (XAS), Mössbauer, and nuclear magnetic resonance (NMR) spectroscopies have provided a number of important molecular and mechanistic insights of the chemical and physical nature of HS binding sites and metal–HS interactions.

**IR, UV-VIS, and fluorescence spectroscopies** Fourier transform IR (FT IR) and diffuse reflectance FT IR (DRIFT) spectroscopies, often in conjunction with selective group blocking techniques and/or with

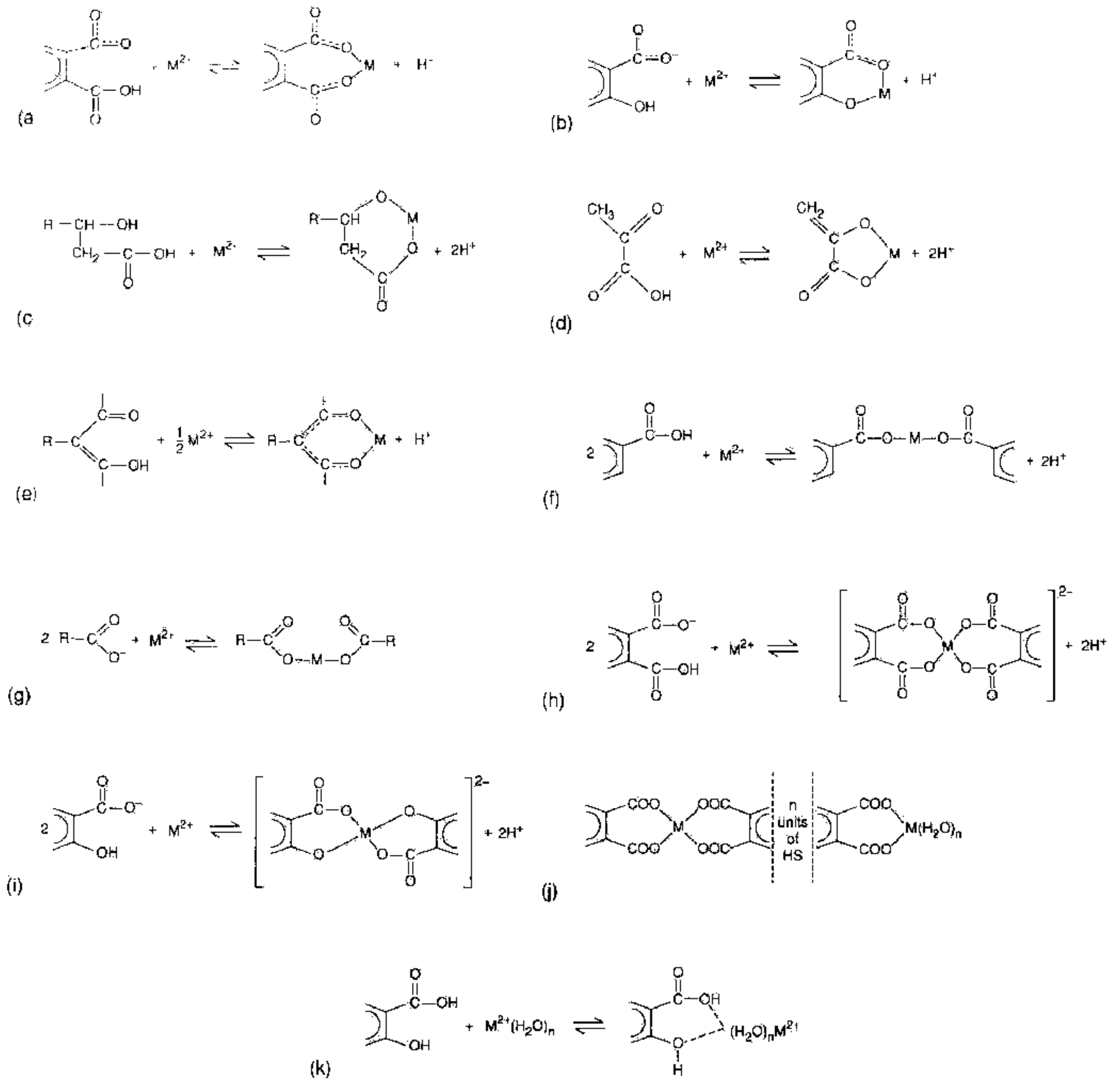


**Figure 1** The most common binding sites for metal ions in humic substances and stoichiometries of complexation: (a) a phthalic-type binding site; (b) a salicylate-type binding site; (c) a pyruvic acid binding site; (d) a glycolic binding site; (e) a conjugated ketonic binding site; (f) a 2:1 aromatic carboxyl complex; (g) a 2:1 aliphatic carboxyl complex; (h) a 2:1 phtalic-type chelate; (i) a 2:1 salicylate-type chelate; (j) a chain structure; (k) a salicylate-type outer-sphere complex.

chemical derivatization methods, have provided useful information about the nature and reactivity of HS functional groups that interact with metal ions.

The prominent role played by COOH groups of HA and FA in the complexation of several metal ions, including Cu$^{2+}$, Mn$^{2+}$, Zn$^{2+}$, Pb$^{2+}$, Co$^{2+}$, Ni$^{2+}$, Ca$^{2+}$, Mg$^{2+}$, Sr$^{2+}$, Fe$^{3+}$, and Al$^{3+}$, has been demonstrated by the strong reduction or disappearance of the typical IR absorption bands of COOH groups at approximately 1710 and 1200 cm$^{-1}$, and the simultaneous reinforcement or appearance of the typical IR bands of COO$^-$ groups near 1600 and 1380 cm$^{-1}$, upon reaction of HA and FA with the metal (**Figure 2**). The extent of COO$^-$ to COOH conversion varies in the decreasing order: Fe$^{3+}$ > Cu$^{2+}$ > Al$^{3+}$ > Ca$^{2+}$ > Mg$^{2+}$. The increase in COO$^-$ vibration frequency for a number of metal–HA and –FA complexes suggests that high-covalent bonds are preferentially formed at low levels of metal ion, whereas bonding becomes increasingly ionic as the system is enriched with the metal. The large separation measured between the frequencies at approximately 1600 and 1380 cm$^{-1}$ for the COO$^-$ group complexed by Cu$^{2+}$, Fe$^{3+}$, Co$^{2+}$, and Zn$^{2+}$, with respect to that of

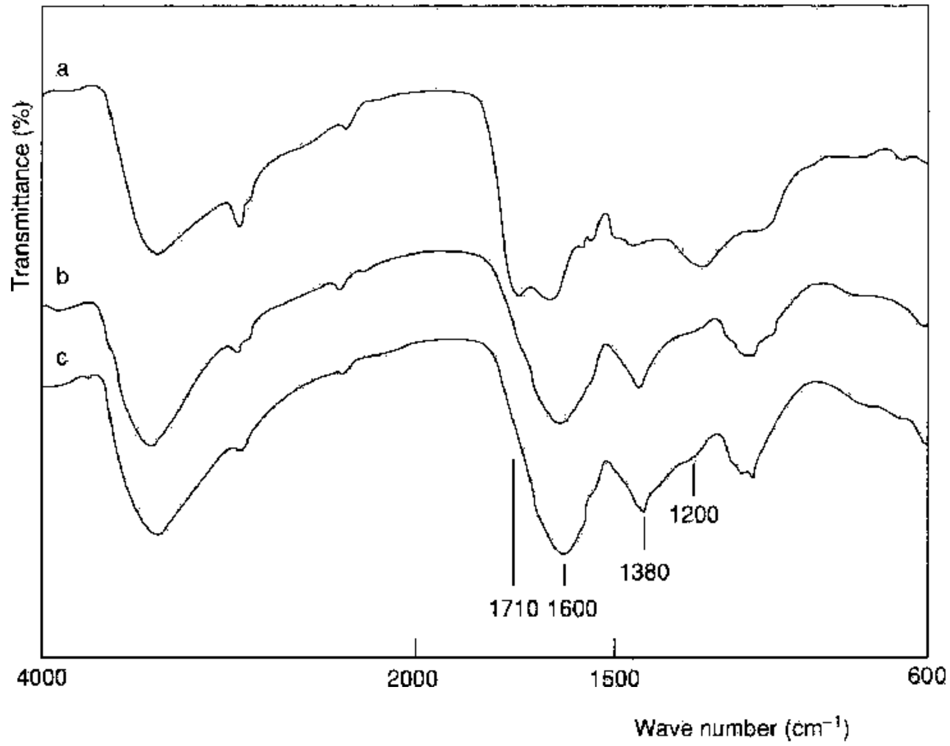


**Figure 2** Infrared spectra of a soil humic acid (HA) (a) and its complexes with $Cu^{2+}$ ($Cu^{2+}$:HA = 0.04) (b), and $Cu^{2+}$ and $Fe^{3+}$ (($Cu^{2+}$ + $Fe^{3+}$):HA = 0.08) (c). (Adapted from Senesi N, Sposito G, and Martin GP (1986) Copper (II) and iron (III) complexation by soil humic acids: an IR and ESR study. *Science Total Environment* 55: 351–362, with permission.)

the uncomplexed $COO^-$ ion, suggests the preferential formation of unidentate metal complexes.

The net IR shift of the OH band, from 3500–3400 cm$^{-1}$ in HA and FA to 3300–3200 cm$^{-1}$ in their complexes with $Zn^{2+}$, $Fe^{3+}$, $Cu^{2+}$, $Pb^{2+}$, $Co^{2+}$, and $Mn^{2+}$ is ascribed to the involvement of phenolic and/or alcoholic OH groups in metal binding. The extent of the OH stretching shift toward lower frequencies follows the order: Fe > Cu > Co > Mn. Modification of the typical band at approximately 1070 cm$^{-1}$ ascribed to OH groups of polysaccharide components of FA suggests that these structures are also involved in metal complexation. The shift to a lower frequency measured for the band at 1610 cm$^{-1}$ in $Cu^{2+}$–FA complexes is attributed to the involvement of conjugated ketones in $Cu^{2+}$ binding. IR evidence is also provided for metal binding sites involving amide N (and possibly amide C=O) and sulfonic groups by N- and S-rich HS. Finally, the very sharp IR bands observed in the region 1130–1080 cm$^{-1}$ and at between 890 and 697 cm$^{-1}$ for HA and FA complexes with $Cu^{2+}$, $Zn^{2+}$, $Al^{3+}$, or $Fe^{3+}$ are attributed to the metal–O vibration of bound hydroxylated and/or hydrated metal ions.

The stoichiometry of metal–HS complexes has been investigated by UV-VIS spectroscopy based on the Job method. For example, the ions $Cu^{2+}$, $Fe^{3+}$, and $Al^{3+}$ form 1:1 complexes with FA at pH 3; whereas, at pH 5, $Cu^{2+}$ and $Fe^{3+}$ form 2:1 molar complexes with the FA, and $Al^{3+}$–FA complex composition remains at 1:1.

The capacity of paramagnetic transition metal ions such as $Cu^{2+}$, $Fe^{3+}$, $Fe^{2+}$, $Mn^{2+}$, $Cr^{3+}$, and $VO^{2+}$ to quench efficiently the natural fluorescence of some HS ligands is attributed to their capacity to form strong, inner-sphere complexes with HS, involving strongly acidic carboxyl groups in bidentate salicylic- or phthalic-type chelation sites. In contrast, the low fluorescence-quenching ability exhibited by cations such as $Mn^{2+}$, $Co^{2+}$, $Ni^{2+}$, $Pb^{2+}$, and $Al^{3+}$ suggests the formation of weak, outer-sphere complexes with HS. The quenching effect generally increases with increasing pH (**Figure 3**). Fluorescence-quenching studies also suggest the involvement of aromatic amino acidic groups such as tyrosine and phenylamine moieties, and citrate and malonate moieties in the complexation of $Cu^{2+}$ by FA.

**ESR, XAS, Mössbauer, and NMR spectroscopies** The ESR technique applied to HA and FA complexes with paramagnetic transition metals of major importance in soil, including Fe, Cu, Mn, and V, has provided useful information about oxidation states of metals bound, symmetry and type of coordination sites in HS, binding mechanisms of metals to HS, and identity of ligand atoms and groups involved in metal complexing.

The representative ESR spectrum of a soil HA shown in **Figure 4a** exhibits a number of patterns

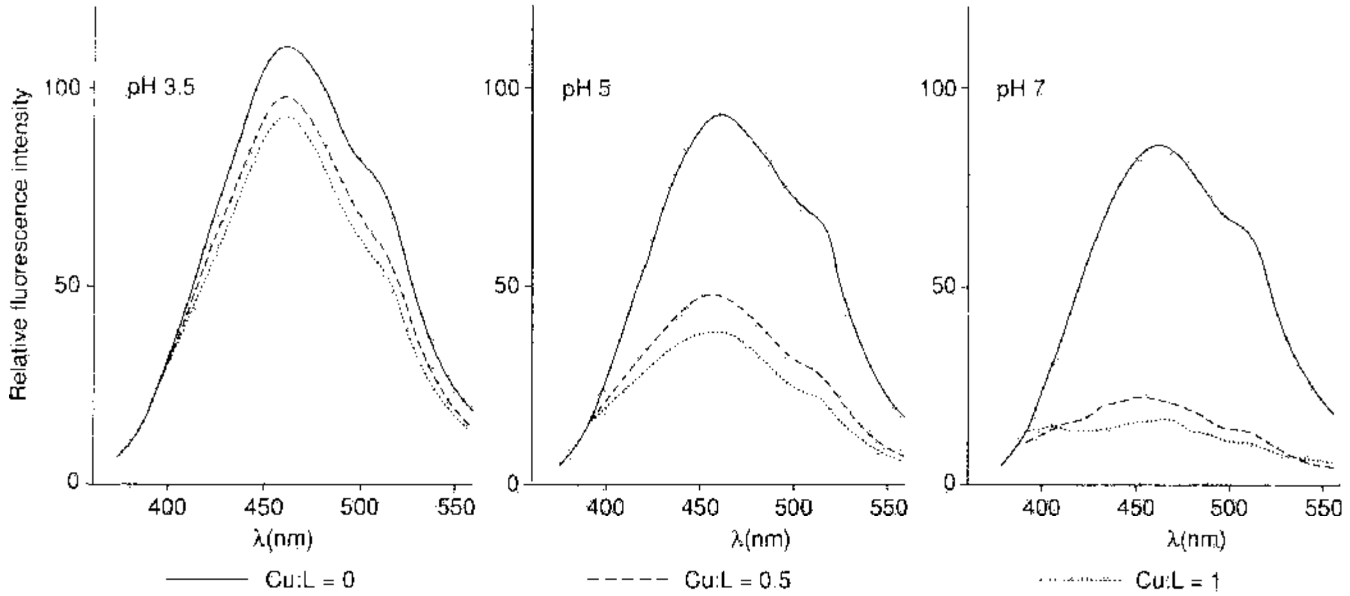


**Figure 3** Fluorescence emission spectra ($\lambda_{exc}$ = 355 nm) of a soil fulvic acid (FA) and its complexes with $Cu^{2+}$ obtained at three pHs and two different $Cu^{2+}$-to-FA ligand (L) ratios. (Reproduced with permission from Bartoli F, Hatira A, Andre JC, and Portal JM (1987) *Soil Biology Biochemistry* 19: 355–362.)

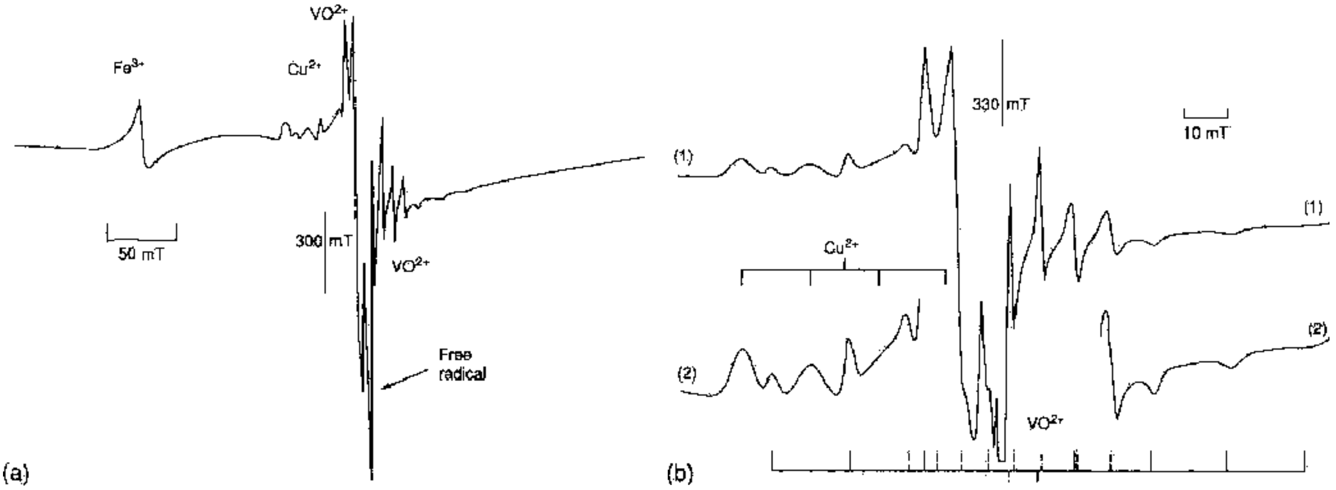


**Figure 4** Electron-spin resonance (ESR) spectra at 77 K of a loam soil HA recorded on scan ranges of 800 mT (a) and 200 mT (b) (1, high-field region; 2, higher gain). (Reproduced with permission from Senesi N (1990) Application of electron-spin resonance (ESR) in soil chemistry. In: Stewart BA (ed.) Propriétés fluorescentes et colloïdales d'une solution organique de podzol au cours du processus de complexation par le cuivre. *Advances in Soil Science*, vol. 14, pp. 77–130. New York: Springer-Verlag.)

that can be resolved, analyzed, and attributed to various metal–HS complexes. The asymmetrical, isotropic resonance line appearing at low magnetic field is consistent with high-spin $Fe^{3+}$ ions in tetrahedral or octahedral sites of rhombic symmetry bound to O functional groups, possibly carboxylic and/or phenolic groups of HA. This form of Fe exhibits considerable resistance to proton and metal exchange, and to chemical reduction, thus suggesting that $Fe^{3+}$ is strongly bound and protected in inner-sphere complexes in HS. The broad signal often exhibited by soil HS at higher field is preferentially ascribed to $Fe^{3+}$ ions in slightly distorted octahedral sites. Iron in such sites is easily reduced by chemical agents and easily extracted by complexing agents, thus suggesting a weak binding of $Fe^{3+}$ on external surfaces of HS.

The richly structured pattern exhibited at higher field (**Figure 4a**), which is better analyzed using the enlarged spectrum of the region (**Figure 4b**), can be resolved into two distinct, overlapping rigid-limit spectra of the 'axial' type. One comprises the typical anisotropic pattern ascribed to $Cu^{2+}$ ions held in inner-sphere complexes with ligands arranged in a square planar (distorted octahedral) coordination site (tetragonal symmetry), involving either only O functional groups (carboxyls, phenolic hydroxyls, carbonyls, and often water molecules) of HS, or both O and N ligands, or even only N atoms (i.e., a tetraporphyrin site). The other pattern is consistent with a $VO^{2+}$ ion rigidly bound in HS as an inner-sphere complex in a square planar coordination site to O ligands, mostly phenolate, or possibly water molecules and, more rarely, N ligands.

ESR spectral parameters also provide evidence of a high covalent bond for $Cu^{2+}$ and $VO^{2+}$ complexed in HS, which renders $Cu^{2+}$ and $VO^{2+}$ ions difficult to remove from these sites, even with acid leaching. Covalent bonding is favored by low metal loadings in HS, where complexation to N-containing groups is preferred to O-containing ligands, whereas at high loadings the metal is bound largely to O-containing ligands and has a higher degree of mobility.

The ESR spectra of some soil HA and FA feature a well-resolved isotropic pattern consistent with high-spin, hexahydrated $Mn^{2+}$ bound by electrostatic forces as outer-sphere complexes in a distorted octahedral environment to six O atoms of negatively charged carboxylate and phenolate groups. However, at $pH > 8$, or temperature $T > 50°C$, $Mn^{2+}$ can enter inner-sphere, multiligand complexation sites. Thus, the type and stability of $Mn^{2+}$–HS complexes and, in turn, their ease of exchangeability and bioavailability in natural systems are strongly dependent on pH and temperature.

ESR evidence has been provided that HS possess a high residual binding capacity toward added $Cu^{2+}$, $Fe^{3+}$, $VO^{2+}$, and $Mn^{2+}$ ions that can form complexes of various stability against various physical and chemical treatments. Most of the added metal is bound to surface sites in relatively labile and exchangeable forms.

In general, ESR analysis shows that high pH values that generate a greater availability of negatively charged O ligands favor inner-sphere complexation for metals, whereas, at lower pH, metals are preferentially retained as hydrated ions. Inner-sphere coordination is also preferred when competing water ligands are removed by dehydration, thus forcing the metal to enter into direct bonding with HS ligands. ESR data confirm that the more electronegative the metal ion, the stronger the metal bound to HS, the higher the degree of bond covalency. The ESR-related, electron-nuclear double resonance (ENDOR) and electron spin-echo envelope modulation (ESEEM) spectroscopies have the potential to overcome some limitations and yield more information than the classic ESR approach about the chemical environment of paramagnetic metal ions.

The XAS techniques, X-ray absorption near-edge structure (XANES) and extended X-ray-absorption fine-structure (EXAFS) spectroscopies, have yielded additional details about the coordination of the central ion to HS ligands and values of bond lengths and distances of metal-ligand also for nonparamagnetic metal ions that cannot be revealed by ESR. Analysis of XANES spectra, radial structure function (RSF) plots derived from EXAFS spectra, and corresponding simulations for $Cu^{2+}$ bound to HS at various pHs (Figure 5) indicate the presence of a tetragonally distorted octahedral binding site for $Cu^{2+}$ in HS, involving four equatorial O (or N) atoms at a mean distance of 1.92–1.97 Å and two axial O atoms at a mean

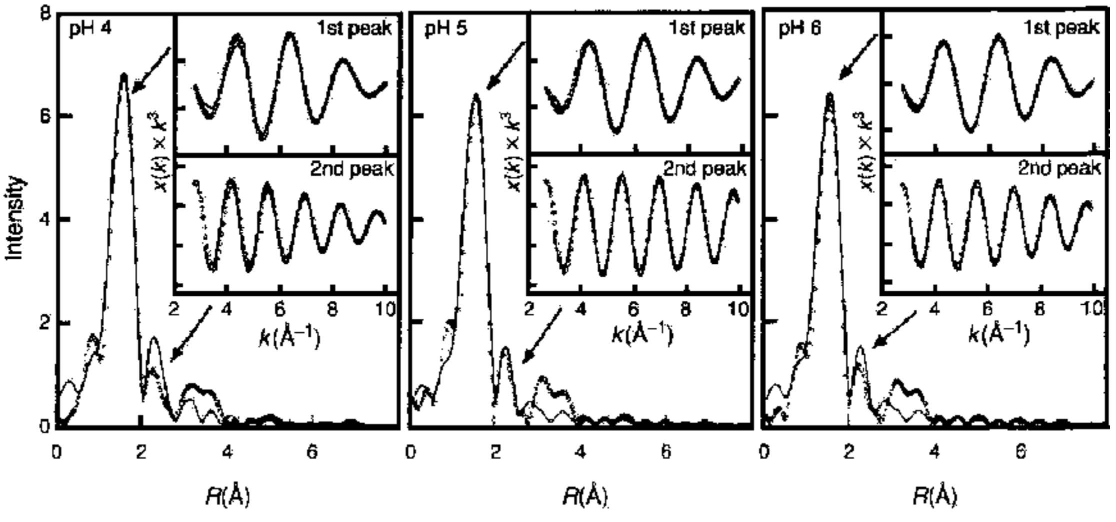


**Figure 5** Experimental radial structure function for $Cu^{2+}$–humic substance complexes at pH 4, 5, and 6 (dotted line) and simulations (solid line) for an adjusted model of the coordination site derived from bond network analysis. The inset shows plots of experimental (dotted line) and fitted (solid line) inverse Fourier-transformed scattering curves for the first atomic shell (Cu–O) and second atomic shell (Cu–C). *R*, distance (in Å) of the ligand atom from $Cu^{2+}$. (Reproduced with permission from Xia K, Bleam WF, and Helmke PA (1997) Studies of the nature of $Cu^{2+}$ and $Pb^{2+}$ binding sites in soil humic substances using X-ray absorption spectroscopy. *Geochimica Cosmochimica Acta* 61: 2211–2221.)

distance of 2.02 Å in the first coordination shell, and four C atoms at a mean distance of 3.13 Å in the second shell. Similar XANES and EXAFS studies of $Pb^{2+}$–HS complexes indicate that $Pb^{2+}$ is bound to four O atoms in the first shell at mean distances, decreasing from 2.46 to 2.32 Å, with increasing pH from 4 to 6, and two C atoms at a mean distance of 3.26 Å in the second shell. These results suggest inner-sphere bonding of $Pb^{2+}$ to two O-containing, possibly salicylate- and catechol-type functional groups of HS. Complexes of $Ni^{2+}$, $Co^{2+}$, and $Zn^{2+}$ with soil HS at pH 4 are interpreted as inner-sphere octahedral coordination, with six O atoms in the first shell for $Ni^{2+}$ and $Co^{2+}$, and with four O and two S atoms for $Zn^{2+}$. The bond distances for Ni–O and Co–O are 2.10 Å and 2.04 Å, respectively, and for Zn–O and Zn–S are 2.13 Å and 2.33 Å, respectively. There is one C atom in the second shell for Co, at a distance of 2.87 Å, and there are two for Ni and Zn, at a distance of 2.94 Å and 3.29 Å, respectively. At very high $Zn^{2+}$ concentration, however, most of the Zn (80–90%) is found to be bound as an outer-sphere complex. XANES and EXAFS spectra of $Fe^{3+}$ and $Mn^{2+}$ complexes with HA suggest that both ions are held in HA by six nearest-neighbor O or N atoms at a distance of 2.01 ± 0.05 Å and 2.20 ± 0.04 Å, respectively, and that $Fe^{3+}$ occupies more than one binding site. Detailed EXAFS studies of $Hg^{2+}$ complexation by soil HA suggest a two-coordinate binding environment for $Hg^{2+}$, with one O atom and one S atom in the first coordination shell at distances of 2.02 and 2.38 Å, respectively, and one C atom and a second S atom in the second shell at 2.78 and 2.93 Å, respectively. These results suggest thiol, disulfide/disulfane, COOH, and phenol OH groups of HA as ligands for Hg complexation.

Although Mössbauer spectroscopy is highly specific for the isotope $^{57}Fe$, this technique is able to distinguish unambiguously between $Fe^{3+}$ and $Fe^{2+}$ ions that occur mostly in high-spin forms in combination with HS. Mössbauer parameters calculated from the representative Mössbauer spectrum of a soil HA shown in **Figure 6a** suggest the presence of three sites for $Fe^{3+}$ ions in the HA, two with octahedral and one with tetrahedral coordination. Chemical reduction with hydrazine of the HA sample results in a Mössbauer spectrum in which one of the two doublets assigned to $Fe^{3+}$ ions in octahedral sites disappears (3a in **Figure 6a**), and a new doublet appears that is typical of $Fe^{2+}$ ions in the same type of sites (3b in **Figure 6b**). Mössbauer studies of several $Fe^{2+}$–HA complexes suggest the presence of $Fe^{2+}$ ions either as a completely hydrated form in outer-sphere complexes or as a partially hydrated form bound to $COO^-$ and phenolic and alcoholic OH- and N-containing groups of HA in elongated octahedral arrangements in inner-sphere, chelate-type complexes. Mössbauer evidence of partial oxidation of $Fe^{2+}$ to $Fe^{3+}$ is obtained on exposure to air of some $Fe^{2+}$–HA complexes.

The slight decrease observed in the line widths of the water NMR signal in a water solution of soil FA and $Mn^{2+}$ suggests the formation of outer-sphere complexes of hexahydrated $Mn^{2+}$ ion with FA in a slightly distorted octahedral environment. Differently, the great decrease measured in solutions of FA with $Fe^{3+}$ or $Cu^{2+}$ suggests the formation of inner-sphere coordination of these ions with FA. The involvement of O- and N-containing groups is suggested by $^{113}Cd$-NMR in Cd complexation by soil FA in acidic conditions. Modern, two-dimensional NMR techniques such as total correlation (TOCSY), nuclear Overhauser effect (NOESY), heteronuclear Overhauser effect (HOES), and inverse-detection heteronuclear multiple quantum coherence (HMQC) spectroscopies are very promising in the studies of metal–HS interactions.

## Quantitative Aspects

### Metal-Binding Capacity of HS

In general, the binding or complexation capacity (CC) of a multiligand system such as HS is defined as the maximum capacity of the system for binding metal ions. To a good approximation, the CC of HS may be considered as the weighted average of the complexation capacities of the individual ligands $(CC)_i$ in the system, according to:

$$CC = \frac{\sum (CC)_i [\text{weight}]_i}{\sum [\text{weight}]_i} \qquad [1]$$

where $[\text{weight}]_i$ is a weighting factor related to the relative abundance of the ligand in the multiligand system. Given the relative abundance of acidic functional group ligands, primarily COOH, and the low concentrations of other O-, N-, and S-containing ligands in HS, the maximum binding capacity of HS approximates the total acidic functional group content.

In practice, the measured complexation capacity of HS is dependent on experimental conditions, being a function of various factors that include pH, ionic strength, HS concentration and properties, and nature of the metal ion. The effect of pH on metal–HS complexation results from both changes in the extent of ionization of COOH groups and hydrolysis reactions of the metal ion involving the formation of oxyhydroxides. In general, the complexation capacity

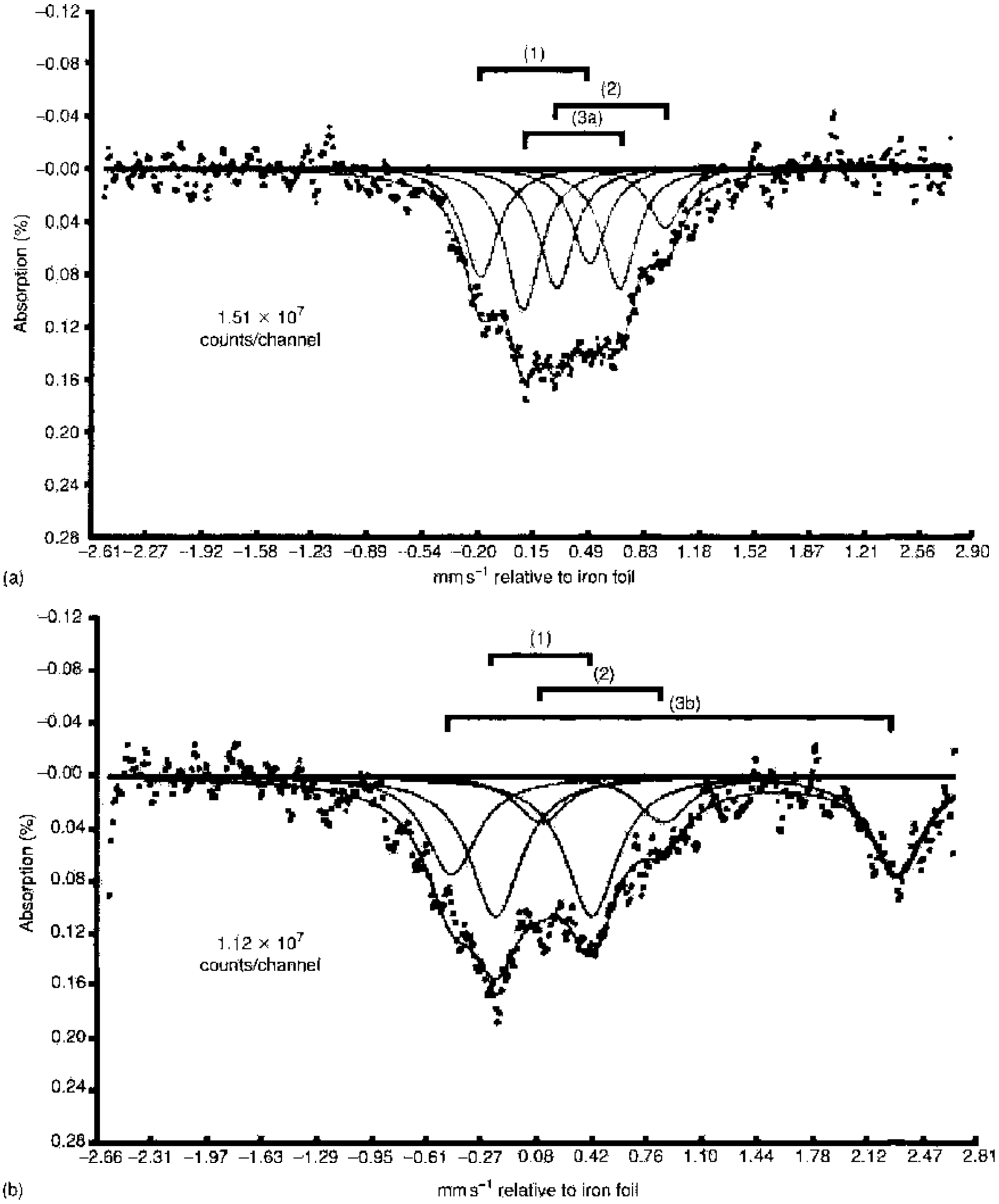


**Figure 6** Mössbauer spectra of a soil humic acid (HA) (a) and the same HA after reduction with hydrazine (b). Doublets (1), (2), (3) refer to three different sites for Fe in HA. Doublets (3a) and (3b) refer, respectively, to $Fe^{3+}$ and $Fe^{2+}$ ions in similar octahedral sites. (Reproduced with permission from Senesi N, Griffith SM, Schnitzer M, and Townsend MG (1977) Binding of $Fe^{3+}$ by humic materials. *Geochimica Cosmochimica Acta* 41: 969–976.)

of HS is reported to increase with increasing pH from acidity up to approximately neutrality and increasing HS concentrations, to decrease at high ionic strength, and to vary with the nature of the metal ion. For any given pH and ionic strength, trivalent cations are bound in greater amounts than divalent ones, of which those forming strong coordination complexes (e.g., $Cu^{2+}$) are bound to a greater extent and at stronger binding sites than those forming weak complexes (e.g., $Mn^{2+}$).

## Stability Constants and Related Parameters of Metal–HS Complexes

The stability constant is probably the most important quantitative parameter for the characterization of a metal–ligand complex in that it provides a numerical index of the affinity of the metal cation for the ligand. Several different theoretical and experimental approaches have been attempted for the determination of stability constants of metal–HS complexes.

The overall reaction of a single metal ion (M) and a single ligand (L) to form the complex $M_mL_n$:

$$mM + nL \rightleftharpoons M_mL_n \qquad [2]$$

can be described by:

$$K = \frac{\{M_mL_n\}}{\{M\}^m\{L\}^n} = \frac{[M_mL_n]}{[M]^m[L]^n} \times \frac{\gamma_{M_mL_n}}{(\gamma_M)^m(\gamma_L)^n} = K_c \times \Gamma \qquad [3]$$

where braces and square brackets denote activities and concentrations, respectively, $\gamma$-values are activity coefficients, $K$ is the true 'overall thermodynamic stability (equilibrium) constant,' $K_c$ is the 'concentration quotient,' and $\Gamma$ is the activity coefficient ratio. For convenience, the charges on the metal, ligand, and complex are not shown explicitly, even if their values are extremely important.

Most experimental investigations of metal–HS complexation reactions yield concentrations, rather than activities, of reactants and products. Thus, thermodynamic constants cannot be measured directly but must be indirectly obtained by calculation or by extrapolation methods from the concentration quotients whose values are not constant but depend, as well as those of the activity coefficient ratios, on the ionic concentration (ionic strength) of the solution. Basic electrostatic considerations (Debye–Hückel theory) indicate that $\Gamma$ equals 1, and thus $K_c$ values equal $K$ values, at zero ionic strength; and $\Gamma$ increases, and $K_c$ tends to decrease, with increasing ionic strength. The concentration of the complex $M_mL_n$ is thus expected to decrease upon addition of a background electrolyte to the solution of metal and ligand.

Competition from side reactions, especially the acid–base chemistry of both the metal ion and the ligand, i.e., hydrolysis of the metal ion to produce hydroxy complexes and protonation of the ligand, is another factor that affects the extent of complexation of M by L, which must be duly considered. In general, protonation of ligands occurs at low pH, and hydrolysis of metals at high pH. Thus, the most favorable condition for measuring the complexation of M by L is at intermediate pH values.

Metal–ligand complexation reactions are usually investigated at constant pH and using experimental techniques that are able to distinguish the metal–ligand complex from the uncomplexed metal and all forms of the ligand that are not bound to the metal ion. Thus, a 'conditional stability constant,' or more precisely a 'conditional concentration quotient,' $K_c^*$, can be written:

$$K_c^* = \frac{[M_mL_n]}{\frac{[M]^m}{(\alpha_M)^m} \times \frac{[L]^n}{(\alpha_L)^n}} = (\alpha_M)^m(\alpha_L)^n \frac{[M_mL_n]}{[M]^m[L]^n}$$
$$= (\alpha_M)^m \times (\alpha_L)^n \times K_c \qquad [4]$$

where $\alpha_M$ and $\alpha_L$ are the side-reaction coefficients of the metal ion and the ligand, respectively. Eqn [4] describes the formation and stability of the complex relative to all other forms of the metal and ligand that may exist under the actual experimental conditions. Further, eqn [4] shows that the conditional concentration quotient, $K_c^*$, includes a pH-invariant but ionic strength-dependent term, $K_c$, and pH-dependent terms, $\alpha_M$ and $\alpha_L$.

Assuming for simplicity a 1:1 stoichiometry for all $ML_i$ complexes, the complexation reaction of a single metal ion M with a multiligand system such as HS containing a number of ligands $L_i$ can be described by an 'average stability constant,' or, more appropriately, an 'average concentration quotient':

$$\bar{K}_c = \frac{\sum[ML_i]}{[M]\sum[L_i]} \qquad [5]$$

where $\sum[ML_i]$ and $\sum[L_i]$ are, respectively, the sum of the concentrations of all 1:1 complexes and all uncomplexed ligands in the system. In practice, the metal ion will distribute itself among the various ligands on the basis of its stability constant with each ligand and of the relative concentration of that ligand. The metal ion will thus tend to complex preferentially with ligands having the highest stability constant and/or highest concentration.

Similar to single-ligand systems, an 'average conditional concentration quotient,' also called 'stability function,' $\bar{K}_c^*$, that takes pH effects into account, can be defined for multiligand systems such as HS. This quantity is a measure of the overall binding of a single metal ion to a multiligand system. However, it cannot be calculated rigorously for HS, because there is not sufficient information available to enable the full description of side reactions and the calculation of side-reaction coefficients of HS for which individual components are unknown. Thus the function $\bar{K}_c^*$ is often calculated directly from experimental data as:

$$\bar{K}_c^* = \frac{C_M - [M]}{[M](C_L - C_M + [M])} \qquad [6]$$

where $C_M$ and $C_L$ are the total stoichiometric concentrations of the metal and ligand in the system examined and $[M]$ is the concentration of the free metal ion. The value of $(C_M - [M])$ corresponds to the sum of the concentrations of all complexes formed between M and the multiligand system, and is calculated by neglecting the presence of inorganic metal complexes. The quantity ($C_L - C_M + [M]$) represents the sum of the concentrations of all binding sites that are not associated with M, and is calculated assuming an average 1:1 metal-to-ligand stoichiometry for the total of binding sites. Stability functions can also be defined for complexes of stoichiometries other than 1:1.

Mean $\bar{K}_c^*$ values are ultimately functions of ionic strength, pH, and degree of saturation of the multiligand system with the metal ion. At a given pH and ionic strength, $\bar{K}_c^*$ will decrease steadily as the total metal-to-ligand ratio ($C_M$:$C_L$) increases. At low metal-to-ligand ratios, the occurrence of preferential reactions of stronger ligands will determine the functional nature of $\bar{K}_c^*$. Thus, it is not possible to obtain true stability constants for metal–HS complexation, and reported values are not actually constant, because the functional dependence is a fundamental feature of systems such as HS, which cannot be eliminated by any experimental method unless total fractionation be achieved into pure ligands that can be studied separately.

In conclusion, the ill-defined nature of multiligand and polyelectrolytic HS systems and of the types, concentrations, and strengths of their several nonidentical binding sites prevent metal–HS complexation reactions to be described in rigorous mathematical terms. However, the stability function concept provides an important overview and useful insights into the nature of metal binding by HS, and may also provide useful indications for the most appropriate experimental methods with which to investigate metal–HS interactions.

## Modeling and Models

The large body of field and laboratory data accumulated on the binding of metal ions and protons by HS has determined the development of chemical speciation models that could encapsulate the ample knowledge and information acquired, and allow applications to 'real-world' situations. The prevalent effort of any modeling approach is the quantitative description of the relative concentrations and strengths of the many nonidentical binding sites contained in HS. The various models applied to describe metal–HS complexation can be classified as 'discrete' ligand models and 'continuous' multiligand models.

In the discrete ligand approach, only a few ligands or binding sites are required to fit the experimental data. Since early modeling work, the competition for HS binding sites between metal ions and protons has been recognized, and metal binding has been estimated from displacements of pH titration curves. Successively modified and refined discrete models have incorporated site heterogeneity effects, i.e., variability in the intrinsic binding strengths among sites, and electrostatic effects related to the net negative charge of HS. Recently, a simplified and improved model (model VI), also based on the discrete-site/electrostatic formulation but moving toward a more distributional approach in describing binding sites for metals, has been developed and applied with success to describe published data sets on a range of metals.

The continuous multiligand distribution models are based on the assumption that a large number of heterogeneous ligands with a range of binding affinities are involved in metal binding with formation of complexes characterized by a continuous (e.g., Gaussian) distribution of the stability constants. The initial, noncompetitive Gaussian distribution models that do not account for proton–metal competition have been extended in the competitive Gaussian distribution models that characterize HS binding-site heterogeneity by using a continuous distribution of equilibrium constants, together with electrostatic submodels, and by assuming monodentate binding of metals to proton-binding sites. The most recent of these models is the NICA (nonideal competitive adsorption)–Donnan model that fits very well extensive experimental data sets obtained for a wide range of metal ion concentrations. The incorporation of multidentate sites into the NICA-Donnan model is under consideration.

A conceptually simple and calculation-convenient model is the 'quasiparticle' model, which includes previous concepts and models, and describes mathematically an aqueous system containing HS by replacing it with a set of hypothetical, average noninteracting molecules whose behavior mimics that of the actual HS system.

Since the primary purpose of any metal–HS binding model is to perform chemical speciation calculations for field situations, validation of the model on field data is required. Models formulated have been applied with variable success. More recent models reproduce trends quite well but lack of knowledge still exists on some key factors, and empirical adjustment of these factors is necessary to fit the data. Thus, more extensive field-testing is required to establish the usefulness or otherwise of any model of metal–HS binding.

## Experimental Methods

A variety of separation and nonseparation methods have been used to speciate metal ions in the presence of HS, to assess the complexing capacity of HS, and to calculate conditional stability constants or quotients of metal–HS complexes. Most commonly used separation approaches include: (1) proton-release titration; (2) cation exchange with synthetic resins; (3) gel filtration or permeation chromatography (GPC); (4) high-performance cation exchange chromatography (HP-CEC); (5) conventional dialysis and equilibrium dialysis–ligand exchange (EDLE); (6) ultrafiltration (UF); and (7) capillary electrophoresis (CE).

Separation approaches generally present the great advantage of allowing measurement of nearly any metal ion by sensitive techniques such as atomic absorption and plasma spectrometries. However, dialysis and chromatographic methods suffer from the disadvantages of disturbance of complexation equilibria and adsorption of interacting species on the membrane or chromatographic material, whereas equilibrium dialysis and cation exchange present uncertainties in the definition of chemical forms actually measured by the method. An important advantage of the UF methods is that they do not take such a long time to carry out as dialysis and do not disturb the complexation equilibria.

Most important nonseparation methods include: (1) ion-selective electrode (ISE) potentiometry; (2) voltammetric techniques such as anodic-stripping voltammetry (ASV); and (3) various spectroscopic techniques, including fluorescence, ESR, and NMR spectroscopies. Some nonseparation methods such as ISE and ASV, although being the most sensitive and advanced, also suffer from the problems of adsorption and shifts in equilibria. Further, these methods are applicable adequately only to a limited typology of metal ions, e.g., $Cu^{2+}$, $Cd^{2+}$, $Pb^{2+}$, and $Zn^{2+}$.

The fluorescence method presents a number of advantages over other methods. Titration curves of HS fluorescence quenching versus concentration of added metal quencher have allowed direct measurement of the complexing capacity of HS ligands and the stability constants of HS–metal complexes. The method is relatively rapid, since no separation is required between bound and free metal ion, thus errors associated with the separation step in most speciation methods are avoided. Neither supporting electrolyte, nor buffer, nor adsorbing material are required to be added to the sample; it is even more sensitive than ASV and ISE potentiometry and is sensitive enough for application to unmodified, natural organic ligands without preconcentration. The major disadvantage of fluorescence spectroscopy is, however, its efficiency only with strongly binding, paramagnetic metal ions such as $Cu^{2+}$. This limitation can be overcome by the use of fluorescent probes, such as in the lanthanide ion probe fluorescence spectroscopy (LIPS), where HS samples are titrated by $Eu^{3+}$ ions, and the titration plot of the ratio of the intensities of two emission lines of $Eu^{3+}$ is used to estimate the amount of bound and free species of the probe ion.

The ESR approach allows the determination of the free metal ion concentration and therefore of its degree of complexation with the ligand. ESR spectroscopy has been applied to obtain stability constants and stoichiometries in aqueous solution of $VO^{2+}$ and $Mn^{2+}$ complexes with FA. On these bases, the $VO^{2+}$–FA complexes are modeled as bis(phthalato)(salicylato)-oxovanadium (IV) and mono(salicylato)-oxovanadium (IV). The $K_c$ values measured for $Mn^{2+}$–FA complexes by ESR are in excellent agreement with $K_c$ values determined by an ion-exchange method. Major advantages of ESR spectroscopy are its high sensitivity and the minimal or no-sample pretreatment required, whereas its intrinsic limitation is applicability only to paramagnetic metal ions such as Cu, Fe, Mn, and V.

Few NMR experiments have been conducted to relate structure of HS to their metal-binding capacity. The positive correlation found between aromaticity of several HA determined by solution-state $^{13}C$-NMR and stability constants of their complexes with $Cd^{2+}$ and $Zn^{2+}$ ions suggests the involvement of phenolic groups in the complexation of these metals. The higher content of carboxylate groups in a soil FA than in a soil HA measured by distortion enhancement by polarization transfer (DEPT)-$^{13}C$-NMR and quaternary-C only (QUAT) subspectra is consistent with the higher binding constant measured for $Eu^{3+}$ with FA than with HA. Dipolar dephasing cross-polarization magic angle spinning (CPMAS) $^{13}C$-NMR has shown that the stability constant of a $Cd^{2+}$–sediment HA is greater than that of $Cd^{2+}$–soil HA, probably due to the higher content and stronger binding of amino acidic N groups, in addition to carboxylic groups, of the sediment HA. Further, great metal stability constants of $Zn^{2+}$–HA and $Cd^{2+}$–HA complexes are found to be associated with a higher degree of substituted aromatic C groups, possibly due to chelation effects.

## List of Technical Nomenclature

| | |
|---|---|
| $\alpha_L$ | Side-reaction coefficient of the ligand |
| $\alpha_M$ | Side-reaction coefficient of the metal ion |
| $\mathbf{\Gamma}$ | Activity coefficient ratio |
| Å | Ångstrom ($10^{-8}$ cm) |
| $K$ | Thermodynamic stability (equilibrium) constant |
| $\bar{K}$ | Mean stability constant or mean concentration quotient |

| | |
|---|---|
| $K_c$ | Concentration quotient |
| $K_c^*$ | Conditional stability constant or conditional concentration quotient |
| $\bar{K}_c^*$ | Mean conditional concentration quotient or stability function |
| mT | Millitesla |

*See also:* **Chemical Equilibria; Colloid-Facilitated Sorption and Transport; Electron-Spin Resonance Spectroscopy; Fluorescence Spectroscopy; Fourier Transform Infrared Spectroscopy; Heavy Metals; Organic Matter:** Principles and Processes; **Oxidation–Reduction of Contaminants; Sorption:** Metals; **Surface Complexation Modeling**

## Further Reading

Buffle J (1988) *Complexation Reactions in Aquatic Systems: An Analytical Approach.* Chichester, UK: Ellis Harwood.

Clapp CE, Hayes MHB, Senesi N, Bloom PR, and Jardine PM (eds) (2001) *Humic Substances and Chemical Contaminants.* Madison, WI: SSSA Publications.

MacCarthy P and Perdue EM (1991) Complexation of metal ions by humic substances: fundamental considerations. In: Bolt GH, De Boodt MF, Hayes MHB, and McBride MB (eds) *Interactions at the Soil Colloid–Soil Solution Interface*, pp. 469–492. NATO ASI Series, vol. 190. Dordrecht, The Netherlands: Kluwer.

Senesi N (1990) Molecular and quantitative aspects of the chemistry of fulvic acid and its interaction with metal ions and organic chemicals, part II. The fluorescence spectroscopy approach. *Analytica Chimica Acta* 232: 77–106.

Senesi N (1990) Application of electron-spin resonance (ESR) spectroscopy in soil chemistry. In: Stewart BA (ed.) *Advances in Soil Science,* vol. 14, pp. 77–130. New York: Springer-Verlag.

Senesi N (1992) Metal–humic substance complexes in the environment. Molecular and mechanistic aspects by multiple spectroscopic approach. In: Adriano DC (ed.) *Biogeochemistry of Trace Metals*, pp. 429–496. Boca Raton, FL: Lewis.

Sposito G (1986) Sorption of trace metals by humic materials in soils and natural waters. *CRC Critical Reviews in Environmental Control* 16: 193–229.

Stevenson FJ (1976) Binding of metal ions by humic acids. In: Nriagu JO (ed.) *Environmental Biogeochemistry. Metals Transfer and Ecological Mass Balances,* vol. 2, pp. 519–540. Ann Arbor, MI: Ann Arbor Science.

Stevenson FJ (1994) *Humus Chemistry. Genesis, Composition, Reactions*, 2nd edn. New York: Wiley Interscience.

Stevenson FJ and Fitch A (1986) Chemistry of complexation of metal ions with soil solution organics. In: Huang PM and Schnitzer M (eds) *Interactions of Soil Minerals with Natural Organic Microbes*, pp. 2–58. Madison, WI: SSSA Publications.

# ORGANIC RESIDUES, DECOMPOSITION

**A J Franzluebbers**, USDA Agricultural Research Service, Watkinsville, GA, USA

Published by Elsevier Ltd.

## Introduction

Organic residues are carbon-containing compounds of biological origin. Decomposition is the breakdown of these complex organic materials into simpler components. Decomposition of organic residues in soil is an important ecological function whereby heterotrophic organisms consume various components, resulting in the physical and biochemical breakdown of organic materials and transformation and cycling of constituent elements. Biochemical composition, environmental factors, and diversity of organisms (*See* **Biodiversity**) play major roles in the dynamics and fate of organic residues in soil (*See* **Humification; Organic Matter:** Principles and Processes; Genesis and Formation). Decomposition is an important soil process that affects nutrient availability on a local scale (*See* **Nutrient Availability**) and environmental quality on a global scale by controlling carbon transformations (*See* **Carbon Cycle in Soils:** Dynamics and Management; **Carbon Emissions and Sequestration**), greenhouse gas emissions (*See* **Greenhouse Gas Emissions**), and nutrient fluxes (*See* **Nitrogen in Soils:** Cycle). The most common methods of determining decomposition are by mass loss of residues or by complete mineralization of organic constituents into end products of inorganic carbon ($CO_2$, $CH_4$), nitrogen ($NH_4$, $NO_3$), or phosphorus ($PO_4$).

## Types of Organic Residues

Organic residues that enter the soil ecosystem are derived primarily from: (1) autotrophic organisms, including a wide diversity of plants, algae, and lichens; (2) animal tissue and manures; and (3) various organic additions from human activities, including pesticides, oil products, and synthetic compounds (*See* **Pollutants: Biodegradation**). The presence of heterotrophic organisms in soil, such as fauna and microorganisms, contributes directly to the type of organic residues through synthesis of unique compounds in their bodies, as well as indirectly through transformations of organic residues during decay and subsequent humus synthesis (*See* **Bacteria: Soil; Fauna; Fungi; Nematodes**).

Plant residues are, by far, the most dominant input of organic residues to the soil ecosystem. The quantity of residues produced varies with environmental conditions, but typically ranges from 200 to 2000 g · m$^{-2}$ per year. The allocation of total plant biomass between roots and shoots depends upon a number of factors, including annual or perennial vegetation, soil fertility, climate, and various other soil physical and chemical characteristics. Generally, annual plants tend to have shoot to root ratios of more than 1, while perennial plants tend to have lower shoot to root ratios. Rooting depth of plants depends upon factors similar to those affecting shoot to root ratio. In general, about 50% of total root biomass is found in the surface 20 cm of soil. Accurate characterization of root inputs is difficult because of the intimate contact between roots and the soil-decomposer community, which acts quickly on substrates that are unprotected in the moist soil environment. The portion of root biomass that decomposes rapidly and does not persist in soil is often called 'rhizodeposition.' Rhizodeposits are typically composed of sloughed root cap cells, organic acids actively secreted by growing roots, lysates of root tissues, and high molecular weight root mucilage.

Plant residues are composed of: (1) highly labile intracellular compounds such as protein, starch, fructan, and chlorophyll and other pigments; (2) moderately resistant structural compounds such as cellulose and hemicellulose; and (3) resistant structural compounds such as lignin, polyphenols, lipids, and cutin. If isolated from the whole plant, individual components would decompose at different rates due to their susceptibility to microbial utilization (**Figure 1**). The rate of decomposition of whole-plant materials varies in response to the relative proportion of different intracellular and structural compounds, as outlined in the following.

For many field crop residues, C concentration is relatively constant at 41% of dry matter (**Figure 2a**). Protein concentration (estimated as 6.25 times the N concentration), however, varies greatly among residues by species, plant part, age, and management. The amount of C and N mineralized from crop residues varying in N concentration during incubation in soil is highly related to the C to N ratio of residues (**Figure 2b**).

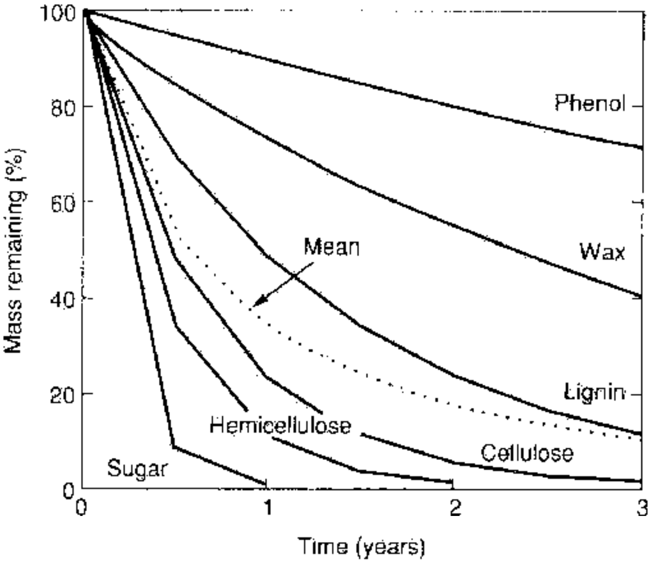


**Figure 1** Theoretical decomposition of various components of plant cells, expressed as mass remaining during a 3-year period. Phenol, wax, and lignin groups decompose slower than the average rate of decomposition of plant material, and cellulose, hemicellulose, and sugar decompose faster. Mean proportion of the whole plant as phenol is 5%, as wax is 5%, as lignin is 40%, as cellulose is 20%, as hemicellulose is 15%, and as sugar is 15%. These proportions will vary depending upon plant species, age, fertility, and environmental conditions. Decomposition of organic residues typically follow first-order kinetics ($y_t = e^{-k \cdot t}$, where $y$ = mass remaining at time $t$ and $k$ is a nonlinear rate constant). (Source: Minderman G (1968) Addition, decomposition and accumulation of organic matter in forests. *Journal of Ecology* 56: 355–362.)

Leguminous herbaceous and woody plant species are often applied to soil as mulch or incorporated as green manure to improve soil fertility, especially in the humid tropics where decomposition conditions are ideal. Decomposition of these materials is dependent not only upon N concentration (positive), but also on the concentration of polyphenolics (negative), which are reactive compounds that can form stable polymers with various N compounds and slow decomposition (**Figure 3**).

The amount of lignin in plant tissue has a negative impact on the rate of decomposition, although some tissues may have a high protein concentration (**Figure 4**). Lignin is a high molecular weight polymer composed of phenyl propane units in cell walls, giving plants rigidity and resistance to microbial attack. Aside from polysaccharides, lignin is one of the most abundant biopolymers in nature. Lignin is relatively resistant to microbial decomposition, being limited primarily to a group of white-rot fungi that can completely decompose it (*See* **Fungi**). In addition, a diverse community of soil organisms is able to work in concert to decompose parts of the lignin structure slowly (*See* **Biodiversity**).

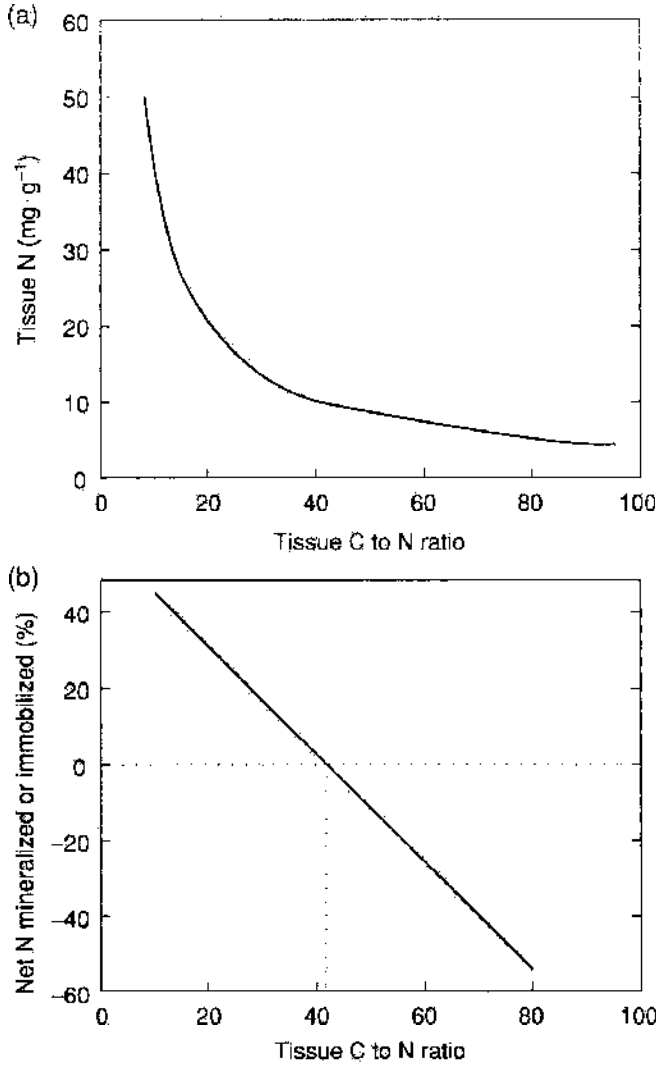


**Figure 2** Relationship of C to N ratio of various crop residues with (a) N concentration and (b) net N mineralization during aerobic incubation. The close relationship between C to N ratio and N concentration (a) indicates uniform C concentration (410 mg · g$^{-1}$) and varying N concentration. Under defined incubation conditions (i.e., optimum temperature and moisture for 20 ± 10 weeks), the percentage of N mineralized from plant tissues varying in C to N ratio (b) is positive when the C to N ratio is less than 42 (i.e., net mineralization of N) and is negative when the C to N ratio is more than 42 (i.e., net immobilization of N). (Source: Vigil MF and Kissel DE (1991) Equations for estimating the amount of nitrogen mineralized from crop residues. *Soil Science Society of America Journal* 55: 757–761.)

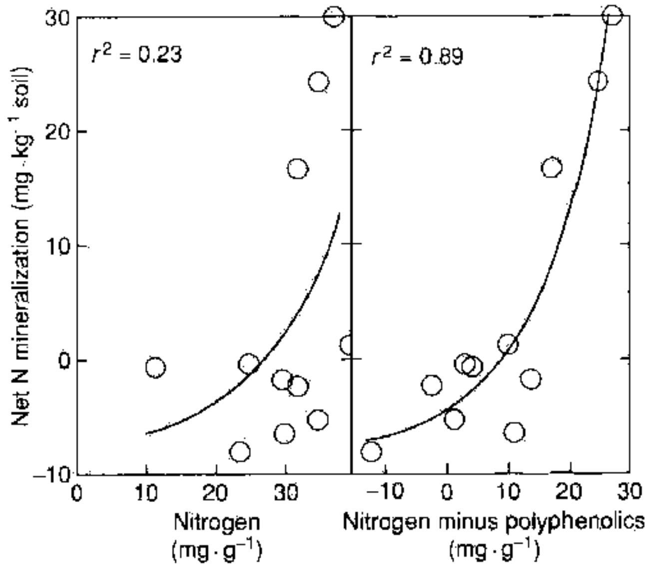


**Figure 3** Relationship of N mineralization from different legume green manures (a) with N concentration and (b) with N concentration minus polyphenolic concentration. Although N concentration alone is typically a good predictor of potential N mineralization, condensation of N with polyphenolics can inhibit mineralization. Therefore, information on polyphenolic content of residues is needed to obtain a better prediction of N mineralization. (Source: Palm CA and Sanchez PA (1991) Nitrogen release from the leaves of some tropical legumes as affected by their lignin and polyphenolic contents. *Soil Biology and Biochemistry* 23: 83–88.)

## Environmental Factors

Water directly affects decomposition through its controls on activity and transport of soil microorganisms, solubilization of organic constituents, oxygen supply, and soil pH (*See* Microbial Processes: Environmental Factors). An optimum water content for decomposition is within a range that is not too high to limit oxygen availability and that is not too low to limit substrate availability and mobility of soil microorganisms. The ideal soil moisture condition for decomposition is between −10 and −50 kPa (or 30–60% water-filled pore space). Fluctuating moisture conditions in soil are common, but they do not necessarily have a negative impact on the ability of microorganisms to respond quickly to available substrates upon rewetting (Figure 5).

Temperature controls the biochemical activity of intra- and extracellular enzyme activities, as well as metabolic activity of most soil organisms responsible for the biodegradation of organic residues. For every 10°C rise in temperature, decomposition generally increases one- to two-fold.

Oxygen is a requirement for the dominant heterotrophic soil organisms responsible for decomposition. Oxygen serves as the most efficient electron acceptor during microbial respiration. When oxygen becomes limiting in soil, decomposition proceeds at a much slower rate. Alternative electron acceptors, including $NO_3^-$, $Fe^{3+}$, $Mn^{4+}$, $SO_4^{2-}$, and $CO_2$, can be utilized by anaerobic bacteria such as denitrifiers, sulfate reducers, and methanogens.

Soil pH affects decomposition of organic materials because of its direct effect on the activity of heterotrophic microorganisms. As soil pH decreases to less than 4, microbial activity is severely retarded and accumulation of organic matter can occur, as

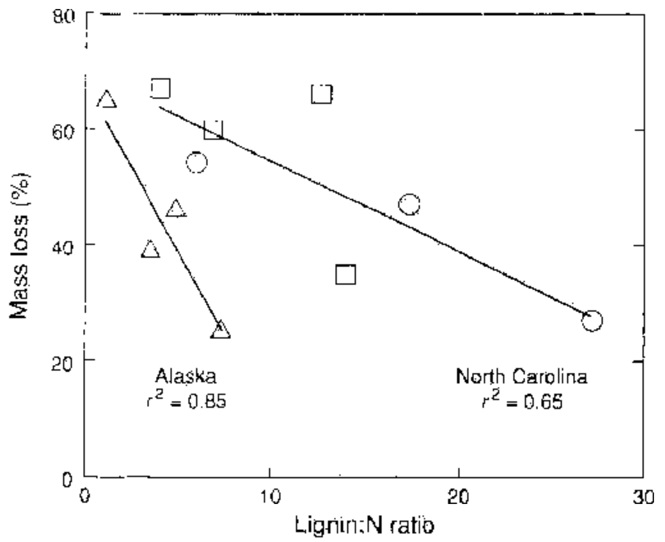


**Figure 4** Decomposition of surface-placed residues as affected by their lignin-to-N ratio. Lignin is a polymer that is difficult for most soil microorganisms to decompose. Therefore, high lignin content reduces decomposition rate. At very low lignin content, decomposition of plant material would be similar whether the climate would be cold (Alaska) or warm (North Carolina). However, with increasing lignin content, decomposition is greatly reduced in a cold climate (Alaska) relative to a warm climate (North Carolina). Data from North Carolina are mass loss at the end of 1 year (circles: Blair JM (1988) Nitrogen, sulfur, and phosphorus dynamics in decomposing deciduous litter in the Southern Appalachians. *Soil Biology and Biochemistry* 20: 693–701.) and C loss at the end of 44 weeks (squares: Buchanan M and King LD (1993) Carbon and phosphorus losses from decomposing crop residues in no-till and conventional till agroecosystems. *Agronomy Journal* 85: 631–638.). Data from Alaska are mass loss at the end of 1 year. (Koenig RT and Cochran VL (1994) Decomposition and nitrogen mineralization from legume and non-legume crop residues in a subarctic agricultural soil. *Biology and Fertility of Soils* 17: 269–275.)

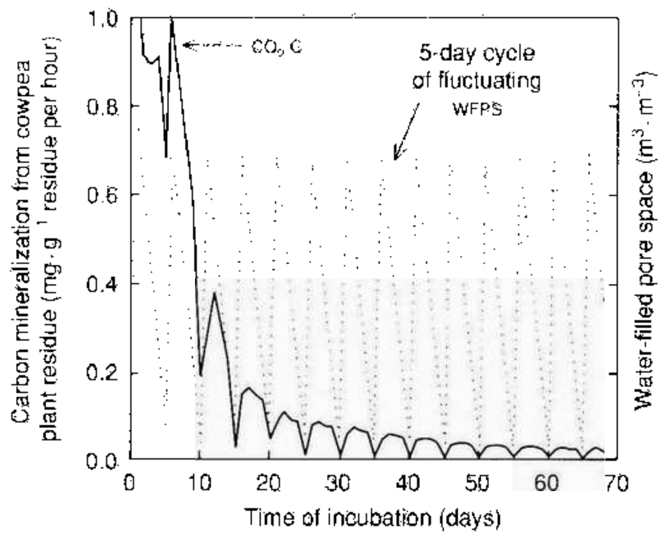


**Figure 5** Carbon loss rate of decomposing leaf tissue during successive cycles of drying and wetting. Lowest points of each cycle have a water potential of −10 MPa. Extreme drying reduces decomposition, but decomposition resumes very rapidly upon rewetting. WFPS, water-filled pore space. (Source: Franzluebbers K, Weaver RW, Juo ASR, and Franzleubbers AJ (1994) Carbon and nitrogen mineralization from cowpea plants part decomposing in moist and in repeatedly dried and wetted soil. *Soil Biology and Biochemistry* 26: 1379–1387.)

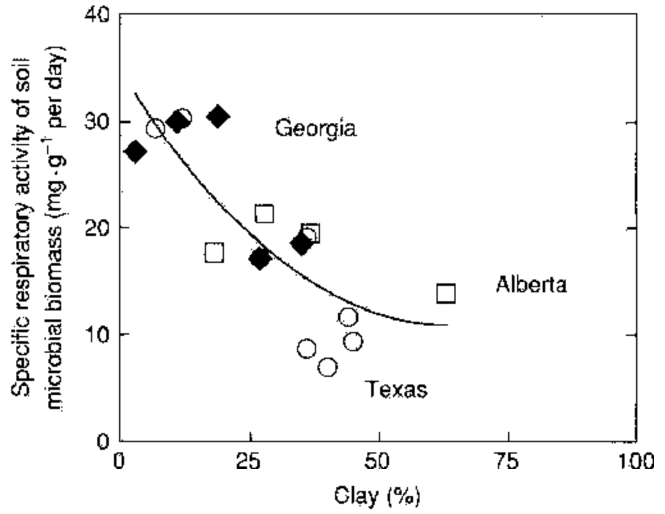


**Figure 6** Relationship of microbial-specific mineralization with soil clay content (AJ Franzluebbers, unpublished data). Soils with fine texture have higher clay content and typically more aggregated soil that can protect a portion of the soil organic matter within aggregates. The higher specific respiratory activity of microbial biomass in coarse-textured soil than in fine-textured soil suggests a lower level of protection of the microbial community against predation by other microorganisms and by soil fauna.

especially evidenced by the thick litter layer that often develops under forests on acidic soils (e.g., mor).

Organic materials can become isolated, and even protected from decomposition to some degree, once they become incorporated into water-stable soil aggregates (*See* **Aggregation:** Microbial Aspects). Protection within these aggregates may be due to the physical barrier that limits soil faunal activity and, at times, due to low oxygen concentration that limits soil microbial activity (*See* **Structure**). Soil texture (*See* **Texture**) offers a similar physical limitation on organic matter decomposition, where the pores in coarse-textured soils are freely accessed by a wide variety of micro-, meso-, and macrofauna. Pores in fine-textured soils are often more diverse, with micropores that limit faunal activity and protect organic matter from decomposition typically not present in coarse-textured soil (**Figure 6**).

Soil management can have large impacts on residue decomposition by altering the physical environment and controlling the supply of substrates. Whether organic residues are left on the soil surface or incorporated into soil (*See* **Cultivation and Tillage; Conservation Tillage**) affects the moisture and temperature regimes they occupy. Generally, incorporated residues decompose faster owing to more consistent moisture and temperature conditions in the soil than on the soil. The type of cropping can alter the timing, quantity, and quality of residues added to soil (*See* **Crop Rotations**). Residues that become available to soil organisms for decomposition during dry or cold periods will decompose at a different rate and possibly undergo different transformations because of different participating organisms than residues that become available during hot or wet periods of the year. Whether crops are perennial or annual can influence: (1) the level of soil disturbance, (2) the type, timing, and quantity of substrates supplied to the soil, and (3) the placement of residues.

## Organism Interactions

Various types of organisms feed upon specific components of organic materials in soil. This complexity of interactions among organisms can be illustrated as a food web (**Figure 7**), where bacteria and fungi are principal decomposers at many points within the web and progressively larger organisms tend to feed on more specific components of organic material within the food web (*See* **Food–Web Interactions**). The soil microbial biomass plays a pivotal role in the decomposition of organic inputs from plant and animal residues and the breakdown and transformation of organic matter to and from slow and passive pools (Figure 8).

The presence of soil fauna (*See* **Fauna**) often facilitates decomposition by: (1) comminuting plant residues, which exposes a greater surface area to attack by soil microorganisms; (2) transporting organic residues to new locations in the soil, which facilitates decomposition, interaction with soil nutrients, and isolation from certain environmental conditions; (3) inoculating partially digested organic residues with specific bacteria and enzymes (*See* **Enzymes in Soils**); and (4) altering physical characteristics of soil by creating burrows, fecal pellets, and distribution of soil particles, which influence water (*See* **Water, Properties**), air (*See* **Aeration**), nutrient (*See* **Nutrient Availability**), and energy (*See* **Energy Balance**) retention and transport in the soil.

## Nutrient Cycling

Nitrogen is a key element that is held within organic residues primarily as proteins and amino acids and released to the soil environment as $NH_4^+$ through the process of mineralization (*See* **Nitrogen in Soils: Cycle**). Decomposition of organic residues is

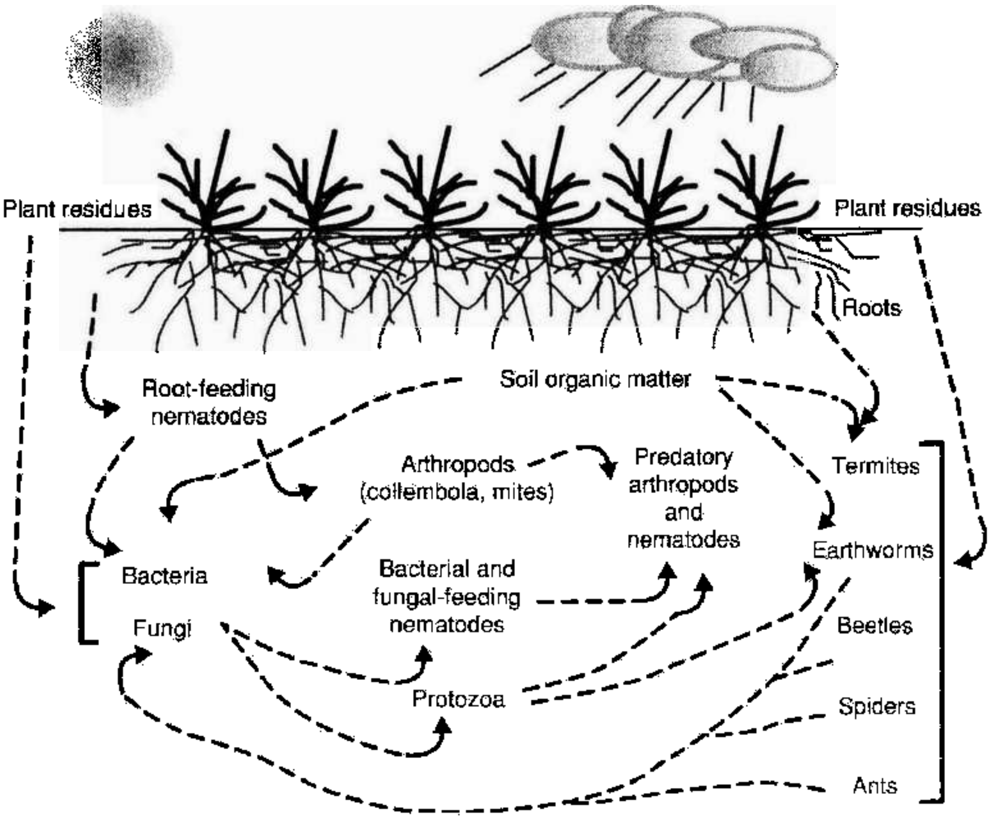


**Figure 7** A soil food web. Plant residues are fed upon by various organisms residing in soil. In turn, the growth and remains of these organisms are fed upon by other heterotrophic organisms in soil until nearly all energy embedded within the organic molecules has been exhausted.

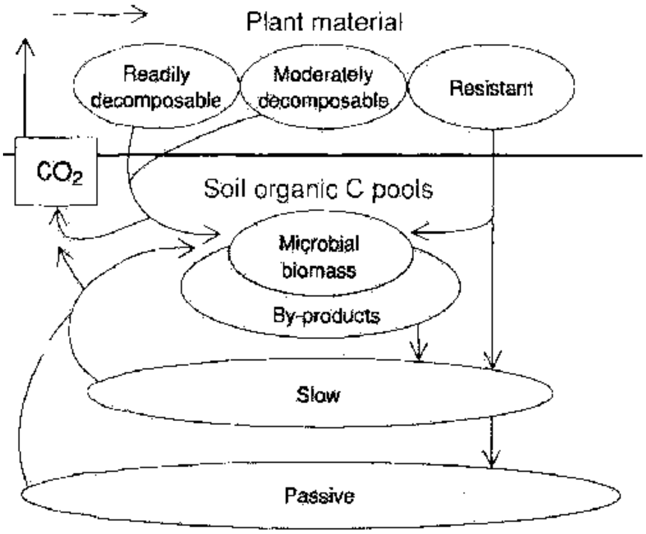


**Figure 8** Pools of carbon in soil and the central role of microbial biomass in transforming organic residues into soil organic matter and releasing carbon dioxide into the atmosphere.

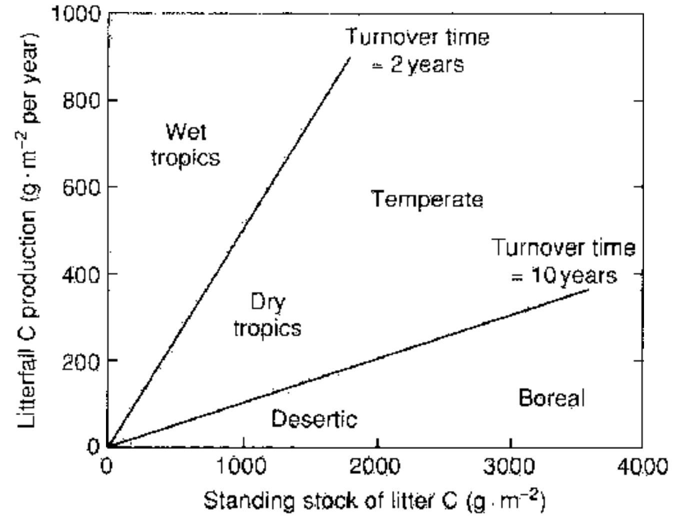


**Figure 9** Decomposition estimates based on annual litterfall and standing stock of litter at the forest floor under different environmental conditions in the world. (Source: Olson JS (1963) Energy storage and the balance of producers and decomposers in ecological systems. *Ecology* 44: 322–331.) With high stock of litter at the forest floor and low litterfall production, slow decomposition occurs due to limited moisture or cold temperature. With low stock of litter and high litterfall production, rapid decomposition occurs due to plentiful moisture and warm temperature.

necessary to release organically bound N into inorganic N, which can then be taken up by plants.

Phosphorus is bound organically in plants primarily as deoxyribonucleic acid (DNA) and the energy-transfer molecules adenosine diphosphate (ADP) and adenosine triphosphate (ATP). Mineralization of P from organic residues can supply not only a portion of plant requirements, but also the rapidly growing microbial population that has a high demand for P to meet its ATP production requirements (*See* **Phosphorus in Soils:** Overview).

Carbon is the primary element in organic materials, representing 35–45% of the dry weight of most plant materials, 40–50% of the dry weight of animal manures, and 45–55% of the dry weight of soil humus. In general, C undergoes complete cycling from atmosphere to plants to soil organic matter to atmosphere (*See* **Carbon Cycle in Soils:** Dynamics and Management; **Greenhouse Gas Emissions**). The residence time of C in the terrestrial sphere can be relatively short in warm–moist environments and longer in cold–dry environments (**Figure 9**), but globally the quantity of C fixed within plant biomass each year is believed to be balanced by the quantity of C released to the atmosphere through decomposition.

Decomposition is an important process that returns organically bound nutrients back to nature for utilization and cycling. On a global scale, this vital ecosystem function balances the autotrophic production of organic matter through photosynthesis. An imbalance in these opposing processes would lead to either: (a) accumulation of organic residues, with a consequent reduction in plant-available nutrients, which could ultimately reduce net primary productivity without external nutrient inputs; or (b) depletion of soil organic matter reserves, which could facilitate greenhouse gas emissions and global warming, thereby depleting inherent soil fertility and reducing net primary productivity. Decomposition is a vital ecosystem function that sustains life on Earth.

## Future Advances

Decomposition of organic residues has been historically and extensively studied, and yet much remains to be learned about the decomposition mechanisms and fate of residues in different environments. Of particular deficiency in our current knowledge are the dynamics and interactions with soil of plant root production and decomposition. More detailed chemical composition and fate of different organic residues are needed. Analytical techniques have relied in the past upon acid digestion procedures that are not particularly well suited for analyzing specific molecular classes of compounds. Future analytical techniques will probably rely more on a combination of spectroscopic, chemolytic, and thermolytic analyses to describe both initial chemical properties and the temporal changes in properties during decomposition.

## List of Technical Nomenclature

| | |
|---|---|
| % | Percent |
| ADP | Adenosine diphosphate |

| | |
|---|---|
| ATP | Adenosine triphosphate |
| C | Carbon |
| C | Celsius |
| d | Day |
| DNA | Deoxyribonucleic acid |
| g | Gram |
| kg | Kilogram |
| kPa | Kilopascal |
| m | Meter |
| mg | Milligram |
| Mpa | Megapascal |
| N | Nitrogen |
| P | Phosphorus |
| yr | Year |

*See also:* **Aeration**; **Aggregation:** Microbial Aspects; **Bacteria:** Soil; **Biodiversity**; **Carbon Cycle in Soils:** Dynamics and Management; **Carbon Emissions and Sequestration**; **Conservation Tillage**; **Crop Rotations**; **Cultivation and Tillage**; **Energy Balance**; **Enzymes in Soils**; **Fauna**; **Food–Web Interactions**; **Fungi**; **Greenhouse Gas Emissions**; **Humification**; **Microbial Processes:** Environmental Factors; **Nematodes**; **Nitrogen in Soils:** Cycle; **Nutrient Availability**; **Organic Matter:** Principles and Processes; Genesis and Formation; **Phosphorus in Soils:** Overview; **Pollutants:** Biodegradation; **Structure**; **Texture**; **Water, Properties**

## Further Reading

Alexander M (1994) *Biodegradation and Bioremediation.* San Diego: Academic Press.

Coleman DC and Crossley DA Jr (1996) *Fundamentals of Soil Ecology.* San Diego: Academic Press.

Kögel-Knabner I (2002) The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter. *Soil Biology and Biochemistry* 34: 139–162.

Paul EA and Clark FE (1996) *Soil Microbiology and Biochemistry.* San Diego: Academic Press.

# ORGANIC SOILS

**D L Mokma**, Michigan State University, East Lansing, MI, USA

## Introduction

Organic soils consist of organic materials derived primarily from plants, in contrast to mineral materials, from which most soils of the world formed. These soils form where production of organic matter exceeds decomposition, usually under conditions of water saturation; as a result, most organic soils are wetlands. Organic soil materials have more than 12–18% organic C depending on the clay content of the mineral fraction (Figure 1). Organic soils are commonly called peatlands. Several terms are commonly used to separate different organic soils. Highly acid organic soils are referred to as bogs. Less acid organic soils are called fens if they lack trees and are dominated by grasses and/or sedges, or swamps if they are at least partially forested. Marsh refers to wet mineral soils that are dominated by grasses and/or sedges.

## Formation

The main soil-forming process active in organic soils is accumulation of organic materials. Paludization is the accumulation of organic materials under anaerobic conditions where preservation allowed a net gain

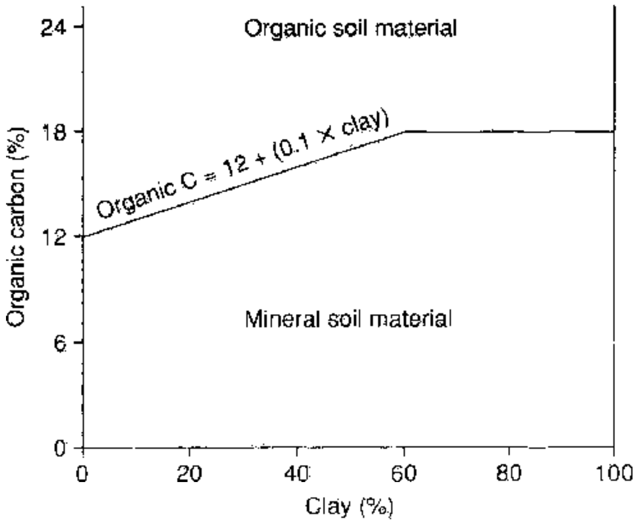


**Figure 1** Organic carbon requirements for organic soil materials based on clay content.

through time. Little transformation of parent materials has occurred because of the anaerobic environment in which they form. Decomposition of organic materials is controlled by interrelated factors, primarily moisture content, temperature, acidity, microbial activity, composition of the deposit, and time. Organic soil formation began when microbial activity was promoted by entry of air into the organic deposit. The processes are physical, chemical, and biological ripening. Physical ripening involves a decrease in volume and depends on the nature of organic matter, content of mineral material, and depth of the water table. Chemical ripening is the decomposition of organic materials. Biological ripening involves the reduction in particle size and mixing of organic materials by living animals. The earth can be divided in two organic soil-forming zones, temperate zone and tropical zone. In temperate zones the median rate of organic soil formation is about 0.6 mm year$^{-1}$ and the carbon contribution is about 21.4 g m$^{-2}$ year$^{-1}$. In tropical zones the rate of formation is about 2.2 mm year$^{-1}$ and the carbon contribution is about 128 g m$^{-2}$ year$^{-1}$.

### Parent Materials

Accumulated organic materials are primarily composed of plant remains. Small amounts of animal materials are included. The initial deposit of organic materials is considered geogenic rather than pedogenic. The deposit is the parent material. The dominant pedogenic process is the alteration of recognizable organic forms of leaves, stems, and roots, to unrecognizable organic materials. The characteristics of organic soils depend primarily on the nature of the vegetation that was deposited in the water and the degree of decomposition. Vegetation growing on organic soils varies with local climatic conditions, groundwater, and topography. Vegetation may differ in different portions of a single area. Changes in vegetation may also occur as the organic materials accumulate.

### Topography

Organic soils may form in closed depressions, coastal marshes, on slopes where there is seepage, or blanket a dissected landscape. The common feature is water that may come from several sources. Climate controls where organic soils form on a regional scale; however, topography determines where they form on a specific landscape. Organic soils form on parts of landscapes where there is concentration of run-on, retention of precipitation, or discharge of groundwater. These conditions most often occur in topographic depressions or very flat areas.

### Organisms

Organic soils in their natural environment can exhibit large numbers of contrasting flora. The flora includes those submerged and floating on water, growing along the shoreline and on the bog surface. The sequence of flora on the bog surface is usually meadow, heath, shrub, and finally forest association. Chemical composition of the organic and inorganic fractions, water level, and pH greatly affect plant associations. As a result, a single area of organic soil may have several plant associations at any one time. Differences in vegetative cover change organic soil characteristics that influence utilization of the soil. Generally, the first stage in organic soil formation is plankton growth in water and chara, hornwort and others in deep water and pond weeds and water lilies in shallow water (**Figure 2**). After some accumulation of organic materials, conditions become favorable for growth of sedges, reeds, and grasses (**Figure 3**). Later the area is covered with heath. Finally, succession leads to shrubs and forests (**Figure 4**).

Plant communities with high primary production usually have high decomposition rates, therefore, organic matter accumulation is negligible. Communities with low or modest primary production frequently have slow decomposition rates, therefore, organic matter accumulation is significant. The anaerobic environment of organic soils preserves pollen, therefore they contain a record of past climates and flora.

### Climate

Organic soils can form in virtually any climatic region; however, water must be available in the region for the organic materials to accumulate. Accumulation of organic materials is favored by cold and wet climates because they inhibit decomposition of the materials. In the high latitudes of Asia, Europe, and North America average precipitation exceeds annual potential evaporation. In the northern part of these latitudes organic soils are not common because short growing seasons inhibit the production of organic materials. In lower latitudes organic soils are primarily limited to coastal plains that are nearly level with high annual precipitation and no dry season or to environments with high water tables maintained by tidal waters or coastal marshes. In arid, semiarid, and arctic areas, organic soil formation is very slow or negligible due to climate.

### Time

Most organic soils have formed since the end of the last ice age. In high latitudes they are concentrated in regions that were covered by glaciers and formed

**Figure 2** First stage of organic soil development: water lilies.

**Figure 3** Sedges and grasses on organic soil.

following retreat of glaciers. In lower latitudes they were also impacted by glaciation. Large quantities of water were incorporated in glacial ice. Melting of glaciers caused sea level to rise at a rate that was more rapid than that of initial vegetation colonizing tidal regions. When the sea level rise slowed, marsh vegetation could become established and organic materials began to accumulate, producing the parent materials for organic soils. About 3.6 m of organic materials accumulated between 4400 BP and 1914 near Belle Glade, FL. In this low-latitude region the average rate of accumulation was about 1.4 mm year$^{-1}$. In Michigan, a more northern latitude, the average rate was 0.55 mm year$^{-1}$. In the slightly glaciated region of southwest Wisconsin, the average rate of accumulation was about 0.68 mm year$^{-1}$. In Iceland it was 0.39 mm year$^{-1}$.

## Distribution

Organic soils comprise about 2% of the Earth's ice-free land surface. This includes 0.8% organic soils with permafrost. Organic soils can form in virtually any climate, even in arid areas, as long as water is available. They are most prevalent in the cool, humid boreal forest areas of northern Asia, Europe, and

**Figure 4** Leather leaf on acidic organic soil.

North America. Every province of Canada and almost every state in the USA has organic soil. More than 95% of the total peat reserves of the world are located in the temperate zone of the northern hemisphere. In these areas precipitation usually exceeds evapotranspiration and summers are relatively cool. At lower latitudes organic soils occur locally on humid coastal plains, for example, South-East Asia, Indonesia, and east North America. These coastal areas are inundated by tidal waters.

In many central European countries organic soils have been extensively used and no longer qualify as organic soils. Nearly all organic soils in the Netherlands, Germany, and Switzerland have been lost. In more northern European countries the organic soils are relatively virgin. About 75% of Europe's remaining organic soil resources are located in Finland, Norway, and Sweden. Because of the loss of organic soils resulting from intensive utilization, it is difficult to obtain an exact amount of organic soil resources remaining.

## Utilization

Most organic soils occur in wetland environments, thus they provide wetland functions of protecting surface and ground water quantity and quality, preserving biological diversity, reducing surface water runoff, assisting in flood protection, and providing erosion and sedimentation control. Early humans avoided organic soils because they were low, wet, and harbored pests. Malaria was a major problem in some areas. Some people believed evil spirits resided in these areas. Individuals seeking refuge from slavery or the law frequently used them. About the turn of the eighteenth century, interest in organic soils (Figure 5) was stimulated by the need for fuel. Wood,

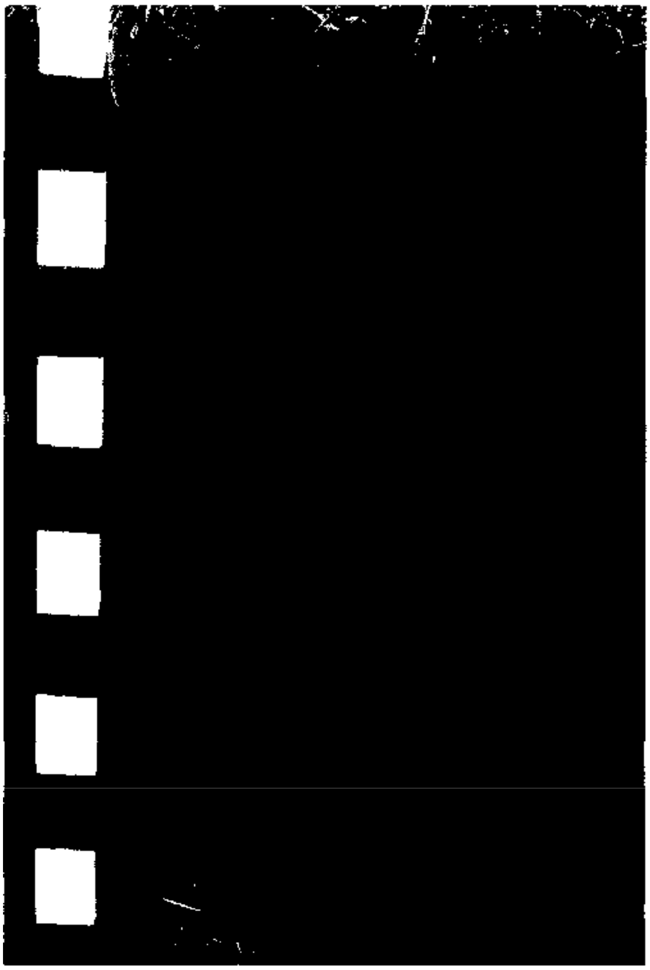


**Figure 5** Profile of organic soil in Finland.

**Figure 6** Use of organic soil for fuel.

the common fuel of that day was in short supply. Peat is still used for fuel (**Figure 6**) and for generation of electricity in some areas.

Utilization of organic soils is affected by several of their properties: soil reaction, cation exchange capacity, bulk density, water retention, water content, hydraulic conductivity, shear strength, and shrinkage. A profile description of a cultivated organic soil is given in **Table 1**. Selected physical and chemical properties of the same pedon are given in **Table 2**. The pH of organic deposits is largely a function of the nature of the water that flowed into the organic soil area and the vegetation. Organic soils in landscapes dominated by acid, sandy Spodosols are typically very acid and have a pH of less than 4.5. Also, organic soils that receive water only from precipitation, no groundwater flow, tend to be acidic. Coastal organic soils that received sulfur from seawater are usually very acid and may be acid sulfate soils. Organic soils in landscapes that receive subsurface flow that contains significant amounts of bases and/or bicarbonate are usually less acidic. Because organic soils are composed of primarily decomposed vegetative materials, composition of vegetation exerts a strong influence on soil reaction. Acid-loving plants, such as sphagnum moss (**Figure 7**), produce more acidic materials than base-loving plants. Very acid organic materials have low levels of microbial activity compared to less acidic materials. Less decomposed organic soil materials are more acidic than more decomposed materials. The availability of phosphorus and many micronutrients is dependent on pH. Acid organic soils are usually not limed above pH 5.5 in order to reduce the likelihood of causing micronutrient deficiencies in crops.

**Table 1** Profile description of a cultivated organic soil

| | |
|---|---|
| Oap | 0–23 cm. Black (N 2/0), black (10YR 2/1) rubbed; about 5% fibers, 1% rubbed; weak fine granular structure; friable; few fine roots; estimated mineral content 20%; abrupt smooth boundary |
| Oa1 | 23–48 cm. Black (N 2/1), very dark brown (10YR 2/2) rubbed; about 20% fibers, 4% rubbed; weak coarse blocky structure; friable; few fine roots; estimated mineral content 25%; clear smooth boundary |
| Oa2 | 48–76 cm. Very dark brown (10YR 2/2) broken face and rubbed; about 40% fibers, 10% rubbed; weak thick platy structure; friable; very few fine roots; estimated mineral content 15%; clear smooth boundary |
| Oa3 | 76–115 cm. Very dark brown (10YR 2/2) broken face and rubbed; about 25% fibers, 4% rubbed; weak thick platy structure; friable; estimated mineral content 15%; clear smooth boundary |
| Oa4 | 115–190 cm. Dark brown (7.5YR 3/2), very dark brown (10YR 2/2) rubbed; about 35% fibers, 7% rubbed; weak thick platy structure; friable; estimated mineral content 12%; clear smooth boundary |

The cation exchange capacity of organic soils results primarily from carboxyl radicals. Other functional groups that contribute are phenolic, enolic, and quinonic hydroxyls and heterocyclic nitrogen structures. Organic soils have high cation exchange capacities compared to mineral soil. The cation exchange capacity is pH-dependent and ranges from 10 to 20 mmol per 100 g at pH 3.5–4.0 to 100–200 mmoL per 100 g at pH 7.0.

One of the most outstanding characteristics of organic soils is their generally low bulk density. Bulk densities range from as little as 0.02 to 0.30 $\mathrm{g\,cm^{-3}}$. Some studies found that highly decomposed materials

**Table 2** Selected properties of a cultivated organic soil

| Horizon | Depth (cm) | Bulk density ($g\ cm^{-3}$) | Water content (%) | $CaCl_2$ pH | Organic C (%) | Fiber Unrubbed (%) | Fiber Rubbed (%) | Extract color[a] |
|---|---|---|---|---|---|---|---|---|
| Oap | 0–23 | 0.30 | 255 | 6.2 | 45.0 | – | 5 | 7.5YR 3/2 |
| Oa1 | 23–48 | 0.17 | 498 | 6.2 | 43.8 | 24 | 3 | 7.5YR 3/2 |
| Oa2 | 48–76 | 0.12 | 699 | 5.8 | 39.0 | 20 | 3 | 7.5YR 4/2 |
| Oa3 | 76–115 | 0.13 | 667 | 5.7 | 32.0 | 38 | 2 | 7.5YR 4/2 |
| Oa4 | 115–190 | 0.12 | 705 | 5.6 | 49.5 | 40 | 2 | 7.5YR 4/3 |

[a] Color of pyrophosphate extract.

**Figure 7** Sphagnum moss.

have a greater bulk density than less decomposed materials; however, a recent study found no difference in bulk density with difference in degree of decomposition. Bulk density of organic soils increases with increasing mineral content.

Organic soils are wet not only because they exist in saturated conditions, but also because they have a great affinity to take up and hold water. Water retention is related to degree of decomposition. Large pores in slightly decomposed organic soil materials drain easily at low suction, whereas small pores in more decomposed materials retain water at low suctions. Water contents of organic soils are related to degree of decomposition. Fibric soil materials retain from 8.5× to more than 30× on a weight basis, hemic soil materials hold between 4.5 and 8.5×, and sapric soil materials usually retain less than 4.5×. Water contents ranged from 100% (by volume) for undecomposed, near-surface sphagnum moss peat to less than 80% for deeper, more decomposed peat. Hydraulic conductivity of organic soils is also related to degree of decomposition of the organic materials because of differences in pore size distribution. Saturated hydraulic conductivity of sapric material ranges from 0.5 to $15\,cm\,h^{-1}$, that of hemic materials ranges from 1.5 to $15\,cm\,h^{-1}$, and that of fibric materials is greater than $15\,cm\,h^{-1}$. Weakly or undecomposed peats had rather rapid hydraulic conductivities, whereas more decomposed peats had slow conductivities. After drainage, the hydraulic conductivity of an organic soil decreases, due in part to greater decomposition and in part to the decrease of the buoyant effect of the groundwater.

Shear strength is a measure of a soil's internal resistance to stress. In mineral soils, shear strength decreases as water content approaches saturation. In organic soils, this relationship is not as clear. Draining a fibrous peat reduces the water content but increases the shear strength because of the 'fibrous interlock.' However, drainage also increases decomposition that destroys the 'fibrous interlock.' As decomposition increases, organic soils behave more like mineral soils with respect to shear strength.

Most organic soils undergo considerable shrinkage upon drying. Well-decomposed organic soil materials may shrink 70% or more depending on the mineral content of the material. The degree to which the shrinkage is irreversible is partially related to degree of decomposition. Organic soils are vast reservoirs of carbon. Carbon is stored during formation of these soils. During warmer periods small amounts of organic materials accumulate, whereas during cool, moist periods considerable amounts accumulate. Excavation of organic materials for various purposes (fuel, manufacture of peat products for agriculture and horticulture, medical applications) accelerates decomposition. Drainage leads to increases in decomposition of organic materials and emission of carbon dioxide. Methane and carbon dioxide stored in organic soils are released by drainage. Release of carbon dioxide as a result of these activities increases the concentration of greenhouse gases.

One of the early more intensive uses of organic soils was grazing, especially during the summer when water tables were naturally lower. Low bearing capacity and strength limited the usefulness of these soils. Artificial drainage not only lowered the water table but increased decomposition, thereby increasing bearing capacity and strength. Initially drainage was accomplished with open ditches and later with tile systems. Ditches have been constructed in some forests to increase productivity (**Figure 8**). Tile systems were more costly and required more intensive use of the land. Agronomic crops in high-latitude areas suffered frost damage in the spring after planting (**Figure 9**) and fall before maturity, thus the change from low-value crops fed to animals to high-value crops that tolerate frost or have a shorter growing season and are consumed by humans. Rice, carrots, onions, celery, potatoes, lettuce, mint, sugarcane, and

**Figure 8** Ditch installed in forest to lower water table and increase production.

**Figure 9** (see color plate 45) Frost damage of crop on organic soil.

cranberries are grown extensively on organic soils (**Figures 10** and **11**). Peat is also mined for use as a soil amendment in agriculture and horticulture. The low bulk density of organic soils makes them ideal for grass sod production (**Figure 12**) where large volumes of soil are moved long distances.

Utilization of organic soils requires careful management. Wise management of water is essential. In their natural state, organic soils are too wet. When they are drained, they become aerobic and organic materials decompose, causing rapid subsidence (**Figure 13**).

The most rapid subsidence occurs immediately after drainage. Subsidence in Hunts, England was 3.4 cm year$^{-1}$, Quebec, Canada 2.1 cm year$^{-1}$, Ontario, Canada 3.3 cm year$^{-1}$, Minnesota 5.1 cm year$^{-1}$, Indiana 3.0 cm year$^{-1}$, Florida 3.0 cm year$^{-1}$, and California 2.3 cm year$^{-1}$. An area in Michigan had to be retiled 28 years after it was established as a result of subsidence (about 3 cm year$^{-1}$). Organic soils have limited life expectancies. Rate of subsidence, degree of decomposition, and depth of organic materials determine the life of organic soils. Careful attention to wind erosion control and water table

**Figure 10** Carrots growing on organic soil.

**Figure 11** (see color plate 46) Organic soil used for cranberry production.

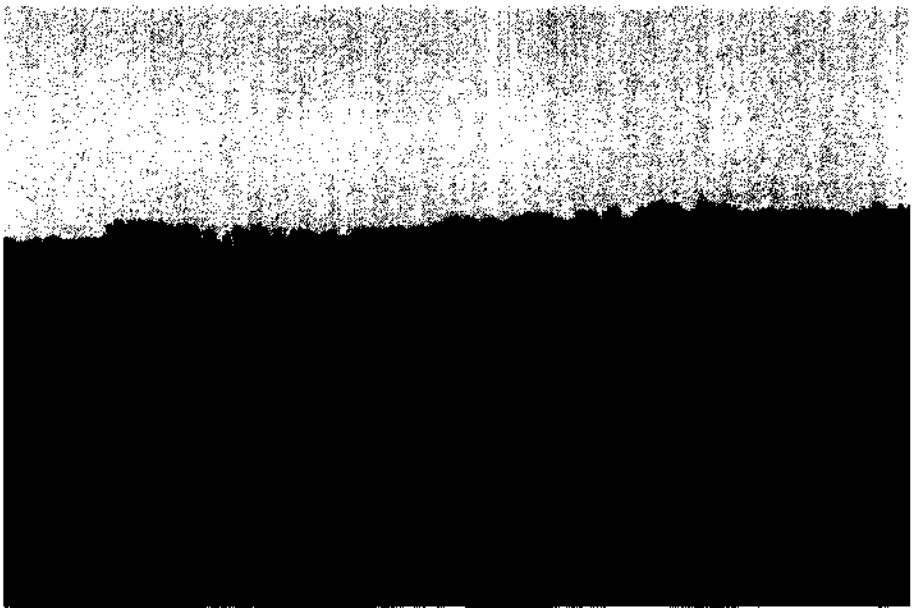


**Figure 12** Low bulk density makes organic soil ideal for sod production.

**Figure 13** In 1924 this 9-feet (3-m) post was placed on the underlying bedrock with the top of the post at the surface of the organic soil at Belle Glade, FL. This photograph, taken in 1972, shows that subsidence has been about 2.5 cm year $^{-1}$.

height can pay dividends in conservation of organic soils. To minimize subsidence water tables should be maintained as high as crop and field conditions will permit. Raising the water tables to at or near the soil surface (**Figure 14**) during the noncrop period can be effective in reducing subsidence and wind erosion. After drainage, organic soils are very susceptible to wind erosion because of the low bulk density. Relatively low wind velocities cause movement of the light-weight materials if unprotected. When wind erosion is very severe, organic soil material in the air reduces visibility, thereby causing highways to be closed.

Fire hazard increases after drainage of organic soils. These fires are difficult to control and may burn underground for several months. Some fires were accidental, caused by sparks from rail locomotives or farm tractors. Other fires were deliberately started to reduce excess acidity and/or increase nutrient levels. Some pasture and old hay lands were burned to destroy unwanted plant residues and brush. As a result of fires, undesirable materials such as sand, gravel, clay, and marl were exposed. These fields usually had marked variability in pH. Some burned areas had greater than pH 8 because of marl or other calcareous materials. Marl in the plow layer

**Figure 14** Raising water tables to at or near the organic soil surface will reduce subsidence and wind erosion.

**Figure 15** Low bearing capacity of organic soils causes cracks in roads that require patching.

causes P and micronutrients to be unavailable for plant growth. In some Michigan fields available soil P levels ranged from 2 to 300 $\mathrm{kg\,ha^{-1}}$. Levels below 10 $\mathrm{kg\,ha^{-1}}$ occurred in plow layers that contained marl, whereas the high levels occurred in plow layers that lacked marl. Fertilizer application in a field had been based on soil test analyses of the problem area.

The major engineering problem with the use of organic soils is compressibility. Compressibility is related to nature of the organic materials, degree of decomposition, water content, structure, and density. Compression and subsidence of organic soil materials continue for many years after loading. Construction of structures and roads on organic soils is usually difficult. Structures are usually constructed on piles that were driven into the underlying mineral soil to support foundations. Floating road designs have been used on large areas of organic soils. These roads frequently crack because of the low bearing capacity (**Figure 15**). If the area is not large, the organic materials can be removed and replaced with sand and gravel. Ultimately, structures are left standing above the ground level. Drainage accelerates decomposition and shrinkage of the organic materials, therefore subsidence of driveways, sidewalks, and/or lawns leaves structures perched above soil surfaces

**Figure 16** Subsidence of lawns and sidewalks in Michigan necessitated addition of steps built on pilings to enter the apartment building.

**Figure 17** Subsidence of lawn and driveway in New Orleans, Louisiana, requires capping of steps and decreases the accessibility of carport.

(Figures 16 and 17). Garages and carports may stand above the driveway and be inaccessible to cars. Many of these garages and carports have been converted into another room, since both the house and garage or carport were built on piles. Acidity of organic soils causes corrosion of concrete and metal materials. Organic soils have high porosity and high adsorptive capacity, therefore they are useful for treatment of wastewater. They have the potential to remove heavy metals and hydrocarbon. Mined peat is used for treatment of septic tank effluent before dispersing into soil.

All these activities have one thing in common – they bring stored carbon into atmospheric circulation.

## Classification

Organic soils are saturated and have 40 cm or more of organic soil materials according to Food and Agriculture Organization of the UN–Unesco Soil Map of the World, Revised Legend and Soil Taxonomy. If the organic materials consist of at least 75% moss or have a bulk density of less than 0.1 g cm$^{-3}$, the thickness must be at least 60 cm. The World Reference

**Table 3** Criteria used in the classification of organic soils

| *Food and Agriculture Organization of the UN Unesco* | *Soil Map of the World. Revised Legend* |
|---|---|
| Soil units | Presence of permafrost |
| | Presence of a sulfuric horizon or sulfidic materials |
| | Degree of saturation |
| | Degree of decomposition of organic materials |
| *World Reference Base* | |
| Soil units | Presence of permafrost |
| | Presence of a sulfidic or sulfuric horizon |
| | Presence of a salic horizon (highly soluble salts) |
| | Presence of a folic horizon (well-aerated, organic materials) |
| | Degree of decomposition of organic materials |
| *Soil Taxonomy* | |
| Suborder | Degree of saturation |
| | Degree of decomposition of organic materials |
| Great group | Soil temperature regime |
| | Special materials (sphagnum fibers, sulfuric horizon, sulfidic materials, humilluvic materials) |
| | Soil moisture regime |
| Subgroup | Presence of water |
| | Presence of lithic contact (bedrock) |
| | Thickness of organic materials |
| | Presence of mineral materials within organic materials |
| | Presence of sphagnum |
| | Presence of limnic layers |
| | Degree of decomposition of organic materials |
| | Electrical conductivity |

Base for soil resources broadened the requirement by decreasing the minimum depth to 10 cm. The criteria used in the classification of organic soils differ in the three soil classification systems (Table 3). Organic soil materials that are not saturated must contain at least 20% organic C.

## Further Reading

Aandahl AR, Buol SW, Hill DE, and Bailey HH (eds) (1974) *Histosols: Their Characteristics, Classification and Use.* Soil Science Society of America special publication 6. Madison, WI: Soil Science Society of America.

Everett KR (1983) Histosols. In: Wilding LP, Smeck NE, and Hall GF (eds) *Pedogenesis and Soil Taxonomy. II. The Soil Orders*, pp. 1–53. Amsterdam: Elsevier.

FAO (1990) FAO-UNESCO *Soil Map of the World, Revised Legend.* World Soil Resources Report 60. Rome: Food and Agriculture Organization.

FAO-ISRIC-ISSS (1994) *World Reference Base for Soil Resources.* World Soil Resources Report 84. Rome: Food and Agriculture Organization.

Farnham RS and Finney HR (1965) Classification and properties of organic soils. *Advances in Agronomy* 7: 115–162.

Gore AJP (ed.) (1983) *Mires: Swamp, Bog, Fen and Moor.* Amsterdam: Elsevier.

Heinselman ML (1970) Landscape evolution, peatland types and the environment in the Lake Agassiz Peatlands Natural Area, Minnesota. *Ecology Monograph* 40: 235–261.

Lucas RE (1982) *Organic Soils (Histosols). Formation, Distribution, Physical and Chemical Properties and Management for Crop Production.* MI State University Agricultural Experimental State Research Report 435. East Lansing, MI: Michigan State University.

Moore PD and Bellamy DJ (1974) *Peatlands.* New York: Springer Verlag.

Rabenhorst MC and Swanson D (2000) Histosols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-183–E-209. Boca Raton, FL: CRC Press.

Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys.* USDA Agricultural Handbook 436. Washington, DC: US Government Printing Office.

# OVERLAND FLOW

**T S Steenhuis, L Agnew, P Gérard-Marchant, and M T Walter**, Cornell University, Ithaca, NY, USA

## Introduction

For many years, it was believed that the occurrence of surface runoff is primarily controlled by the infiltration characteristics of the ground. Specifically, runoff is generated whenever rainfall or irrigation water is applied to the ground at a higher rate than the soil's infiltration capacity. Robert E. Horton was responsible for some of the early foundational work behind this concept and hence the process is ubiquitously referred to as 'Hortonian runoff.' The synonymous term 'infiltration-excess overland flow' is also widely used, again emphasizing that surface runoff occurs when precipitation exceeds the infiltration capacity of the soil. This process is very important in many areas of the country: arid areas (where significant soil crusting and/or surface sealing occurs during rain events), in irrigated fields, in urban areas, and more generally during storms with very high rainfall intensities.

However, the Hortonian runoff concept does not meaningfully explain storm runoff in many regions where the infiltration capacity of the ground is typically much greater than the average rainfall intensities. Many researchers have found that hydraulic conductivities published in soil surveys, usually based on disturbed soil samples, underestimate, by a factor of 10 or more, the conductivity in the field for vegetated soils which contain preferential flow paths in the form of worm channels and root passages. Consequently, only the most intense summer storms can produce runoff under the Hortonian runoff scenario. This is in direct contrast to observations of stream response to almost all rain, even small storms, during the set periods of the year, which means that there must be some other mechanism generating runoff.

In many regions, runoff is most commonly generated on relatively small portions of the landscape that are susceptible to becoming completely saturated. Once the soils in these areas saturate to the surface, any additional rainfall (irrespective of intensity) becomes overland flow. This process is termed 'saturation-excess overland flow.' Therefore, the propensity of an area to produce runoff is largely independent of rainfall intensity. Instead, the total rainfall amount and landscape factors such as soil depth (i.e., available water storage capacity), upland watershed area, and local topography are the important factors determining whether or not a particular area in a watershed will generate runoff. Moreover, as rainfall continues, the saturated area grows in extent, increasing the area generating runoff (hence the term 'variable source area,' VSA). This is in contrast to Hortonian or infiltration-excess runoff generation, which depends on soil type (infiltration rate versus rain intensity) and is independent of position in the landscape, i.e., with the runoff generating area is always the same. Examples of regions in the USA where VSA hydrology is significant include the Northeast and the Pacific Northwest, as well as forested mountain areas. It also occurs throughout the world in many similar areas.

## Saturation-Excess Overland Flow

Saturation-excess overland flow comes from two distinguishable sources. In one case, rain falling on already saturated soil has no option but to run off – this case is termed 'direct precipitation on saturated areas' (DPSA). The other source, termed 'return flow,' occurs if the rate of interflow entering a saturated area from upslope exceeds the capacity for interflow to leave the area by flowing downhill through the soil. The excess interflow, thus, 'returns' to the surface as runoff, hence the term. Whereas DPSA runoff only occurs during and just after a rainfall event, return flow seepage can continue as long as an interflow excess exists.

Areas prone to saturation either have a high groundwater table or hard pan (fragipan) at shallow depth. Interflow results in the formation of saturated areas at the bottom of slopes, usually in concave areas, or it quickly resurfaces in seeps and ditches. Some hydrological processes observed in upstate New York, primarily the Catskill Mountains region, can further serve to elucidate mechanisms resulting in saturation. Soils in the Catskills are generally permeable (relative to rainfall intensity) and underlain by a shallow, low-permeable, restrictive layer, typically bedrock or fragipan. Rainwater easily permeates the soil and, by-and-large, runs laterally as interflow on top of the restrictive layer downslope. Many researchers have observed evidence of the accumulation of interflow water at the bottom of slopes in the Catskills during wet periods in the form of increased moisture content at hill bottoms relative to the steep parts of hills. Some common locations where saturation occurs are areas where the soil above the restricting layer is shallow, in places where the downhill topographic slope decreases such as the toe-slope of a hill, or in topographically converging areas (**Figure 1**). All three incidences are locations in the landscape where the Darcy flow capacity, or interflow capacity, is reduced either by a decrease in hydraulic transmissivity or hydraulic gradient. When interflow capacity is sufficiently restricted, the soil will saturate. During periods of enhanced rainfall, interflow will be higher and often expand the extent of saturation around saturation-prone areas; conversely, dry periods will decrease interflow and extent of saturation. This is illustrated for wet and dry seasons in **Figure 2**.

VSA hydrology is an extension of the saturation-excess concept, recognizing that the extent of saturated areas in a watershed will expand and contract, i.e., vary temporally. The variation in the extent of saturated areas has been studied over a range of temporal scales, ranging from storm duration, hours and days, to seasons (**Figure 2**). Although the original, underlying concept for VSA hydrology was developed by the US Forest Service, the term 'variable source area' is often attributed to Hewlett and Hibbert. As stated above, VSA hydrology is applicable to many landscapes throughout humid regions that have undulating topography.

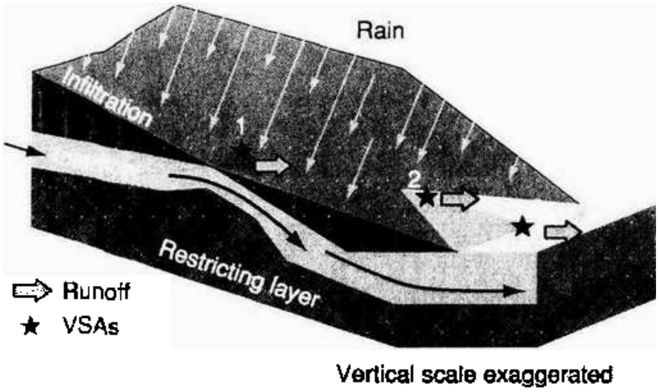


**Figure 1** Incidence of saturation-excess hydrology: 1, shallow soil; 2, convergence area; and 3, decreasing downhill slope. VSA, variable source area.

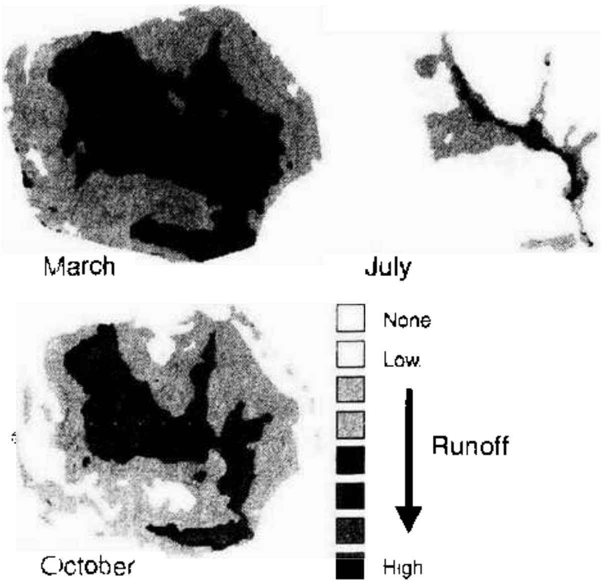


**Figure 2** (see color plate 47) Seasonal changes of saturated areas (variable source areas) in a watershed in New York state. Color denotes areas more prone to saturation.

The VSA hydrology concept, thus, goes beyond conceptualizing the water–soil interface at or below the soil surface and making the simple assumption that water either runs off or infiltrates. The VSA hydrology concept includes the soil reservoir in a spatial landscape context to describe locations where the soil reservoir becomes saturated and begins to generate runoff. The soil reservoir methodology utilizes the bedrock, impermeable soil layers (relative to above layers), and/or the depth to the water table as the underlying hydrologic boundary instead of the soil surface. Thus, both hydrologic and soil water (i.e., porous media) concepts are combined to evaluate potential runoff areas in the landscape.

## Calculation Method

TOPMODEL is used most often to predict the saturated-excess overland flow. It is based on calculating a groundwater table height that, when it intercepts the soil surface, will result in runoff. In TOPMODEL, the assumption is made that the watershed is underlain by a groundwater table. This is not realistic for many mountainous areas. However, since most of the saturated areas occur near streams, this limitation is not severe as a groundwater table is present in these areas. How the water gets to the groundwater is often inconsequential. An alternative method to find the saturated areas is the Soil Moisture Routing (SMR) model, a physically based, fully distributed, geographic information system (GIS) integrated code specifically designed for soils underlain by a hardpan. This model solves the water balance for cells in a grid that covers a watershed. What makes it different from most other models is that the spatial location of the grid cells is preserved in the landscape and water is routed by interflow from one cell to the next, resulting in the simulation of saturated areas within the landscape. This model, despite its simplicity, predicts the saturated areas in the landscape amazingly well.

TOPMODEL and SMR are admittedly more complicated than the popular Soil Conservation Service 'Curve Number' method (SCS-CN), which was developed in the 1950s and 1960s by the USDA Soil Conservation Service (currently the Natural Resource Conservation Service). The traditional SCS-CN method is a rainfall-runoff model originally developed for predicting flood-flow volumes from ungauged watersheds for hydraulic engineering design. Despite the limited scope of intended application and several identified theoretical problems with the traditional SCS-CN method, it is ubiquitously used in water quality models to generate the fraction of rainfall that runs off. The original model was justified by Victor Mockus, to whom the traditional SCS-CN method is largely attributed, "on grounds that it produces rainfall-runoff curves of a type found on natural watersheds." Subsequently, hydrologists have shown that the basis of the method can be described in ways that are nominally consistent with both infiltration-excess theory and saturation-excess or VSA hydrology.

## Water Quality Impacts of Variable Source Area Hydrology

VSA hydrology has been shown to have significant consequences on surfacewater quality, having direct impacts on specific areas such as animal waste management, drinking water and human health, nutrient and pesticide management, pollution assessment and prevention, and watershed management. It is important that the development and implementation of the next generation of land-management practices for improving or preserving surface water quality be consistent with our most current hydrological understanding. Interestingly, many current practices aimed at reducing contaminant transport are based on Hortonian concepts. However, the effectiveness of these practices is diminished where Hortonian processes are not governing runoff generation.

Wherever VSA hydrology is a dominant process, there will be regions within a watershed that are more susceptible to producing runoff and delivering it to surfacewater bodies than other regions. These areas can be considered hydrologically sensitive areas (HSA). Recognizing the existence of HSAs allows watershed-scale water quality efforts to be focused on those areas where HSAs coincide with land uses that potentially contribute pollutants. The intersection of an HSA and a pollutant loading area is referred to as the 'critical management zone.' The most obvious best-management practice for this area would be to limit or prohibit potentially polluting activity from this region. A contrasting approach is finding methods and means of eliminating the hydrological sensitivity of the critical zones, although recent research suggests attempts to remove hydrologic sensitivity through drainage practices may simply reroute pollutants from overland flow to subsurface flow, with little reduction in concentration. In either case, the HSAs warrant primary attention when trying to preserve or improve water quality.

**Figure 3** Manure deposition in exactly the wrong place from a hydrologically sensitive-area standpoint (courtesy of Glenn Warner).

One example of where potential management practices derived from a recognition that VSA hydrology is in direct conflict with currently mandated management practices is whether dairy operators in the New York city watersheds should spread manure on steep slopes or in flat areas. The current dogma is to avoid spreading on steep slopes and to maximize spreading in flat areas. Under Hortonian flow, steep areas might arguably produce the most rapidly moving runoff and, therefore, the greatest potential for erosion and transport of manure. However, VSA hydrology is the dominant process in these watersheds and the steep slopes infiltrate essentially all rainwater and drain very rapidly, resulting in almost no substantial runoff. Conversely, the flat areas, especially at the base of hillslopes, are especially prone to saturation and, thus, prone to runoff generation. **Figure 3** shows manure deposition in, from a HSA standpoint, exactly the wrong place. A manure-spreading policy that is more consistent with the recognized hydrology would promote spreading high in the watershed and minimize spreading on low, flat areas susceptible to runoff generation.

## Further Reading

Beven KJ (2001) *Rainfall-Runoff Modelling: The Primer.* Chichester, UK: Wiley.

Dingman SL (2002) *Physical Hydrology*, 2nd edn. Englewood Cliffs, NJ: Prentice Hall.

Dunne T (1970) *Runoff Production in Humid Areas.* US Department of Agriculture Publication ARS-41-160. Washington, DC: USDA.

Dunne T and Leopold LB (1978) *Water in Environmental Planning*. New York: W.H. Freeman.

Frankenberger JR, Brooks ES, Walter MT, Walter MF, and Steenhuis TS (1999) A GIS-based variable source area model. *Hydrological Proceedings* 13: 804–822.

Hewlett JD and Hibbert AR (1967) Factors affecting the response of small watersheds to precipitation in humid regions. In: Sopper WE and Lull HW (eds) *Forest Hydrology*, pp. 275–290. Oxford, UK: Pergamon Press.

Horton RE (1933) The role of infiltration in the hydrologic cycle. *Transactions of the American Geophysical Union* 14: 446–460.

Kirkby MJ (1978) *Hillslope Hydrology*. New York: Wiley.

Kuo WL, Steenhuis TS, McCulloch CE, Mohler CL, Weinstein DA, DeGloria SD, and Swaney DP (1999) Effect of grid size on runoff and soil moisture for a variable-source-area hydrology model. *Water Resources Research* 35: 3419–3428.

# OXIDATION–REDUCTION OF CONTAMINANTS

**C J Matocha**, University of Kentucky,
Lexington, KY, USA

## Introduction

In the soil environment, inorganic and organic contaminants can be transformed by reduction and oxidation (redox) processes that can occur chemically, photochemically, or biologically. Historically, microbial oxidation of organic matter in soils and sediments ranks the sequence of oxidants as molecular oxygen ($O_2$) reduction, nitrate ($NO_3^-$) reduction, manganese dioxide ($MnO_2$) reduction, iron hydroxide ($Fe(OH)_3$) reduction, sulfate ($SO_4^{2-}$) reduction, and methane ($CH_4$) production. This reduction sequence is based on the oxidant yielding the most free energy change per mole of organic carbon oxidized and proceeds until all oxidants or all oxidizable organic matter are consumed in well-defined depth zones. It is well accepted that, in certain field situations, this thermodynamic reduction sequence is an oversimplification and does not coincide with the kinetic reduction sequence, because it provides no information about reaction rates and other potentially competing abiotic pathways.

In abiotic reaction pathways, the role of catalysis is important. Naturally occurring solid Fe and Mn oxides have been shown abiotically to catalyze transformations of organic and inorganic pollutants, which themselves could serve as electron donors and acceptors. Most reactions involving these metal oxides are heterogeneous, with complex, multistep processes such as adsorption, precipitation, electron transfer, and diffusion. The focus here is the role of reactive soil Fe and Mn species in the redox transformations of inorganic and organic contaminants.

## Redox Status of Reactive Fe and Mn Species

Selected half-reactions and standard reduction potentials of important, solid-phase Mn(III,IV) and Fe(III) oxides in soils are compared below with other important redox half-reactions:

$$\frac{1}{2}\,\delta\text{-}MnO_x(s) + 2H^+ + e^- \rightarrow \frac{1}{2}\,Mn^{2+} + H_2O \qquad E_h^0 = 1.29\,V \qquad [1]$$

$$Fe(OH)_3(s) + 3H^+ + e^- \rightarrow Fe^{2+} + 3H_2O \qquad E_h^0 = 1.057\,V \qquad [2]$$

$$\frac{1}{4}\,O_2(g) + H^+ + e^- \rightarrow \frac{1}{2}\,H_2O \qquad E_h^0 = 1.229\,V \qquad [3]$$

$$\frac{1}{8}\,SO_4^{2-} + \frac{5}{4}\,H^+ + e^- \rightarrow \frac{1}{8}\,H_2S + \frac{1}{2}\,H_2O \qquad E_h^0 = 0.303\,V \qquad [4]$$

Birnessite ($\delta$-$MnO_x(s)$) is the most commonly identified Mn(III,IV) oxide in soils, controlling solubility of plant-available Mn(II). Ferrihydrite, an amorphous Fe hydroxide ($Fe(OH)_3(s)$), is depicted as controlling Fe(II) solubility. Half-reactions are useful to represent the stoichiometry and thermodynamics of redox reactions. The standard state potential ($E_h^0$) in volts fixes the activity of reactants, products, and $H^+$ at unity by definition. The Nernst equation can be written for each half-reaction and $E_h$–pH diagrams constructed under more realistic environmental conditions (($Mn^{2+}$), ($Fe^{2+}$), and ($SO_4^{2-}$) $= 10^{-5}$ mol l$^{-1}$ and partial pressures of $O_2$ and $H_2S$ of 0.02 and 0.0001 MPa):

1. $\delta\text{-MnO}_x(s)/Mn^{2+}$: $E_h = 1.378 - 0.118\,pH$
2. $Fe(OH)_3(s)/Fe^{2+}$: $E_h = 1.352 - 0.177\,pH$
3. $O_2(g)/H_2O$: $E_h = 1.219 - 0.059\,pH$
4. $SO_4^{2-}/H_2S$: $E_h = 0.362 - 0.7375\,pH$

When summing two half-reactions, it is assumed that they are in equilibrium with one another (coupled). As long as the calculated $E_h$ for the oxidant half-reaction is higher than that of the reductant half-reaction, the forward reaction is favorable. The free energy change, $\Delta G$, of a redox reaction, can be assessed by knowledge of the standard reduction potentials of the half-reactions involved. $E_h$ is directly related to $\Delta G$ of the system by:

$$\Delta G = -nFE_h \qquad [5]$$

where $n$ is the number of electrons involved in the redox reaction and $F$ is the Faraday constant ($F = 96\,500\,C\,mol^{-1}$). Redox reactions where $E_h > 0$ would give a negative $\Delta G$ and be thermodynamically favorable. Over a wide range of pH values, including those often encountered in soil environments (pH 4–8), oxidant strength from a thermodynamic point of view decreases in the order $O_2 > \delta\text{-MnO}_x > Fe(OH)_3 > SO_4^{2-}$ (**Figure 1**). Several lines of field evidence have shown that relying solely on thermodynamics to predict oxidant consumption can be misleading, and kinetic and catalytic factors need to be considered.

Sulfate reduction preceding Fe(III) reduction in lake pore-water sediments has been reported (**Table 1**). Overlap of these reduction sequences was ascribed to a wide range of Fe(III) oxide stabilities, giving rise to Fe(III) reduction rates that vary considerably and to the type of oxidizable organic C. The most unstable, poorly crystalline Fe oxides such as ferrihydrite are reduced preferentially to well-crystallized phases such as goethite ($\alpha$-FeOOH). This situation is further complicated in soils containing Fe-rich phyllosilicate minerals that can undergo reduction and oxidation cycles.

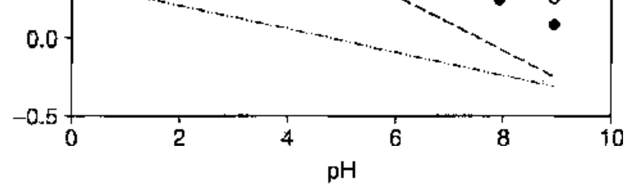


**Figure 1** Relationship of redox potential ($E_h$) to pH for several important redox couples in soils compared with representative inorganic and organic (*ortho*-quinone (O-qu.)/catechol (cat.) couple) contaminants. Dissolved Mn, Fe, S, Cr, and organic species were assigned an activity of $10^{-5}\,mol\,l^{-1}$, and dissolved oxygen and sulfide were given values of 0.02 and 0.0001 MPa.

**Table 1** Pore-water chemistry with depth of Lake Greifen. Note the occurrence of sulfate reduction precedes that of Fe(III) reduction

| *Depth (cm)* | *$SO_4^{2}$ (μM)* | *[$H_2S$] (μM)* | *Fe(II) (μM)* |
|---|---|---|---|
| 7 | 170 | 0 | 0 |
| 1 | 145 | 0 | 0 |
| −1 | 67 | 3.0 | 0 |
| −2 | 10 | 10.0 | 0 |
| −5 | 9 | 9.0 | 12.0 |
| −10 | 5 | 4.0 | 100 |
| −25 | 10 | 2.0 | 300 |
| −30 | 10 | 2.0 | 350 |

Adapted from Wersin P, Hohener P, Giovanoli R, and Stumm W (1991) Early diagenetic influences on iron transformations in a freshwater lake sediment. *Chemical Geology* 90: 233–252.

Reduced products from the primary redox reactions can participate in abiotic secondary reactions, affecting oxidant availability and organic matter degradation pathways. Consumption of dissolved $O_2$ is dependent on microbial activity and availability of suitable electron donors such as dissolved Fe(II) and Mn(II). Dissolved Fe(II) and Mn(II) can diffuse upwards from anaerobic environments into oxic zones. The thermodynamic driving force for capture of downward-diffusing $O_2$, which is in ample supply from the atmosphere, by dissolved Fe(II) and Mn(II) is favorable over a wide soil pH range. Yet at pH >6, Fe(II) oxygenation is rapid, while Mn(II) oxygenation occurs only after years and coexistence of dissolved Mn(II) and $O_2$ is not uncommon. Surface catalysis by hydrous oxides and microorganisms increases Mn(II) oxygenation rates, thus, only heterogeneous oxygenation of Mn(II) is included in reactive transport models because of the slow kinetics of homogeneous Mn(II) oxygenation. Depending on the environmental setting, as much as 50% of the dissolved $O_2$ can be consumed by these rate-limited secondary reactions and the remainder by microbial respiration.

## Kinetics and Mechanism

The goal of collecting fundamental kinetic data is to derive an elementary rate expression with a rate constant and to propose a mechanism. The usefulness of the kinetic approach is that it expresses the rate as changes in reactant and product concentration over time and empirically determines the reaction orders. Consider the following simple forward reaction:

$$A \xrightarrow{k_1} P \quad \text{rate} = -\mathrm{d}[A]\mathrm{d}t^{-1} = \mathrm{d}[P]\mathrm{d}t^{-1} = k_1[A]^a \qquad [6]$$

where $a$ is the power dependency or reaction order and $k_1$ is the rate constant, assuming the reaction stoichiometry is known and constant over the time period investigated. The value of $k_1$ should depend only on temperature, thus it should be specified, and be time-invariant. The rate expression is defined once the value of $a$ is determined. If unity, then the reaction is first-order in $A$ and first-order overall. Fractional-order rate expressions are not uncommon, particularly in multistep processes such as heterogeneous redox reactions. The rate, $-\mathrm{d}[A]\mathrm{d}t^{-1}$, is usually given in terms of moles per liter per second, and the units of $k_1$ for a first-order reaction are per second, per minute, or per hour. This information is used in reactive transport models that attempt to predict spatial and temporal distributions of redox-active species. It is important to realize that the empirical rate expressions may not always obey the overall reaction stoichiometry. Identification of the rate-limiting step, represented by the rate expression, is important because it represents the slowest step, thus any manipulation of it can either speed up (for instance adding a catalyst) or slow down the overall reaction.

For redox reactions to occur, the oxidant and reductant must encounter one another in order to allow the reaction to proceed. This is the case for both homogeneous and heterogeneous reactions. The general mechanism for heterogeneous redox reactions includes the following steps: (1) precursor-surface complex formation between oxidant and reductant, (2) electron transfer, and (3) product release. Both electronic and steric effects will influence the rate and extent of precursor complex formation, especially when considering sorption behavior at a surface. The precursor complex can involve a direct, inner-sphere bond between oxidant and reductant or an outer-sphere bond, where no bond breaking or making occurs and the transition state is unaffected. The rate-limiting step for outer-sphere electron transfer mechanisms is the electron-transfer step, while either precursor complexation or electron transfer can be rate-limiting in inner-sphere mechanisms. Bridging ligands with filled sigma ($\sigma$) and pi ($\pi$) orbitals such as oxo surface groups ($O^{2-}$, $OH^-$, and $H_2O$) on minerals can facilitate electron transfer for inner sphere mechanisms because the lone pairs of electrons function as 'bridges.' If the overlap of molecular orbitals for oxidant and reductant is perfect, then inner-sphere electron transfer mechanisms do not require a bridging ligand. Interfacial characteristics can also determine electron transfer rates, an additional consideration for heterogeneous reactions.

Rates of most redox reactions increase with temperature, and rate constants ($k_{\mathrm{obs}}$) can be described by the Arrhenius equation to determine activation energies ($E_a$):

$$k_{\mathrm{obs}} = A_f \exp^{-E_a/RT} \qquad [7]$$

where $A_f$ is the frequency factor, $R$ is the gas constant, and $T$ is absolute temperature. Low $E_a$ values ($<42\,\mathrm{kJ\,mol^{-1}}$) indicate that the rate-limiting step is a diffusion-controlled process, while high $E_a$ values represent surface chemical control of the rate-limiting step. Conducting kinetic experiments over a temperature range of 4–40°C (277–313 K) can simulate conditions in aquifers and soil environments.

Thermodynamic considerations are an important first step in the appraisal of whether a given inorganic or organic contaminant will undergo redox reactions in a soil environment. Delineation of redox zones by analysis of soil solution and pore water for major redox-active species is also helpful in predicting the fate of the contaminant, particularly if rate expressions have been determined already for the specified redox couple.

## Inorganic Contaminants

### Manganese as an Oxidant

The importance of Mn(III,IV) (hydr)oxides environmentally lies in their ability to adsorb and in some cases oxidize certain trace metals. Chromium (Cr) is a well-known contaminant, and typical oxidation states are Cr(III) and Cr(VI). Hexavalent Cr is mobile in soils because it is an oxyanion and repelled by negatively charged soil particles. In contrast, $Cr^{3+}$ is more strongly adsorbed and forms relatively insoluble oxide and hydroxide compounds. Chromium(III) and Cr(VI) speciation in solution are functions of pH, but will not be accounted for here. The half-reaction and standard reduction potential for the Cr(VI)-Cr(III) couple can be written as:

$$CrO_4^{2-} + 8H^+ + 3e^- \rightarrow Cr^{3+} + 4H_2O$$
$$E_h^0 = 1.51\,\mathrm{V} \qquad [8]$$

where $E_h = 1.51–0.1573\,\mathrm{pH}$.

Oxidation of Cr(III) to Cr(VI) by $O_2$ is more energetically favorable than by Mn(III,IV) (hydr)oxides or Fe oxides (**Figure 1**). However, rates of

Cr(III) oxidation by $O_2$ are slow over a wide pH range. Soil Mn oxides and laboratory-prepared Mn oxides can directly oxidize Cr(III) at faster rates than $O_2$ or Fe oxides, according to the following overall reaction:

$$Cr^{3+} + 1.5MnO_2 + H_2O \rightarrow CrO_4^{2-} + 1.5Mn(II) + 2H^+ \quad [9]$$

Typical rate equations and $E_a$ values (more than $42\,kJ\,mol^{-1}$) indicate that the slowest step is either precursor-surface complex formation or electron transfer. Temporal variability in Cr(III) oxidation rates is observed as initial Cr(III) concentration and pH increases and is attributed to surface precipitation of $Cr(OH)_3(s)$ that inhibits the reaction. In some cases, formation of an intermediate Mn(III) oxyhydroxide phase is reported to retard the reaction also. These findings are significant because Cr(VI) is more mobile and toxic than Cr(III).

The coupling of Mn reduction with Cr(III) oxidation to Cr(VI) can be explained by frontier molecular orbital theory. Oxides and hydroxides of Mn(III) and Mn(IV) are kinetically better oxidants than $O_2$, because the lowest unoccupied molecular orbitals (LUMO) for Mn(III) and Mn(IV) are the $e_g$ orbitals, which are of sigma ($\sigma^*$) symmetry and available to accept electrons from $\sigma$- and $\pi$-donor contaminant metals (Figure 2). For Mn(IV), with a $d^3(t_{2g}^3\ e_g^0)$ electronic configuration, both $e_g$ orbitals are vacant and a concerted two-electron transfer can occur. In contrast, the LUMO for $O_2$ is of $\pi^*$ symmetry ($\pi_{2px}^{1*}$ $\pi_{2py}^{1*}$) and, being singly occupied, can only accept one electron at a time. Oxo groups present on the Mn oxide surface serve as bridging ligands, because they have filled orbitals of $\pi$ and $\sigma$ symmetry to make the Cr(III)-to-Mn(IV) electron transfer facile.

Other examples of Mn oxide-catalyzed oxidations of trace metal contaminants include oxidation of Co(II) to Co(III), As(III) to As(V), and Se(IV) to Se(VI). In all of these reactions, there is a rate dependence on reductant and oxidant concentration. The rates in oxygenated systems are negligible without the presence of the solid-phase Mn oxide.

### Fe(II) as a Reductant

Rate studies of oxidation of Fe(II) in solution at pH >5 shows a second-order dependence on $OH^-$ concentration and first-order dependence on $O_2$ concentration:

$$-d[Fe(II)]/dt = k[Fe(II)][OH^-]^2[O_2] \quad [10]$$

because dissolved $OH^-$ donates electron density to Fe(II) through the $\sigma$ and $\pi$ systems and enhances the metal's reducing power. Thus, $O_2$ reacts predominantly with the hydrolyzed species $FeOH^+(aq)$ and $Fe(OH)_2$ (aq) as the pH increases. Fe(II) coordinated with $\sigma$- and $\pi$-donating ligands, with oxygen as the ligating atom, such as oxo groups in minerals and dissolved organic matter, lowers the aqueous Fe(III)–Fe(II) reduction potential, making coordinated Fe(II) a better reductant than dissolved Fe(II). Naturally occurring, solid-phase Fe(II) minerals such as siderite ($FeCO_3(s)$), pyrite ($FeS_2(s)$), and sulfate green rust ($[Fe_4^{II}\ Fe_2^{III}\ (OH)_{12}]\ [SO_4 \bullet 3H_2O)$ are stronger reductants than aqueous Fe(II) also.

During reductive dissolution of Fe(III) (oxy)hydroxides or Fe-bearing phyllosilicates in in soils and sediments by either microorganisms or abiotic reductants, Fe(II) is released to solution. After initial release, one of the processes that dissolved Fe(II) can undergo is adsorption back on to the surface. Adsorbed Fe(II) is significant in anaerobic soil environments for two reasons. First, in pure mineral systems, dissolved Fe(II) resulting from microbial reduction of goethite has been shown to adsorb to remaining goethite surface sites and limit the extent of reduction. In Fe(III) oxide-rich subsoils with mixed mineralogy, dissolved biogenic Fe(II) adsorbs strongly to kaolinite to limit bioreduction. The second reason is that Fe(II) adsorbed as an inner-sphere surface complex is an effective reductant and can catalyze oxygenation. The conceptual model is shown in Figure 3, where >M–O represents a surface metal (Fe(III)) oxide or phyllosilicate mineral functional group. The oxide or hydroxide group behaves as a bridging ligand and 'pumps' electron density from adsorbed Fe(II) to the $O_2$. Surface OH groups on metal oxides and phyllosilicate minerals can catalyze electron transfer from Fe(II) to $O_2$ analogous to

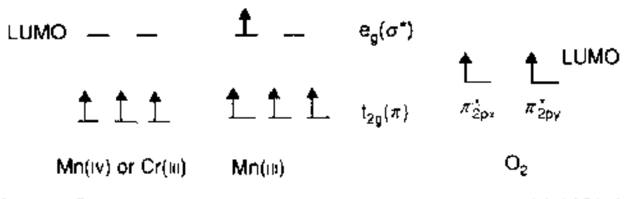


**Figure 2** Lowest unoccupied molecular orbitals (LUMO) for Mn(III) and Mn(IV) are shown as $e_g$ orbitals with $\sigma^*$ symmetry. For $O_2$, the LUMO are antibonding $\pi^*$ molecular orbitals resulting from overlap of 2px and 2py atomic orbitals.

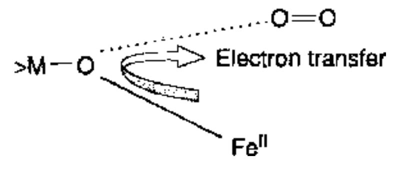


**Figure 3** Conceptual model of adsorbed Fe(II) as an inner-sphere surface complex that enhances oxygenation.

dissolved $OH^-$. Therefore, adsorbed Fe(II) bound to surface OH groups in an inner-sphere surface complex is a better reductant than dissolved Fe(II). A kinetic estimate for the redox potential of adsorbed Fe(II) on goethite is $E_h^0 - 0.36$ V, similar in magnitude to $FeOH^+$ ($E_h^0 - 0.34$ V). Characterization of the reducing power of adsorbed Fe(II) on different mineral surfaces remains a challenge. Adsorbed Fe(II) species as reductants are important to consider in any environmental setting because, at the onset of anaerobic conditions, solid Fe(III) oxides are the most abundant oxidants, in concentrations approximately 60-fold greater than solid Mn oxides.

Investigators have recently shown that adsorbed Fe(II) coordinated to surface hydroxyl functional groups on metal oxides can catalyze the reduction of Cr(VI), nitrite ($NO_2^-$), uranium(U(VI)), and technetium(Tc(VI)). Reduction of toxic Cr(VI) by reduced Fe(II) species is beneficial because the solubility of Cr is lowered by formation of sparingly soluble Cr(III) precipitates. This multistep process is rapid in the presence of dissolved Fe(II), structural Fe(II), and adsorbed Fe(II). Rate expressions are typically second-order:

$$-d[\mathrm{Cr(VI)}]/dt = k_{\mathrm{obs}}[\mathrm{Fe(II)}]_x[\mathrm{Cr(VI)}] \qquad [11]$$

where $k_{\mathrm{obs}}$ is the second-order rate coefficient and $[\mathrm{Fe(II)}]_x$ represents dissolved, adsorbed, or structural Fe(II). The values of $k_{\mathrm{obs}}$ for dissolved, adsorbed, and structural Fe(II) species can be compared qualitatively. Hydrous oxide surfaces such as goethite adsorb Fe(II) in an inner-sphere complex that is very reactive toward reduction of dissolved Cr(VI), with $k_{\mathrm{obs}}$ values ranging from 6.6 to 72 l $\mathrm{mol}^{-1}\,\mathrm{s}^{-1}$ (**Table 2**). Dissolved Fe(II) and solid-phase Fe(II) minerals such as green rust are also effective reductants of Cr(VI) (**Table 2**) and minimize contamination of soil and water supplies.

Nitrite reduction to $N_2O$ by Fe(II) adsorbed to lepidocrocite (γ-FeOOH) is rapid, and the maximum rate of 10 μmol $h^{-1}$ occurs at pH 8.5. The postulated mechanism Fe(II) adsorbed as a bidentate inner-sphere complex with surface oxo groups provides an ideal environment for $NO_2^-$ adsorption and subsequent N–N double-bond formation, similar to the function of the Fe metal center in nitrite reductase enzyme. No reduction of $NO_2^-$ occurs in the absence of lepidocrocite or in the presence of dissolved Fe(II). The $N_2O$ produced may take part in global warming and atmospheric ozone destruction, because it is a more potent greenhouse gas than $CO_2$ and represents an irreversible loss of N from soil.

**Table 2** Chromate (Cr(VI)) reduction rates based on observed second-order rate coefficients ($k_{\mathrm{obs}}$) by naturally occurring reactive Fe(II) species

| *Fe(II) species* | $k_{obs}$ *(l mol$^{-1}$ s$^{-1}$)* |
|---|---|
| Dissolved Fe(II) | 3.7[a] |
| Solid Fe(II) (green rust) | 6.3[a] |
| Adsorbed Fe(II) | |
| Goethite | 6.6–72[a] |
| Aluminum oxide | 3.5–7.5[a] |

[a] Source: Buerge IJ and Hug SJ (1999) Influence of mineral surfaces on chromium(VI) reduction by iron(II). *Environmental Science and Technology* 33: 4285–4291.

## Organic Contaminants

### Background

For reduction and oxidation processes to occur for organic compounds, the same general steps of precursor complexation, electron transfer, and product release are involved. Relevant structural properties of the organic compound that have to be considered are identity and location of functional groups, acidity and basicity of those groups, and overall molecular size and shape of the molecule. These properties in turn affect the water solubility of the organic compound and hence its ability to encounter the relevant oxidant or reductant in the soil environment. Active sites on organic compounds are the functional groups. Electron-donating functional groups that are able to form surface complexes, particularly chelate complexes, are effective reductants of oxidized Fe and Mn species. In general, removal of electrons from organic compounds can occur by two different pathways. The heterolytic pathway involves the transfer of two electrons, while homolytic pathways describe one electron transfer and are often referred to as free-radical pathways. Organic free radicals are usually reactive and unstable.

### Oxidation and Reduction in Soil Organic Matter

Soil organic matter (SOM) consists of a wide range of compounds derived from algal, plant, and microbial material with different structures and consequently different properties. Primary functional groups in SOM include O- and N-containing groups that are electron-donating and undergo processes to form humic substances (HS), a subclass of SOM consisting of high-molecular-weight biopolymers that react profoundly with inorganic and organic contaminants. The exact structure of HS is unknown, and the humification process is not completely understood, yet it is a way of sequestering C. Characteristics of the precursor organic carbon and the transformations it has undergone are important factors influencing the chemical properties of HS. Humification of natural SOM can also tie up organic contaminants which themselves often contain electron-donating functional groups through nucleophilic addition reactions. Implicit in the SOM-decomposition model is the paramount role of microorganisms, because they can derive energy from the reduced C source. Faster SOM degradation rates have been shown to result in greater overlap of reduction sequences with depth, indicating the likelihood of abiotic SOM oxidation.

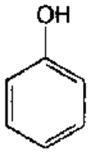
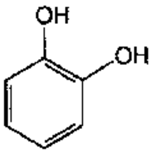
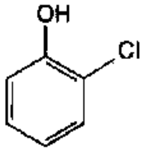
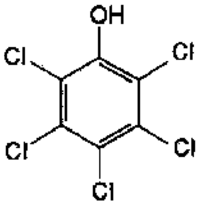
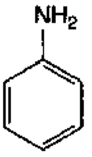
### Manganese and Iron as Oxidants

Direct reaction of organic compounds with $O_2$ is exceedingly slow, especially for organic pollutants. Solid-phase Mn(III,IV) (hydr)oxides can function as important electron acceptors of SOM, chlorinated phenols, and anilines. Catechol (1,2-dihydroxybenzene) is a suitable analogue of SOM, because semiquinones and quinones have been identified in soil HS. Catechol is known for its oxidation and reduction properties and represents a chelating and polymerizing polyphenol.

Naturally occurring birnessite ($\delta$-$MnO_{1.7}$) catalyzes oxidation of catechol more rapidly than $O_2$ and FeOOH based on measurements of Mn(II) release (**Table 3**). Efficient dissolution of birnessite is probably aided by the ability of catechol to form a surface chelate during the precursor-surface complex step with metal center surface sites. In addition, electrons can be transferred immediately along the bonding axis as an inner-sphere process from the filled catechol p orbitals of $\sigma$ symmetry to the vacant LUMO of $\sigma^*$ symmetry of the Mn(III,IV) metal centers after precursor-surface complex formation. This proper symmetry overlap probably explains why Mn oxides are kinetically better oxidants than $O_2$ (see **Figure 2**) and Fe(III) oxides, because the LUMO for Fe(III) is singly occupied ($d^5$, $t_{2g}^3$ $e_g^2$).

Oxidation of substituted phenols by metal ion oxidants is complex, because product distribution and reaction rates are highly dependent on the nature of the oxidant, reactant concentrations, and reaction conditions. Several competing reactions may occur that accompany phenol oxidation and reductive dissolution of Mn(III,IV) (hydr)oxides, such as oxidative coupling of organic radicals or phenolic dimers which form complex mixtures of polymeric oxidation products. In fact, the rapid production of dissolved Mn(II) is accompanied by formation of polymeric reaction products via oxidative coupling reactions from the catechol precursor that resembles soil humic substances (**Table 4**). Spectroscopic indicators of humification are $E_4$:$E_6$ ratios, the ratio of absorbance at 465 nm to that at 665 nm. Lower $E_4$:$E_6$ ratios (approx. <5) correspond to higher organic radical

**Table 3** Initial oxidation rates of soil organic matter (as catechol), chlorophenols, and aniline by naturally occurring Mn and Fe oxides based on Mn(II) and Fe(II) release

| *Organic compound* | *$MnO_x$ d[Mn(II)] dt$^{-1}$ (μmol l$^{-1}$s$^{-1}$)* | *Fe oxides d[Fe(II)] dt$^{-1}$ (μmol l$^{-1}$s$^{-1}$)* |
|---|---|---|
| Phenol  | 0.005[a] | BD[b] |
| Catechol  | 2.1[c] | 0.0009[b] |
| Chlorophenols  | 0.003[a] | |
|  | 0.001[a] | |
| Aniline  | 2.2[d] | |

BD, below detection.

[a] Source: Ulrich HJ and Stone AT (1989) Oxidation of chlorophenols adsorbed to manganese oxide surfaces. *Environmental Science and Technology* 23: 421–428.

[b] Source: Lakind JS and Stone AT (1989) Reductive dissolution of goethite by phenolic reductants. *Geochimica et Cosmochimica Acta* 60: 3169–3175.

[c] Source: Matocha CJ, Sparks DL, Amonette JE, and Kukkadapu RK (2001) Kinetics and mechanism of birnessite reduction by catechol. *Soil Science Society of America Journal* 65: 58–66.

[d] Source: Laha S and Luthy RG (1990) Oxidation of aniline and other primary aromatic amines by manganese dioxide. *Environmental Science and Technology* 24: 363–373.

**Table 4** Spectroscopic indicators of humification based on $E_4$:$E_6$ ratios for products of the birnessite–catechol mixture compared with extracted humic acid from a typical agricultural soil in Kentucky (Woolper silt loam)

| *Sample* | $E_4/E_6$ |
|---|---|
| Birnessite–catechol (pH 4) | 20.3 |
| Birnessite–catechol (pH 6.5) | 3.1 |
| Woolper humic acid | 3.17 |

contents and a greater degree of humification and stabilization of soil humus. The $E_4$:$E_6$ ratios of filtrates for the birnessite–catechol reaction mixture at high pH are comparable with humic acid extracted in a typical Kentucky soil. Direct abiotic coupling of oxidizable organic carbon (C) with solid-phase Mn contributes to sequestration of C through humification processes.

Chlorophenols are well recognized as being highly toxic and represent an organic pollutant class that can be abiotically degraded by solid-phase Mn oxides. Their water solubility and reactivity decrease as the number of chloro substituents added to the phenol increase (Table 3). Detoxification of chlorinated phenols by Mn(III,IV) (hydr)oxides occurs by oxidation, dechlorination, and polymerization processes that can result in less-toxic products that can be biodegraded more readily in soils. During the oxidative coupling (polymerization) reactions, xenobiotic substances such as chlorinated phenols form insoluble precipitates that can be removed from water by sedimentation or filtration. Anilines, another organic pollutant class found in industrial byproducts and components in pesticides, can also be oxidized rapidly in the presence of birnessite (Table 3) and undergo further oxidative coupling reactions to form a variety of polymeric products. Although not as abundant, Mn oxides are more reactive catalysts than Fe oxides for a wide variety of organic pollutant redox reactions.

### Adsorbed Fe(II) as a Reductant

In anaerobic soils and aquifers, Fe(II) species adsorbed on solid Fe(III) oxide sites are important reductants of oxidized organic contaminants such as halogenated solvents and nitroaromatic compounds. A common theme in past research in this area has been the greater reduction rates with Fe(II) adsorbed as an inner-sphere surface complex when compared with the presence of solid-phase or dissolved Fe(II) alone. Surface-bound Fe(II) is not only more reactive but also may be continuously regenerated by the activity of Fe-reducing microorganisms and subsequent adsorption from solution to form new reactive Fe(II) species. The surface density of adsorbed reactive Fe(II) can be manipulated by varying pH values due to the pH-dependent sorption behavior. Thus, reduction rates typically increase with pH because of the greater surface density of reactive adsorbed-Fe(II).

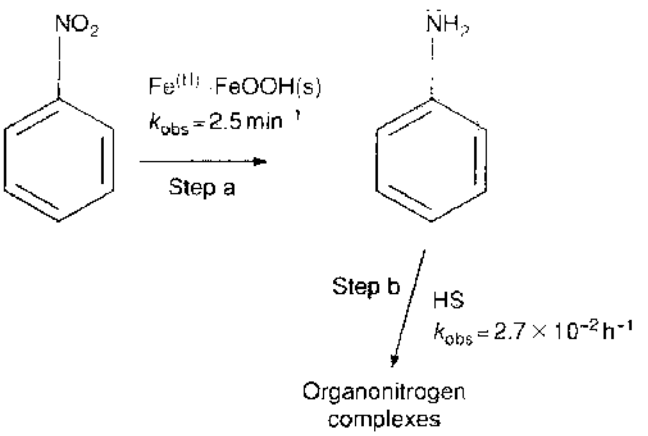


**Figure 4** A two-step remediation strategy for nitrobenzene.

For nitroaromatic compounds that are prevalent contaminants in the soil at munitions facilities, a two-step remediation strategy has been proposed. The first step involves the complete reduction of the nitroaromatic group to the corresponding aromatic amine compounds, followed by nucleophilic addition reactions such as oxidative coupling with natural SOM to form tight covalent bonds (bound residues) that pose minimal environmental threat. Both reduction and covalent binding reactions generally are scaled with pH. A generalized scheme is shown in Figure 4 for nitrobenzene. Adsorbed Fe(II)-goethite species reduce nitrobenzene at rapid rates (step a) to the reduced aniline product. Aniline is then reacted with natural SOM groups and incorporated into humic–organonitrogen complexes. Step b can be catalyzed by addition of Mn oxides to minimize environmental contamination of nitrobenzenes and anilines. Structural analyses of the products of these reactions have been made recently by state-of-the-art $^{15}$N- and $^{13}$C-labeled nuclear magnetic resonance techniques.

*See also:* **Metal Oxides**; **Redox Potential**; **Redox Reactions, Kinetics**

## Further Reading

Buerge IJ and Hug SJ (1999) Influence of mineral surfaces on chromium (VI) reduction by iron(II). *Environmental Science and Technology* 33: 4285–4291.

Fendorf S (1995) Surface reactions of chromium in soils and waters. *Geoderma* 67: 55–71.

Fendorf S, Wielinga BW, and Hansel CM (2000) Chromium transformations in natural environments: the role of biological and abiological processes in chromium(VI) reduction. *International Geology Review* 42: 691–701.

Laha S and Luthy RG (1990) Oxidation of aniline and other primary aromatic amines by manganese dioxide. *Environmental Science and Technology* 24: 363–373.

Lakind JS and Stone AT (1989) Reductive dissolution of goethite by phenolic reductants. *Geochimica et Cosmochimica Acta* 60: 3169–3175.

Luther GW III (1990) Frontier molecular orbital theory in geochemical processes. In: Stumm W (ed.) *Aquatic Chemical Kinetics: Reaction Rates of Processes in Natural water*, pp. 173–198. New York: Wiley Interscience.

Luther GW III, Kostka JE, Church TM, Sulzberger B, and Stumm W (1992) Seasonal iron cycling in the salt marsh sedimentary environment – the importance of ligand complexes with Fe(II) and Fe(III) in the dissolution of Fe(III) minerals and pyrite, respectively. *Marine Chemistry* 40: 81–103.

Matocha CJ, Sparks DL, Amonette JE, and Kukkadapu RK (2001) Kinetics and mechanism of birnessite reduction by catechol. *Soil Science Society of America Journal* 65: 58–66.

McBride MB (1994) *Environmental Chemistry of Soils*. New York: Oxford University Press.

Postma D and Jakobsen R (1996) Redox zonation: equilibrium constraints on the Fe(III)/$SO_4$-reduction interface. *Geochimica et Cosmochimica Acta* 60: 3169–3175.

Schwarzenbach RP, Gschwend PM, and Imboden DM (1993) *Environmental Organic Chemistry*. New York: Wiley Interscience.

Sparks DL (1995) *Environmental Soil Chemistry*. San Diego, CA: Academic Press.

Stumm W (1992) *Chemistry of the Solid–Water Interface*. New York: Wiley Interscience.

Ulrich HJ and Stone AT (1989) Oxidation of chlorophenols adsorbed to manganese oxide surfaces. *Environmental Science and Technology* 23: 421–428.

Wehrli B (1990) Redox reactions of metal ions at mineral surfaces. In: Stumm W (ed.) *Aquatic Chemical Kinetics*, pp. 311–336. New York: John Wiley.

Wersin P, Hohener P, Giovanoli R, and Stumm W (1991) Early diagenetic influences on iron transformations in a freshwater lake sediment. *Chemical Geology* 90: 233–252.

# P

# PADDY SOILS

**C Witt**, Potash & Phosphate Institute/Potash & Phosphate Institute of Canada and International Potash Institute, Singapore
**S M Haefele**, International Rice Research Institute, Metro Manila, Philippines

## Introduction

Wetland rice or paddy soils provide the staple diet for nearly half the world's population. Worldwide, the total harvested rice area covers approximately 150 million hectares, and approximately 90% of this rice area is situated in Asia, with a similar share in rice production. The flooded soil environment is one of the specific characteristics associated with rice cultivation of the traditional rice-growing areas in the river deltas of China, Southeast Asia, and the Indian subcontinent. The term 'paddy soil' does not describe a specific soil type but rather denotes soils in lowland rice production systems with a prolonged period under submergence. 'Water regime' is used to classify the environment of paddy soils into broad ecosystems, distinguishing irrigated and rainfed rice grown predominantly in the lowlands, upland rice grown in the uplands under aerated soil conditions, and deep-water rice in flood-prone river deltas and coastal areas.

## Cropping Environments in Asia's Lowlands

With an increasing demand for rice from a growing population, rice-cropping and paddy-soil management in Asia has undergone substantial intensification since the first modern high-yielding varieties with shorter growth periods, reduced height, and better fertilizer response were released in the early 1960s. Rice production tripled from 1948 to 1990 because of a larger physical land area cropped to rice, an increasing number of crops grown per year, and higher yields as a result of Green Revolution technologies such as improved germ plasm, irrigation infrastructure, water and pest management, and greater use of inorganic fertilizer. The intensification was accompanied by changes in chemical, biological, and physical processes in the soil–floodwater system. The duration of soil submergence and tillage operations increased, especially in the expanding double and triple rice-cropping systems. In such systems, periods of aerobic soil conditions are shorter and are mostly restricted to dry fallows. The duration of the dry fallow between two rice crops varies from a few weeks up to 2–3 months in the typical annual double rice-cropping system in Asia.

In the mid-1990s, the harvested area in Asia's rice ecosystems included 73 million hectares of irrigated rice, 46 million hectares of rainfed lowlands, 9 million hectares of upland rice, and approximately 4 million hectares of flood-prone rice ecosystems. While the overall physical rice area will probably not change much in the future, it is forecast that total rice production in Asia will have to increase by approximately 25% from 530 million tonnes in 1999 to 662 million tonnes in 2020 to meet the increasing rice demand in Asia (**Table 1**). Given the expansion of rice area with irrigation at the expense of upland rice area and the small potential for increasing rice yields in the uplands, yields in irrigated rice would have to increase from $5.3\,\mathrm{t\,ha^{-1}}$ in 1999 to $6.7\,\mathrm{t\,ha^{-1}}$ in 2020. The expected changes in rice demand, environment and socioeconomic conditions, and their effects on productivity and management of paddy soils in Asia's intensive rice systems are depicted in **Figure 1**. Most of the expected production increase will have to come from grain yield increases in the lowland rice ecosystems, and technologies developed in recent years provide substantial opportunities for farmers to increase productivity and profitability as rice demand increases. With sufficient investment, rice production can undergo an intensification that is environmentally sound and sustains the soil resource base. However, in some areas, crop management may change in response to environmental concerns, resource availability or low profitability, with possible adverse effects on rice yields and soils.

**Table 1** Yield, area, and production of irrigated and nonirrigated rice for Asia in 1999, and required changes in yield based on projected changes in area and production requirements until 2020 (Dawe D, IRRI, personal communication.)

| System | Grain yield $t\,ha^{-1}$ 1999 | Grain yield $t\,ha^{-1}$ 2020 | Grain yield Δ% | Area $10^6\,ha$ 1999 | Area $10^6\,ha$ 2020 | Area Δ% | Production $10^6\,t$ 1999 | Production $10^6\,t$ 2020 | Production Δ% |
|---|---|---|---|---|---|---|---|---|---|
| Irrigated | 5.36 | 6.72 | +25% | 76.0 | 78.4 | +3.2% | 406 | 527 | +30% |
| Nonirrigated | 2.08 | 2.39 | +14% | 59.5 | 56.5 | −5.0% | 124 | 135 | +9% |
| Total | 3.92 | 4.90 | +25% | 135.2 | 135.2 | 0.0% | 530 | 662 | +25% |

Projected changes in area and production are adapted from Rosegrant MW, Sombilla MA, and Perez N (1995) *Global Food Projections to 2020: Implications for Investment. Food, Agriculture and the Environment.* Discussion Paper No. 5. Washington, DC: International Food Policy Research Institute.

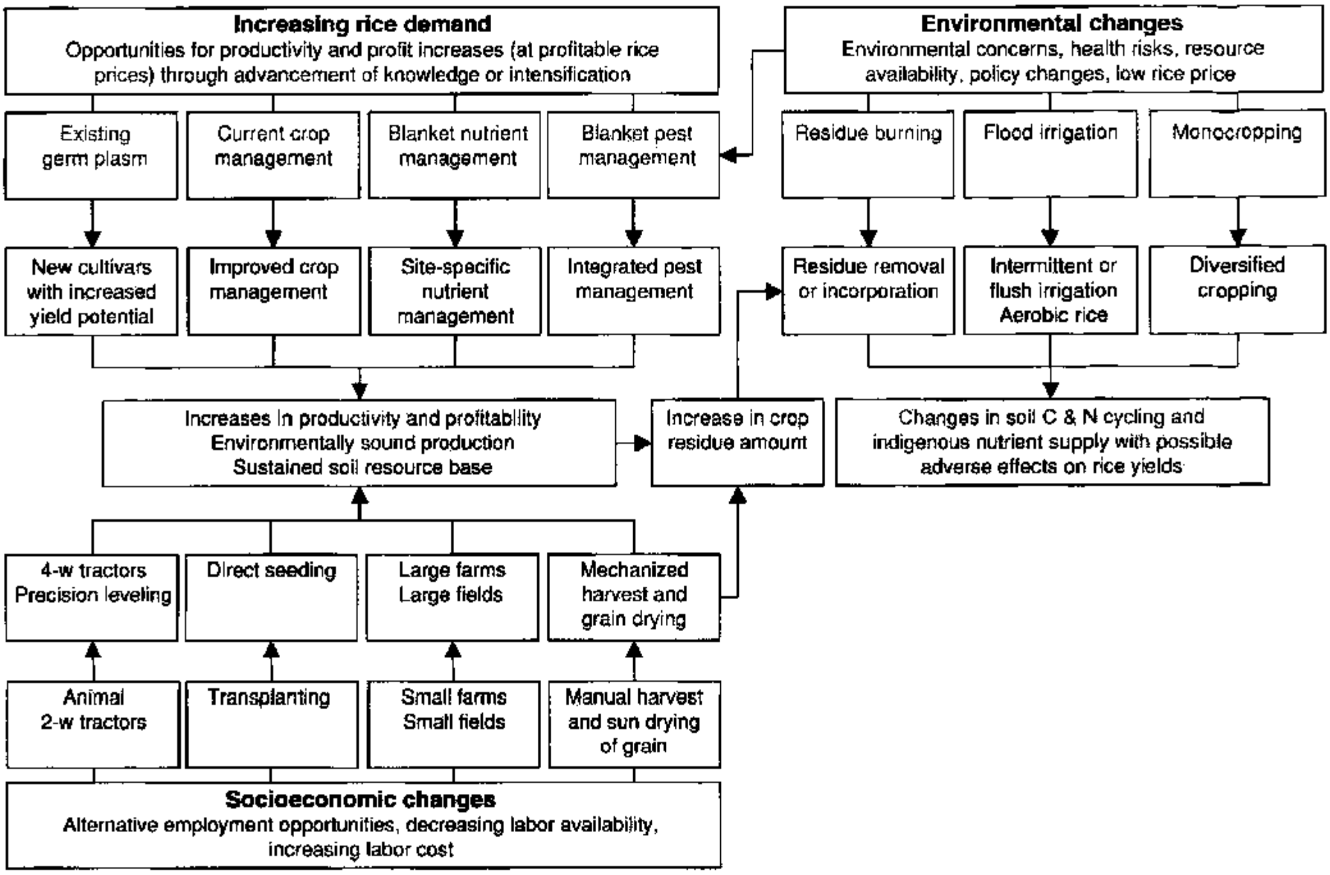


**Figure 1** Expected changes in rice demand, environment, and socioeconomic conditions affecting the productivity and management of paddy soils in intensive rice systems in Asia. Some of these changes have already occurred or are in progress in certain areas. w, wheel.

## Distribution and Important Characteristics of Paddy Soils

Paddy soils constitute an important group within the wetland soils, where water saturation dominates soil development and the types of plant and animal communities. Thus, paddy soils are defined as wetland soils with standing water in bunded and leveled fields used for cultivation of rice (bunds are small dikes around the field to keep the water from running off). Paddy soils may support an upland, nonrice crop in the dry season when soils are aerated, but insufficient drainage capacities often make rice the only possible crop choice in the lowlands of the humid tropics. Free surface water may occur naturally or rainfall, runoff, or irrigation water may be retained by field bunds and/or compacted subsoil layers.

According to the Food and Agriculture Organization of the UN (FAO) World Reference Base for Soil Resources, most paddy soils are designated as Anthrosols (having a puddled surface layer and a plow pan) or as Gleysols, Fluvisols, Planosols, Plinthosols, and Histosols. Smaller areas of paddy soils fall within the gleyic soil units of Arenosols, Andosols, Cambisols, Solonetz, Solonchaks, Luvisols, Lixisols, Acrisols, and Alisols. Although Vertisols, Nitisols, and Ferralsols have no gleyic soil units, these soils may be artificially flooded and used for rice cultivation. The US Department of Agriculture Soil Taxonomy does not recognize wetland soils at the level of soil orders, but classifies soils with aquic conditions at the suborder level and soils with hydromorphism at the subgroup level. Most paddy soils would be assigned to the aquic suborders of Andosols, Oxisols, Vertisols, Ultisols, Mollisols, Alfisols, Inceptisols, and Entisols.

Natural or artificial flooding of paddy soils has marked effects on electrochemical, chemical, and microbial processes, which in turn affect soil fertility and thereby crop growth in a dynamic manner. The main electrochemical changes include a decrease in redox potential and changes in soil and solution pH, microbially mediated processes that are mainly controlled by organic matter and reducible Fe contents of the soil. With decreasing redox potential, oxidized forms of soil redox systems serve as electron acceptors in microbial respiration. The general sequence of reduction is $O_2$–$NO_3^-$–Mn(III, IV)–Fe(III)–$SO_4^{2-}$–$CO_2$–$H^+$. Some of these redox systems are overlapping (e.g., $O_2$, $NO_3^-$, and Mn(III, IV)), whereas others are not (e.g., $O_2$ and Fe(III); $SO_4^{2-}$, and $CO_2$). Reduction of $CO_2$ causes the emission of methane from flooded paddy soils in freshwater systems. The electrochemical changes induce a pH increase in most acid soils, whereas the pH decrease upon flooding of calcareous and sodic soils is due to the accumulation of carbon dioxide. Two distinctly different topsoil layers develop: a thin aerobic surface layer of a few millimeters depth caused by $O_2$ diffusion through the aerobic floodwater layer (and around the rice root surface due to $O_2$ transport through the rice plant's aerenchyma), and the anaerobic soil layer below, with a variable depth. Most roots are concentrated in the first 20 or 30 cm below the soil surface. The presence of aerobic and anaerobic soil compartments causes considerable soil N losses through mineralization–nitrification–denitrification reactions (**Figure 2**). In most soils, flooding increases the plant availability of P, K, Ca, Mg, Na, Mn, Fe, and Si at least temporarily. Reduction of Fe and Mn oxides releases sorbed and occluded elements such as P, Zn, Cu, B, and Mo. Increasing pH reduces plant availability of Zn and Cu, whilst the solubility of Mo increases. Reduction of $SO_4^{2-}$ reduces plant-available S.

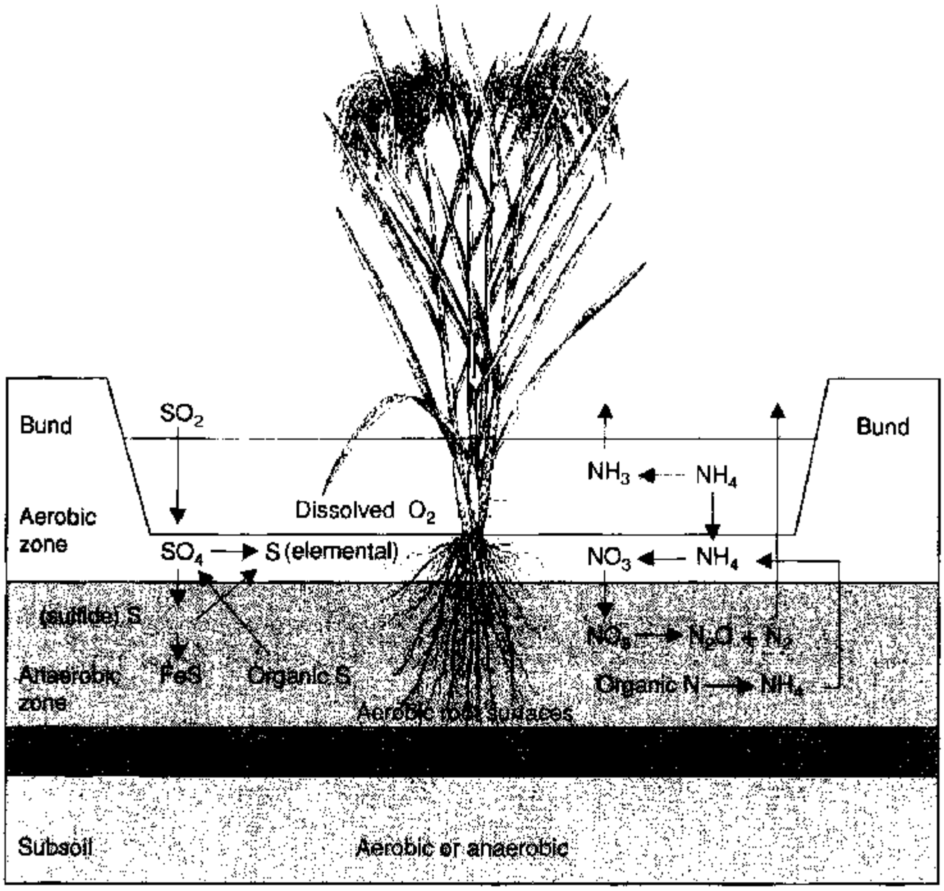


**Figure 2** Aerobic and anaerobic soil compartments in a flooded paddy soil, and related sulfur and nitrogen dynamics. (Adapted from Greenland DJ (1997) *The Sustainability of Rice Farming*. pp. 1–273. Wallingford, UK: CABI Publishing/Manila, Philippines: International Rice Research Institute, with permission.)

A widespread and traditional soil-preparation practice in lowland production systems is puddling. Puddling refers to the practice of mixing surface soil with water to make it soft for transplanting, the traditional crop establishment method for rice. Other short-term puddling effects are the destruction of soil aggregates, irrigation-water savings due to reduced percolation, altered soil bulk density, increased soil water-holding capacity, decreased hydraulic conductivity, an acceleration of the electrochemical processes described above, and good weed control. A long-term effect of puddling is the development of a hardpan (plow pan, traffic pan) in the subsoil below the puddled layer, often associated with the accumulation of Fe, Mn, and Si. The puddled surface layer is often characterized by a coarser texture caused by clay mineral decomposition and/or transportation processes. Hardpan formation may take from 3 to 200 years, depending on soil type, climate, hydrology, and puddling frequency.

## Paddy Soil Fertility and Nutrient Management

The nutrient-supplying capacity of paddy soils in the traditional rice-growing environments is largely determined by their parent material, the alluvial deposits in the river deltas and flood plains. Natural nutrient inputs are derived from sedimentation, irrigation and rainwater, organic residues or manure, biological $N_2$ fixation, and carbon assimilation by floodwater flora and fauna. Efforts to maintain soil fertility have historically centered around maintaining soil organic matter (SOM) through the addition of green manures, animal waste, and crop residues, but their significance has decreased continuously in the last decades.

The total annual C input in a double rice-cropping system may currently range from 4000 to 7000 kg C $ha^{-1}$, including inputs of 1000–2000 kg C $ha^{-1}$ from the photosynthetic and heterotrophic biomass in rice fields. Despite high temperatures in the tropical soil floodwater environment and rapid initial breakdown of organic inputs, average SOM concentrations in paddy soils are higher than in aerated soils, mostly because of high organic matter inputs and a slower decomposition rate of humus. Crop intensification with prolonged soil flooding changes the nature of SOM toward more aromatic compounds (chemical structure with six C-atoms arranged in a ring) because of changes in the breakdown of organic matter under increasingly anaerobic conditions. These changes appear to have limited effects on yield stability under rice monocropping, except in systems with intensive triple cropping and nearly continuous soil submergence.

A recent analysis of yield trends in long-term experiments with intensive rice-based cropping systems in tropical and subtropical regions of Asia has suggested that yields in most experiments have remained stable with recommended fertilizer N, P, and K addition. There is also evidence that organic matter application does not improve grain yield trends in rice-based systems vis-à-vis a balanced mineral fertilizer application. However, the application of organic matter, unless purchased, is often profitable if used as a nutrient source complementary to inorganic fertilizer. The typical irrigated lowland rice soil maintains an average carbon concentration of approximately 1–2% and a nitrogen concentration of approximately 0.1–0.2%, but considerably lower concentrations are found in most rainfed lowland environments.

A unique contribution to the indigenous N supply of paddy soils derives from biological $N_2$ fixation (BNF), providing 15–50 kg N $ha^{-1}$ per cropping cycle from blue-green algae in the floodwater and heterotrophic bacteria in the rice rhizosphere. The mineralization of algae and bacteria provides a constant supply of ammonium ($NH_4^+$), the dominating inorganic N form available to plants under anaerobic soil conditions. Because of efficient plant uptake of mineralized N, larger amounts of extractable $NH_4^+$ are found only after soil flooding during the initial decomposition of residues or if fertilizer N is applied (**Figure 3**).

The major N loss mechanism during the cropping season is ammonia ($NH_3$) volatilization from the flooded soil, particularly when $NH_4^+$ concentrations in the floodwater rise after N fertilizer application. Nitrogen is also leached, denitrified, and/or immobilized by soil microbial biomass and SOM. Approximately 70% of the fertilizer N applied by farmers is currently not recovered by crop N uptake, but N recovery efficiencies of applied fertilizers can reach 50% or more in farmers' fields with adequate real-time N management strategies. Considerable amounts of soil nitrate only accumulate in aerated soil during fallow periods or during upland cropping cycles in rice-based rotations. Almost all nitrate is lost after soil flooding because of leaching and denitrification. Including upland crops in rice-based systems affects mineralization, plant availability, and loss of inorganic N. Soil aeration also accelerates the turnover of SOM and changes its composition toward a less aromatic nature, with possible adverse effects on the amount of soil indigenous N supplied to rice. However, few long-term data are available from experiments that directly compare crop rotations.

Other anticipated future changes in crop management include mechanization of soil preparation and

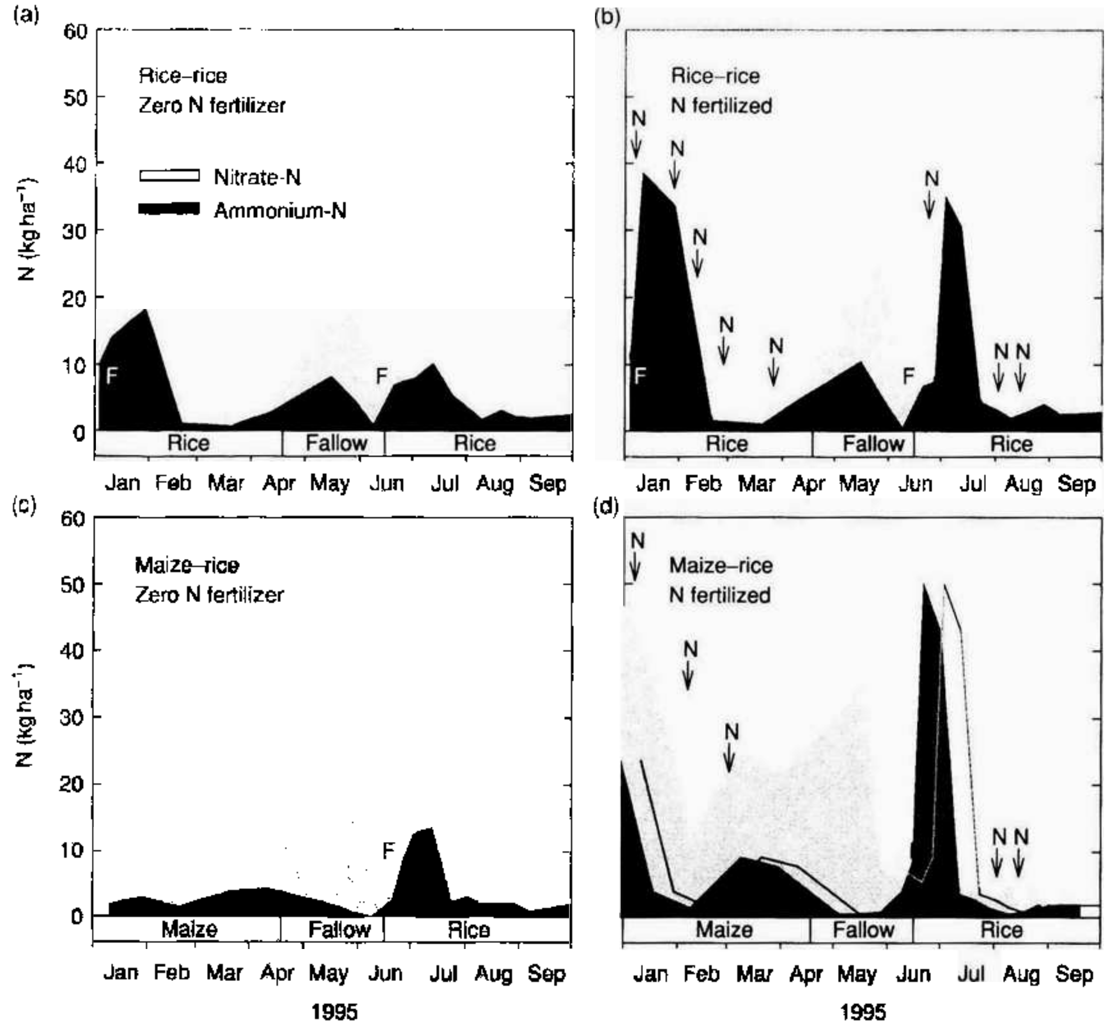


**Figure 3** Nitrate ($NO_3^-$) and ammonium ($NH_4^+$) dynamics in a rice–rice and a maize–rice cropping experiment with and without fertilizer N application, IRRI experimental farm, 1995. Fertilizer N was applied in several split applications with a total of 190 kg N ha$^{-1}$ in the dry season (Jan to mid-Apr) and 100 kg N ha$^{-1}$ in the wet season (mid-Jun to early Oct). F, flooding; N, fertilizer N split application.

harvest operations, changes in water management toward increasing soil aeration, and larger amounts of crop residues remaining in the field for farmers to deal with (Figure 1). These changes will affect soil C and N cycling in paddy soils, and research currently aims to develop a portfolio of technology and management options to mitigate potential adverse effects.

Rice requires an adequate supply of nutrients from various sources to produce high and sustainable yields. A rice crop with a yield of 6 t ha$^{-1}$, for example, accumulates approximately 100 kg N, 18 kg P, 100 kg K, 0.3 kg Zn, 11 kg S, 480 kg Si, 21 kg Mg, 24 kg Ca, 3 kg Fe, and less than 0.1 kg Cu and B. Without fertilizer, the supply of indigenous N limits yield in most paddy soils. Other nutrients that have to be supplied with fertilizer in intensive rice systems include P and K, and in some areas also S and/or Zn. Required fertilizer rates are determined by the deficit between plant nutrient requirement and indigenous nutrient supply. This is particularly relevant for N, because residual effects of fertilizer N application are small and long-term trends show little change in indigenous N supply with time in farmers' fields. For nutrients other than N, fertilizer application is often recommended even in the absence of a yield response (maintenance rates) to avoid depletion of soil nutrient particularly at elevated yield levels at which nutrient removal with grain and straw is high. However, a comparison of current fertilizer rates used by farmers in seven major irrigated rice areas in Asia has shown that fertilizer use is often not in congruence with the deficit between yield level and indigenous nutrient supply (Table 2). On average, farmers use 90–140 kg N ha$^{-1}$, 10–25 kg P ha$^{-1}$, and 10–55 kg K ha$^{-1}$ to achieve grain yields

**Table 2** Average seasonal indigenous N, P, and K supply, fertilizer use, and grain yield in intensive rice systems of South and Southeast Asia. Ranges refer to interquartile ranges (25th, 75th percentiles) of on-farm data

| | *Indigenous nutrient supply (kg ha$^{-1}$)* | | | *Fertilizer (kg ha$^{-1}$)* | | | |
|---|---|---|---|---|---|---|---|
| *Site* | *N* | *P* | *K* | *N* | *P* | *K* | *Yield (t ha$^{-1}$)* |
| JI | 60–79 | 18–24 | 97–141 | 147–192 | 14–23 | 41–75 | 5.3–6.7 |
| HA | 49–68 | 12–19 | 61–102 | 90–118 | 13–30 | 50–80 | 5.5–6.7 |
| AD | 43–59 | 13–20 | 69–109 | 93–137 | 17–27 | 22–54 | 5.3–6.6 |
| TH | 30–44 | 14–19 | 47–65 | 83–124 | 9–25 | 27–45 | 4.6–6.2 |
| SU | 36–74 | 10–14 | 65–91 | 86–148 | 0–17 | 0–0 | 3.6–5.4 |
| MA | 43–65 | 10–18 | 65–100 | 75–142 | 9–20 | 8–31 | 3.9–5.3 |
| OM | 27–49 | 11–18 | 55–94 | 96–123 | 13–25 | 9–37 | 3.4–5.5 |
| Mean | 41–65 | 12–19 | 61–102 | 92–141 | 10–25 | 11–56 | 4.4–6.2 |

JI, Jinhua, Zhejiang (China); HA, Hanoi, Red River Delta (Vietnam); AD, Aduthurai, Old Cauvery Delta, Tamil Nadu (India); TH, Thanjavur, New Cauvery Delta, Tamil Nadu (India); SU, Sukamandi, West Java (Indonesia); MA, Maligaya, Central Luzon (Philippines); OM, Omon, Mekong Delta (Vietnam).
Source: Dobermann A, Witt C, and Dawe D (eds) (2004) *Increasing Productivity of Intensive Rice Systems through Site-Specific Nutrient Management*. Enfield, NH: Science Publishers/Los Baños, Philippines: International Rice Research Institute (IRRI).

of 4.4–6.2 t ha$^{-1}$ in the typical annual double or triple rice-cropping systems of South and Southeast Asia. This is approximately 50–75% of the climate-adjusted, genetic yield potential of 8.3 t ha$^{-1}$ that has been estimated for existing germ plasm across sites, seasons, and years. In general, yields and fertilizer rates are higher in the typical dry season, with greater solar radiation, than in the typical wet season, with greater cloud cover. The indigenous supply of N, P, and K in irrigated rice in Asia was on average 41–65 kg N ha$^{-1}$, 12–19 kg P ha$^{-1}$, and 61–102 kg K ha$^{-1}$, but varied widely with site, year, season, farm, and field. In rainfed lowlands, yields and fertilizer doses are generally lower, decreasing from favorable to unfavorable lowlands with increasing production risk.

**Table 3** Typical input–output balance for P and K during one cropping cycle in intensive irrigated rice systems in South and Southeast Asia

| *Parameters for P and K balance* | *Unit* | *P* | *K* |
|---|---|---|---|
| Input with inorganic fertilizer | kg ha$^{-1}$ | 10–25 | 10–55 |
| Input with irrigation water | kg ha$^{-1}$ | 0–1 | 10–30 |
| Input with rainwater | kg ha$^{-1}$ | 0–1 | 0–10 |
| Losses with percolation | kg ha$^{-1}$ | 0–1 | 5–15 |
| Removal with grain and straw | kg ha$^{-1}$ | 10–15 | 50–65 |
| Nutrient balance | kg ha$^{-1}$ | 0 to +11 | −35 to +15 |

Source: Dobermann A, Witt C, and Dawe D (eds) (2004) *Increasing Productivity of Intensive Rice Systems through Site-Specific Nutrient Management*. Enfield, NH: Science Publishers/Los Baños, Philippines: International Rice Research Institute (IRRI).

Blanket fertilizer recommendations prevail in most of Asia's lowland rice areas without considering amount and spatial distribution of indigenous nutrient supplies. There is also concern that fertilizer P and K rates are not optimally adjusted to long-term needs. A typical input–output balance for P and K in farmers' fields is given in Table 3. On-farm measurements in irrigated rice have confirmed for a wide range of soils and growing environments that farmers' current fertilizer P and K management causes slightly positive P balances, while fertilizer K application is insufficient to balance K removal with grain and straw. Modern site-specific nutrient management (SSNM) strategies have been developed in recent years to manage nutrients more efficiently and supply fertilizers as required by the crop considering indigenous nutrient supplies. Estimates of indigenous nutrient supplies need to be verified if general crop management of paddy soils changes (**Figure 1**). Changes in fallow period and residue management or system diversification are likely to introduce changes in indigenous nutrient supplies that are sufficiently large to require fertilizer adjustments.

The deficiency of Zn is the most widespread micronutrient disorder in rice, not only on soils with a traditionally low supply of Zn (e.g., neutral and calcareous soils containing a large amount of bicarbonate), but increasingly also on other soils with intensive rice cropping. Sulfur deficiency occurs mainly in upland soils containing allophane (an aluminosilicate with primarily short-range structural order which occurs as exceedingly small spherical particles, especially in soils formed from volcanic ash), soils low in organic matter, highly weathered soils containing large amounts of Fe oxides, and sandy soils with high percolation. Lowland rice areas with common S deficiencies are located in Bangladesh, China, India, Myanmar, Pakistan, Sri Lanka, and Thailand. Other nutrient disorders are only of local importance.

## Lowland Rice Systems and Declining Water Availability

Today, nearly 90% of the fresh water diverted for human use in Asia goes to agriculture and, of this, more than 50% is used to irrigate rice. With the growing demand for water for nonagricultural uses (domestic, municipal, industrial, and environmental), the proportion available for agriculture is projected to decline from 90% to 73% in developing countries in the near future. Indicators of a looming water crisis include falling groundwater tables in regions where mainly tube wells are used for irrigation and severe water shortages downstream of some major rivers in Asia (e.g., Yellow River, Indus, and Ganges). **Figure 4** gives the current distribution of irrigated and rainfed lowland rice and major hot spots of groundwater depletion and reduced river water flows. The productivity of intensive rice systems depends strongly on the availability of sufficient irrigation water, and production is likely to decrease with water shortage unless suitable mitigation options are developed. By 2020, physical water scarcity is projected for wet-season irrigated rice in North China (2.5 million hectares), Pakistan (2.1 million hectares), and north and central India (8.4 million hectares), as well as on 2 million hectares of dry-season rice in central India.

Water productivity at the field level can be increased by: (1) increasing yield per unit cumulative evapotranspiration through the use of high-yielding, short-duration varieties and optimal, site-specific agronomic practices; (2) more effectively using rainwater by using dry-seeding rice technologies to synchronize the start of the cropping season with the start of the rainy season; and (3) reducing unproductive water losses (seepage, percolation, evaporation) through adequate soil and water management. Examples of the latter include saturated soil culture (daily flooding up to approximately 0.01 m water depth) or alternate wetting-and-drying regimes (flooding every 2–7 days up to 0.02–0.05 m water depth). These measures frequently cause a reduction in yield and profitability, although they often result in an increase in water productivity.

A more radical approach is to grow rice under permanently aerobic soil conditions by keeping the soil water content in the root zone always at or below field capacity, a technology called 'aerobic rice.' In contrast to 'upland rice,' this system is intended to be high-yielding, but more research is required to evaluate long-term effects on SOM and nutrient cycling. Such systems may gain local importance where water is limiting and rice prices are high, but yields may not reach those achieved in well-managed flooded systems. The adoption of such water-saving irrigation technologies is expected to cause a shift from mainly anaerobic soils to partly or even completely aerobic soils. Changes in the soil-moisture regime of paddy soils will have profound effects on SOM turnover, nutrient dynamics, carbon sequestration, soil productivity, weed ecology, and greenhouse gas emissions in irrigated systems. Water-saving technologies must therefore develop into integrated natural resource management approaches to maintain the productivity, profitability, and sustainability of the evolving systems.

Water productivity in rainfed lowland systems can be increased through more efficient and effective use of all inputs. A general increase in productivity will also reduce the pressure for more intensive management and area expansion of irrigated lands. Direct-seeding techniques offer further opportunities to achieve higher (water) productivity by timely crop establishment. Avoiding drought stabilizes yields, and a second, upland crop can be grown after rice using the residual soil-water content. Such crop intensification must be accompanied by balanced fertilization and adequate residue management on the often poor soils in rainfed systems to maintain or build up SOM and thereby the indigenous nutrient supply.

## Sustainability and Environmental Issues

In general, rice cultivation is sustainable even where rice is continuously grown in monocropping at high production levels. The quality of the soil resource base can be maintained through adequate management practices, including proper crop residue and water management, tillage operations, soil aeration during fallows, and balanced fertilization. The sustainability of rice cropping may be at risk, however, where management-induced changes strongly influence soil characteristics and the soil environment. This may include prolonged soil submergence in intensive systems with three rice crops per year as well as prolonged soil aeration because of crop diversification (**Figure 1**). In the following, we focus on selected issues that are of general importance for sustainable paddy soil management and the environmentally sound production of rice.

When crops are grown with irrigation, the sustainability of agriculture is potentially threatened by several water-related processes such as salinization and alkalization. In the semiarid tropics, irrigation may cause salinization of soils by rising saline groundwater tables. Alkalization can occur when irrigation water with a positive residual alkalinity is used. However, ample irrigation can leach salts from the topsoil through percolation if sufficient drainage is given.

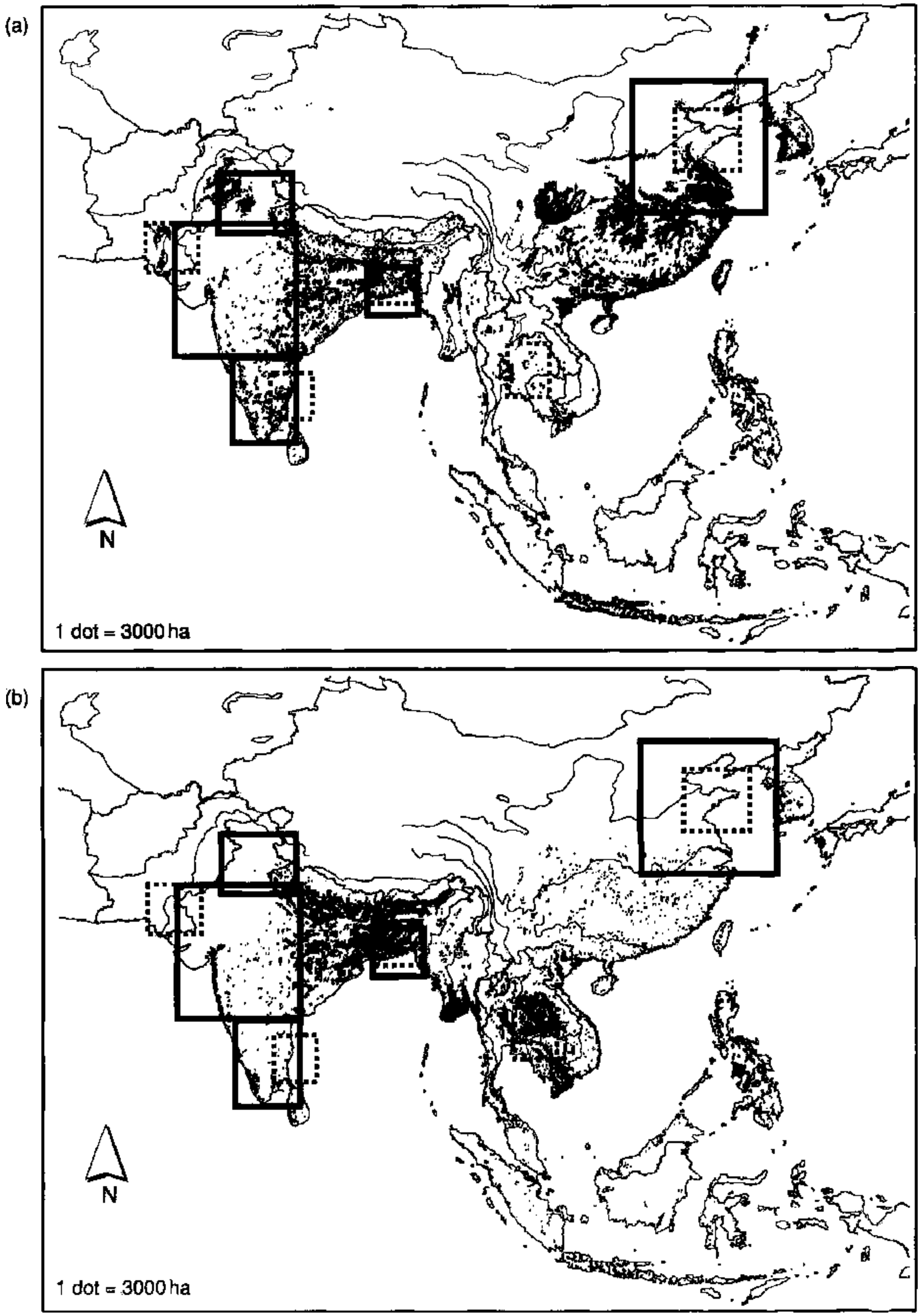


**Figure 4** Current distribution of irrigated (a) and rainfed lowland rice (b) in Asia (Data source: Huke RE and Huke EH (1997) *Rice Area by Type of Culture: South, Southeast, and East Asia.* Manila, Philippines: International Rice Reasearch Institute (IRRI), and major hot spots of groundwater depletion (squares with solid lines) and reduced river water flows (squares with broken lines).

This makes irrigated rice a reclamation crop for such problem soils. A very different water-related problem is accidental flooding caused by inadequate irrigation control in some parts of tropical Asia, taking valuable land out of agricultural production; for example, this occurs in approximately 6 million hectares of arable land in India annually.

Other problem soils in the lowlands include Thionic Fluvisols or acid sulfate soils, which are widespread in Thailand, Vietnam, and Indonesia. Potential acid sulfate soils strongly acidify upon drainage, whereas actual acid sulfate soils are characterized by pH values of approximately 4. Rice cropping under submergence offers one of the few sustainable uses of these soils. Management strategies include: (1) complete soil oxidation through drainage and leaching of acidity from the crop root zone, which offers a permanent solution but may cause a contamination of surface water affecting crops, soils and aquatic organisms in surrounding regions; and (2) the limitation of soil oxidation by maintaining a permanent, high groundwater table, which requires sufficient irrigation water.

Environmental concern is increasing because the intensification of rice cropping in the past has led to an excessive use of fertilizer N and pesticides in many intensive rice-growing areas in Asia. This has been recognized, and much progress has been made in recent years in developing and disseminating adequate management strategies such as real-time N management and integrated pest management. Current changes in today's cropping environment include a shift from irrigated transplanted rice to direct seeding and/or the use of water-saving technologies. This causes greater weed pressure and may lead to an increased use of herbicides, particularly at sites with further intensification and mechanization because of labor shortages. Research efforts aim to integrate better water, nutrient, weed, and pest management approaches to enhance direct environmental and economic benefits for farmers.

Paddy soil environments have also caused global environmental concerns because methane ($CH_4$) is second in importance to $CO_2$ as a greenhouse gas and flooded paddy soils contribute approximately 10% of global $CH_4$ emissions. Although the $CH_4$ production of paddy soils is much smaller than originally assumed, it is still an important contribution. Methods to reduce this share include water-saving technologies as described above, improved crop-residue management, and direct seeding instead of transplanting. Care has to be taken to avoid significant increases in $N_2O$ emissions, another greenhouse gas, which is 10 times more 'effective' than $CH_4$. Soil-moisture regimes keeping the soil redox potential between +100 and −200 mV could minimize both $CH_4$ and $N_2O$ emissions.

## Summary

Paddy soils are defined through their production environment, the typically irrigated and rainfed lowland rice-cropping systems. Cropping intensification since the 1960s has greatly increased rice yield and production, and further intensification is needed to achieve food security at the national and household level. Productivity increases at the farm level could be mainly achieved in two ways. One option is the continuing optimization of the existing rice-based cropping systems with two to three annual rice crops and short fallow periods. There is strong evidence that rice double-cropping in Asia can provide high, sustainable yields despite an increase in the duration of soil submergence and a shift from organic to inorganic nutrient sources since the 1960s. Sufficient inputs of external nutrients with fertilizer are necessary to match the insufficient supply of nutrients from indigenous sources, but inefficient fertilizer use, nutrient imbalances, and excessive pesticide use have become widespread and must be overcome through more site- and season-specific crop, soil, and water management to manage all inputs efficiently. The second option for farmers is to turn to crop diversification, adding more profitable, nonrice crops to the rice-based cropping system or to grow rice under aerobic soil conditions. Conditions imposed in the nonrice season will influence important characteristics of paddy soils and remove typical paddy soil characteristics temporarily or permanently. A key issue is whether the high productivity of current rice-cropping systems can be sustained in such diversified systems.

## List of Technical Nomenclature

| | |
|---|---|
| **Indigenous N, P, and K supply ($kg\,ha^{-1}$)** | The cumulative amount of a nutrient from all indigenous sources circulating through the soil solution surrounding the entire root system during one crop cycle. Indigenous nutrient supply is estimated as aboveground plant nutrient in plots where the element in question is the only limiting factor |
| **Paddy** | (1) Wetland rice; (2) bunded and leveled field used for cultivation of rice. The original meaning of paddy (Malay *padi*s) is 'threshed, unhulled rice' |

*See also:* **Anaerobic Soils**; **Redox Potential**; **Redox Reactions, Kinetics**

## Further Reading

Dawe D, Dobermann A, Moya PF *et al.* (2000) How widespread are yield declines in long-term rice experiments in Asia? *Field Crops Research* 66: 175–193.

Dawe D, Dobermann A, Ladha JK *et al.* (2003) Does organic matter improve the sustainability and profitability of intensive rice systems? *Field Crops Research* 83: 191–213.

Dobermann A and Fairhurst T (2000) *Rice: Nutrient Disorders and Nutrient Management.* Singapore: Potash & Phosphate Institute (PPI), Potash & Phosphate Institute of Canada (PPIC), and Los Baños, Philippines: International Rice Research Institute (IRRI).

Dobermann A, Witt C, and Dawe D (eds) (2004) *Increasing Productivity of Intensive Rice Systems through Site-Specific Nutrient Management.* Enfield, NH: Science Publishers/Los Baños, Philippines: International Rice Research Institute (IRRI).

Greenland DJ (1997) *The Sustainability of Rice Farming.* Manila, Philippines: International Rice Research Institute (IRRI)/Wallingford, UK: CABI Publishing.

Huke RE and Huke EH (1997) *Rice Area by Type of Culture: South, Southeast, and East Asia.* A revised and updated database. Manila, Philippines: International Rice Research Institute (IRRI).

IRRI (1978) *Soils and Rice.* Los Baños, Philippines: International Rice Research Institute.

Kawaguchi K and Kyuma K (1977) *Paddy Soils in Tropical Asia: Their Material Nature and Fertility.* Honolulu, Hawaii: University Press of Hawaii.

Kirk GJD (2004) *The Biogeochemistry of Submerged Soils.* Chichester, UK: John Wiley.

Kirk GJD and Olk DC (eds) (2000) *Carbon and Nitrogen Dynamics in Flooded Soils.* Los Baños, Philippines: International Rice Research Institute (IRRI).

Kyuma K (2004) *Paddy Soil Science.* Kyoto, Japan: Kyoto University Press/Melbourne, Australia: Trans Pacific Press.

Ladha JK, Hill JE, Duxbury JM, Gupta RK, and Buresh RJ (eds) (2003) *Improving the Productivity and Sustainability of Rice–Wheat Systems: Issues and Impacts.* Proceedings of an International Symposium cosponsored by ASA, IRRI, and CIMMYT at Charlotte, NC, USA, 22 October 2001. ASA Special Publication No. 65. Madison, WI: American Society of Agronomy.

Maclean JL, Dawe D, Hardy B, and Hettel GP (eds) (2002) *Rice Almanac.* Los Baños, Philippines: IRRI Bouaké, Côte d'Ivoire: West Africa Rice Development Association (WARDA), Cali, Colombia: International Center for Tropical Agriculture (CIAT), Rome, Italy: FAO.

Moormann FR and van Breemen N (1978) *Rice: Soil, Water, Land.* Los Baños, Laguna, Philippines: International Rice Research Institute (IRRI).

Rosegrant MW, Sombilla MA, and Perez N (1995) *Global Food Projections to 2020: Implications for Investment. Food, Agriculture and the Environment.* Discussion Paper No. 5. Washington, DC: International Food Policy Research Institute.

Tuong TP and Bouman BAM (2003) Rice production in water-scarce environments. In: Kijne JW, Barker R, and Molden D (eds) *Water Productivity in Agriculture: Limits and Opportunities for Improvement. The Comprehensive Assessment of Water Management in Agriculture Series,* vol. 1, pp. 53–67. Wallingford, UK: CABI Publishing.

Wassmann R, Lantin RS, and Neue HU (eds) (2000) Methane emissions from major rice ecosystems in Asia. Reprinted from *Nutrient Cycling in Agroecosystems* (2000) vol. 58(1–3). Los Baños, Philippines: International Rice Research Institute (IRRI)/Dordrecht, the Netherlands: Kluwer Academic Publishers.

**Parent Material** *See* **Pedology:** Basic Principles; **Factors of Soil Formation:** Parent Material

# PEDOLOGY

## Basic Principles

**M J Singer**, University of California–Davis, Davis, CA, USA

### Introduction

Pedology has two meanings according to the *American Heritage Dictionary of the English Language*. The first is the study of the physical and mental development and characteristics of children, and the second is the scientific study of soils, including their origins, characteristics, and uses. The Greek word 'pedon' means ground and 'logy' means study. The International Union of Soil Scientists (IUSS) and many of the member scientists within the organization interpret 'pedology' as encompassing all scientific study of the soil, and hence all soil scientists are pedologists. In this encyclopedia the narrower definition of 'pedology' that characterizes it as the study of the formation, properties, classification, and management of soil is used. Others define pedology as "the earth science that quantifies the factors and processes of soil formation including the quality, extent, distribution, spatial variability and interpretation of soils from microscopic to megascopic scales."

Pedology is the study of soils in their natural setting (Figure 1) and has been equated with the study of soil genesis and soil classification. The people who study soils in these ways are pedologists. Pedology includes pedogenesis, the study of the origin and formation of soil; soil morphology, which is the characterization of the soil as a natural body; the distribution of soil on landscapes, which is soil survey; the organization of soil information into natural and technical classification systems; and the interpretation of soil properties for use and management of the soil resource. Pedology is ". . .an integrative and extrapolative science." Pedologists integrate understanding of landscapes, vegetation patterns, climate, and human activity into knowledge about soils, their distribution, mode of formation, and usefulness. Pedologists also extrapolate information from one scale to the next, to build a coherent model of soils from microscopic observations to the landscape scale. Integration and extrapolation are necessary because soils form a highly variable continuum over the Earth's surface.

### History

Pedology is a relative newcomer to natural science, having emerged as a discipline in the nineteenth century. Although the agricultural value of soil and the importance of the chemical properties of soil were known well before the emergence of pedology, the concept that soil could be studied as a natural body was a nineteenth-century idea. A German scientist, A. Fallou, introduced the term 'pedology' in 1862. Most soil scientists credit the Russian soil scientist V.V. Dokuchaev as the person who made pedology a scientific discipline. He did so by recognizing that the five soil-forming factors (parent material, climate, biota, topography, and time) interact to produce the morphological properties of the three-dimensional soil individual, the pedon (see Figure 2, below). Hilgard is regarded as the first North American-based scientist to recognize pedology as an independent discipline. Hans Jenny is recognized as the nineteenth-century US soil scientist who took the

**Figure 1** Pedologists examining a soil during the 2002 post-meeting field tour of the International Union of Soil Scientists, in Thailand.

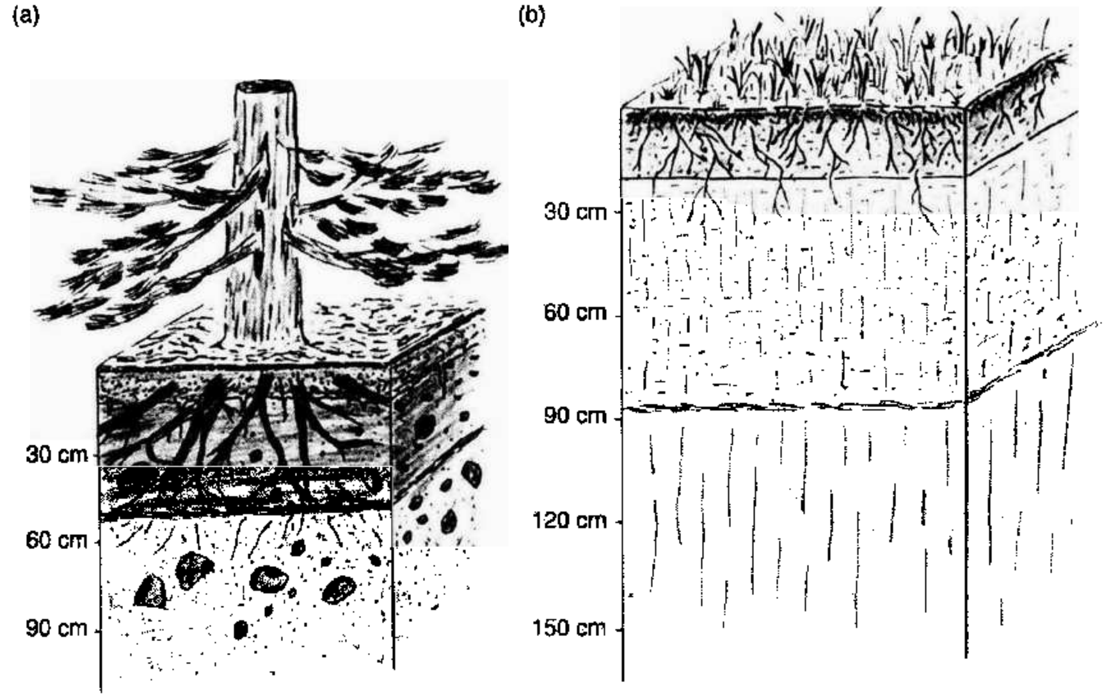


**Figure 2** Soils forming under different climatic conditions with various types of vegetation develop different characteristics that are recognized by pedologists: (a) forest soil; (b) grassland soil.

concept of five soil-forming factors and gave it substance in his 1941 book *Factors of Soil Formation*. He further elaborated on his original text in his 1980 text *The Soil Resource*.

From these early beginnings, which were mostly centered on the mapping of soils and their description, emerged a science that built upon the early foundation and that today is a vibrant mix of individuals who study the soil as a natural body on landscapes, who are continuing to develop concepts about the development of soils through complex biogeochemical pathways, and who are looking for new ways to classify the soil to enhance and protect this resource. As scientists reach out to the moon and beyond, the definition of soil and the role of pedology are expanding.

## Overview

### Soil Classification

Why do we classify soils? The frequently cited reason is to utilize the large body of information about soil. Soil classification systems are grouped into natural and technical systems. Natural systems are those that classify what is seen on the landscape, while technical systems classify soils based on perceived strengths and weaknesses for specific uses. The entries in this encyclopedia focus on the natural classification systems from several countries. These represent diverse methods for categorizing the many kinds of soils that are recognized by pedologists. Although the language of natural soil classification appears foreign and esoteric to the casual reader, soil classification organizes soils with similar properties into useful groups. Most taxonomic systems are hierarchical, with the broadest groupings at the highest levels and the most narrow, most specific groupings at the lowest levels. The informed user can thus learn something about the properties of each group by understanding the construction of the system and the meaning of the terminology.

Classification systems reflect the interests and understanding of the soil scientists who create the systems. Hence the US system of soil classification Soil Taxonomy differs in structure from other systems, because it was developed from its own historical base. The US system has evolved over time, from a system based on perceived processes of soil formation and qualitative definitions that separated broad groupings of soils, to the present system that is based on quantitative definitions. Quantification is the path followed by all sciences and, as the discipline of pedology has matured, the various facets of the science have become more quantitative.

The IUSS has adopted the World Reference Base for Soil Resources (WRB) as the official system for correlating various national soil classification systems. It has also been developed to help users view broad divisions of soils and to facilitate understanding of soils throughout the world. Although each existing soil classification system is different, each is based on carefully defined and quantified criteria that allow the system to be useful to the widest range of individuals.

Technical systems have also evolved throughout the world, with the intent of creating groups of soils with similar capabilities or limitations. These systems are not limited to agriculture; they exist for most of the common uses of soil. For example, systems exist for rating the usefulness of soils for on-site waste disposal, for construction of recreation areas, and as a source of material for road-fill. Because soil underlies and supports much of our civilization, technical systems are critically important guides to the proper selection of soils for different uses.

## Soil Genesis

Soil genesis or pedogenesis is the study of the origin and formation of soil. Pedologists conceptualize soil as a natural, three-dimensional body on the Earth's surface (**Figure 2**). This three-dimensional body, the pedon, is large enough to include all of the properties of the soil individual. Typically the smallest pedon is 1 $m^2$ and 2 m deep. Pedons form within the framework of the five soil-forming factors. 'Parent material' is the term pedologists use to describe the starting material for soil genesis. The parent material may be consolidated (such as rock) or unconsolidated (such as wind-blown silt or sand). The amount of consolidation influences the rate at which water can enter the parent material and therefore the rate of soil formation.

Equally important are the constituent minerals that comprise the parent material. It is the size, crystalline nature, and chemistry of the minerals that determine the rates at which the inherited (or primary) minerals are changed to new minerals or are dissolved and removed from the parent material, thus creating soil. The mineralogy also determines the fertility of the parent material and soil and hence the biological activity on the parent material and soil. This oversimplifies the complexity of pedogenesis, which consists of many interacting processes, which proceed simultaneously at multiple rates.

A special case of how parent material greatly influences soil properties is the formation of soils on parent material rich in clay minerals. Such parent materials form soils that are rich in clay-size minerals and clay minerals. These soils, called 'shrink–swell soils,' tend to change volume greatly as they wet and dry, which makes use of the soils for agriculture and construction difficult.

Topography describes the landscape on which the soil is found. It typically includes the slope angle, slope shape (concave, convex, simple, or complex), and slope direction or aspect. Topography is complex in how it affects pedogenic processes. For example, in the northern hemisphere, a south-facing slope intercepts more solar energy than does a north-facing slope. In dry climates, the south-facing slope will be more water-limited than the north, and the south-facing slope is likely to have less organic matter accumulation, lower biological activity, less organic acid production, and slower mineral weathering than on a north-facing slope. In the southern hemisphere and in moist cold climates, the results are opposite.

The biotic factor includes animal and plant organisms of all sizes and their direct and indirect effects on pedogenic processes. Vegetation is part of the biotic factor of soil formation. Vegetation, under which soils form, mediates the cycle of carbon and nutrients into and through soil. The most obvious effect of plants is the input of carbon into and on to the soil through roots and litter fall. Plant roots and microorganisms respire large amounts of carbon dioxide that decreases soil pH, increases weathering, and has other important effects on soils. Plants also alter the chemistry of precipitation as it falls through the canopy and flows down the stems of plants. Animals mix the soil, consume and modify carbon, and create pores through which flow water and air. All of these processes and more are part of the biotic factor of pedogenesis.

Precipitation and annual temperature are two climatic factors that pedologists consider when studying soil genesis. Others are the timing of the precipitation (e.g., winter versus summer precipitation), rainfall versus snowfall as the dominant precipitation type, precipitation intensity as it affects soil destruction through erosion, and differences between winter and summer temperatures. The climatic factor is usually focused on mean annual temperature and precipitation, because these are the 'macro' climatic factors, but microclimate also influences the many soil-forming processes considered within the context of soil genesis.

Microclimate is the 'on-the-ground' climate that influences small areas of soil, compared with the regional climate. A topographic low that accumulates water and enhances weathering and translocation of weathering products creates a pedologically distinct 'climate' compared with upslope positions. Cold-air drainage into a valley can result in a distinctly different weathering regime than would be characterized by the general climate.

These factors describe the environment and conditions under which a soil forms and determine the properties of the 'product' of these processes. Time is the final natural soil-forming factor. Pedologists have shown that the net rate of soil formation is slow. The time required for horizons to form and for a soil to develop its distinct properties may be thousands of years. Many processes contribute to the differentiation of the parent material into a vertically differentiated soil and they all act at different rates. Thus the question 'How old is this soil?' has no easy answer.

Soils record the history of the Earth's surface and sometimes soils that were once at the Earth's surface are buried. Once buried, they cease to form and they may represent climate and vegetation different from the modern climate and vegetation. These soils are paleosols and they are valuable records of natural history.

The influence of humans on soil development has become a growing area of interest to pedologists. This topic was not of major interest until the end of the twentieth century, when pedologists recognized the profound affect that humans were having on soils. Humans create new areas of soil, reclaim areas that have been destroyed through mining and road building, and destroy soils by modifying them for agriculture or urban uses. In particular, large areas of soil in the USA have been deep-plowed or ripped to remove horizons that are limiting to roots or water penetration. When this is done, the soil, as a natural body, is destroyed and the pedological clock is returned to the starting point.

Observing the vertical distribution of properties made visible by digging a hole or trench or removing vegetation on a vertical road excavation is typically the process used to study soil. Once the properties have been described and the horizons designated, samples are collected by depth increments or by horizon for laboratory analyses. Many chemical, physical, and biological analyses are performed on the samples. Interpretation of the analyses is used to reconstruct the pedogenic history from the time the parent material stabilized and pedogenesis began to the present. This method of studying soil has been referred to as 'static pedology.' Another method of study involves collecting and analyzing the soil solution from within horizons and at horizon boundaries. The data originating from these analyses are interpreted to indicate the types of processes extant in the soil profile. This method has been called 'dynamic pedology.'

## Soil Morphology

Soils form a more or less continuous blanket over the terrestrial surface of the Earth and, in order to distinguish individuals from within this continuum, soil morphology is described in the field. Conceptually the pedon is the soil individual that is described (Figure 2), but in reality what is described is the two-dimensional slice through the pedon that soil scientists call the soil profile. These descriptions begin in the natural environment where the soil is observed, hence the morphological properties are those that can be measured or discerned by a trained observer. The field descriptions are supplemented and enhanced by laboratory analyses and micromorphological observations using thin sections and light and electron microscopes.

The major properties that are described in the field are the thickness, color, texture, and structure of individual layers, called 'soil horizons.' In addition, properties such as the physical behavior at different moisture contents (consistence), pH, boundaries between horizons, and presence of fragments larger than 2 mm in diameter are described. These properties characterize the soil individual, and they are interpreted to predict how soils behave when used for agriculture, recreation, urban development, and waste disposal.

Most pedologists study soils that are well drained. That is, they are well aerated and oxygen-rich for most of the year. The processes that form these soils and their morphology are better understood than the processes and morphology of soils formed under conditions where water remains and oxygen is in short supply for long periods of time. These hydric soils are described using the same properties as well-drained soils, although the soils themselves may look and act quite differently. Another group of soils found under wet, oxygen-poor conditions are the organic soils. Although the vast majority of soils are formed on mineral parent materials in which aluminosilicate minerals are the starting point for soil development, a small but important minority is formed from plant material that is slow to decay because the cold or wet conditions are unsuitable for microbial decomposition.

Organic soils or Histosols consist primarily of dead plant material in various stages of decay. The horizon nomenclature and morphology of these soils are different because of the different processes acting on the parent material. For example, rather than A, B, and C horizons that are standard for mineral soils, the organic soils consist of O horizons. An O horizon is one in which organic material is a major constituent. Each O horizon is given a subscript to denote its organic character. In the US Soil Taxonomy, for example, the subscripts a, e, and i are used to denote strongly decomposed, moderately decomposed, and mostly undecomposed plant material. The terminology used to describe soil morphology has sufficient breadth to encompass all soils. The breadth is accompanied by carefully crafted definitions that allow the specific characteristics of the soil profile to be described.

Modern pedology recognizes that the type of vegetation influences soil morphology, but that morphological properties are not purely a result of the vegetation under which the soil forms. Hence the familiar terms 'forest soil' or 'grassland soil' are more correctly 'soil formed under forest' or 'soil formed under grass' (**Figure 2**).

### Soil Mapping

Pedology developed from the disciplines of agriculture and geology. Both require an understanding of the relationship between the soil individual and the landscape on which the individual is found. Creating soil maps that represent this relationship has been a more or less continuous process in the USA since the end of the nineteenth century. The process accelerated in the twentieth century as aerial photography replaced plane-table mapping and as new concepts of soil formation were learned and applied.

Soil mapping is not the sole provenance of the USA: the developing as well as the developed countries of the world make resource assessments on which development can be based. The Food and Agriculture Organization of the United Nations has led an international effort to create a soil map of the world to help nations and individuals understand the distribution of soil. Maps such as the world soil map are at a low level of resolution, and more detailed maps are required for resource management, but all scales of map are useful in displaying the distribution of soils.

Map scale determines the smallest area that can be easily identified on the map. The scale is selected based on the intended use of the map, the cost, and the finances available. Carefully designed transects across the landscape with many observations at close intervals is the mapping technique used to make very detailed maps. Such soil maps are made for construction projects or very-high-value crops. Fewer observations are made as map scales become less detailed. Soil classification follows the map scale with more general levels in the taxonomy used as map units on generalized maps and more detailed levels in the taxonomy used on more detailed maps.

## Summary

Pedology, according to one broadly accepted definition, is the discipline that studies soil as a part of nature. Pedologists study the origin and formation of soil, the morphology of soil, the distribution of soil on landscapes, and the management of soil. The pedology entries in this encyclopedia bring to life these topics and explain the role of pedology in the study of the most important material of the Earth's surface, soil.

## List of Technical Nomenclature

| | |
|---|---|
| **Aluminosilicates** | Minerals with aluminum and silicon as major constituents. Examples include the clay minerals |
| **Dynamic pedology** | The study of soil formation processes by sampling and analyzing the soil solution in different horizons or at horizon boundaries |
| **Histosols** | Organic soils that have organic soil materials in more than half of the upper 80 cm, or that are of any thickness if overlying rock or fragmental materials that have interstices are filled with organic soil materials (an order in the US system of Soil Taxonomy) |
| **Horizon** | A layer of soil or soil material approximately parallel to the land surface and differing from adjacent genetically related layers in physical, chemical, and biological properties or characteristics such as color, structure, texture, consistency, kinds and number of organisms present, degree of acidity or alkalinity, etc. |
| **Hydric soils** | Soils that are wet for long enough to produce periodically anaerobic conditions, thereby influencing the growth of plants |
| **Micromorphology** | The study of soil morphology by microscopic (light optical and, less frequently, by submicroscopic) methods, often using thin-section techniques |
| **Natural soil classification** | The systematic arrangement of soils into groups or categories on the basis of their characteristics. Broad groupings are made on the basis of general characteristics and subdivisions are made on the basis of more detailed differences in specific properties |
| **Paleosols** | A soil that formed on a landscape in the past with distinctive morphological features resulting from a soil-forming environment that no longer exists at the site. A paleosol (or component horizon) may be classed as relict if it has persisted in a land-surface position without major alteration of morphology by processes of the prevailing pedogenic environment. An exhumed paleosol is one that was formerly buried and has been reexposed by erosion of the covering mantle |
| **Parent material** | The unconsolidated and more or less chemically weathered mineral or organic matter from which the solum of soils is developed by pedogenic processes |

| | |
|---|---|
| **Pedogenesis** | The processes that produce soils from parent materials |
| **Pedologists** | The individuals who study soils |
| **Pedon** | A three-dimensional body of soil with lateral dimensions large enough to permit the study of horizon shapes and relations. Its area ranges from 1 to 10 m$^2$ |
| **Soil-forming factors** | Those external factors that control the kinds and rates of processes that form soils: parent material, climate, biota, topography and time and the five soil-forming factors |
| **Soil taxonomy** | The US system of soil classification |
| **Static pedology** | A method for studying soil formation by sampling the soil solid phase and measuring properties of the solids |
| **Technical soil classification** | A soil classification system that rates soils based on desired use of the soil |

## Acknowledgment

Many of the definitions within this List of Technical Nomenclature were reproduced with permission from the Soil Science Society of America Glossary of Soils Science Terms.

*See also:* **Classification of Soils**; **Factors of Soil Formation:** Biota; Climate; Human Impacts; Parent Material; Time; **Pedology:** Dynamic

## Further Reading

Birkeland PW (1999) *Soils and Geomorphology*, 3rd edn. New York: Oxford University Press.

Buol SW, Hole FD, and McCracken RJ (1980) *Soil Genesis and Classification*, 2nd edn. Ames, IA: Iowa State University Press.

Food and Agriculture Organization of the United Nations (2001) *Lecture Notes on the Major Soils of the World*. World Soil Resources Report 94. Rome, Italy: FAO–UNESCO.

Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw-Hill.

Jenny H (1961) E.W. Hilgard and the birth of modern soil science. In: Berkeley, CA: Farallon Publications/ Pisa, Italy: Agrochimica, Istituto di Chimica Agraria dell' Universita. *Agrochimica*.

Jenny H (1980) *The Soil Resource, Origin and Behavior*. New York: Springer-Verlag.

Tandarich JP, Darmody RG, Follmer LR, and Johnson DL (2002) Historical development of soil and weathering profile concepts from Europe to the United States of America. *Soil Science Society of America Journal* 66: 335–346.

Pickett J (ed.) (2000) *The American Heritage Dictionary of the English Language*, 4th edn. Boston, MA: Houghton Mifflin.

Wilding L (2000) Pedology. In: Sumner ML (ed.) *Handbook of Soil Science*, pp. E1–E4. Boca Raton, FL: CRC Press.

Yaalon DH (1997) History of soil science in context: international perspective. *Advances in Geoecology* 29: 1–15.

# Dynamic

**F C Ugolini**, Università degli Studi di Firenze, Firenze, Italy

The expression 'dynamic pedology' was first introduced in the open literature in 1978. Later, in 1985, the term was adopted and the approach used to examine the genesis of a Gray Luvisol in the boreal forest of Canada. Dynamic pedology singles out pedological investigations conducted with the use of soil solution obtained in the field through lysimetry; it does not include experimental pedology where solution is obtained by leaching or displacement from soil columns, or laboratory experiments conducted to simulate weathering. Dynamic pedology is distinguished from traditional pedology or 'static pedology' in which only the solid phase is investigated. Paraphrasing a biological expression 'dynamic pedology' is equivalent to an *in vivo* versus an *in vitro* study of the soil.

The collection, analysis, and interpretation of the soil solution allow evaluation of the contemporaneity of the pedogenic processes and hence the dynamic aspect of the soil. In contrast, the analyses of the solid phase tend to average the sum total of all the processes that have occurred since the parent material was laid down. Furthermore, the proton donors considered responsible for the formation of soils can be identified in soil solution. If pedogenesis is defined as a combination of weathering and transportation processes occurring in the presence or absence of living and dead biota, proton donors initiate weathering reactions and participate in transport of ions released during weathering and biological activity. Questions inquiring about the nature and the actuality of the processes cannot be answered by traditional pedology. This is particularly crucial when processes are at their limits as the process of podzolization in the High Arctic or in those environments where soil formation is dictated by catastrophic events. In terrestrial ecosystem studies, the watershed approach has been useful for providing a balance of the ionic inputs and outputs; however, it has not provided an insight to soil processes. In these studies the soil is considered as a black box placed between the incoming precipitation and the stream discharge. Dynamic pedology, by examining the soil solution exiting the genetic horizons, could, on the other hand, explain the changes in ionic composition caused by the processes occurring at the horizon level. The dynamic pedology approach is suitable to examine perturbations such as substitution of species, effects of fire, addition of fertilizers, pesticides, and pollutants. During the early 1970s, when acid rain became a large issue in Western Europe and North America, soil solution collected via lysimeters became the best way to monitor soil acidification and pollutants' movement in soils.

## Early Stages of Soil Development on Recent Volcanic Ejecta

An example of soil development comes from a study conducted on an Entisol (**Figure 1**) developed on a pyroclastic flow emplaced during the 1980s eruption of Mount Saint Helens (USA). Soil solution collected between 1984 and 1986 shows that vapors of strong acids ($H_2SO_4$, HCl, $HNO_3$, and HF) produced by volcanic emanations during the eruption reacted with the airborne tephra, resulting in the formation of soluble salts that were promptly leached once the tephra was deposited on the ground. In addition, acid precipitation (pH 3.6–5.2) from gaseous emissions produced during the dome-building stage after the eruption has contributed to the introduction in the soil of proton donors, mostly $H_2SO_4$ and HCl. These strong mineral acids have been neutralized within the first 5 cm of soil (**Figure 2**). Calcium, $Mg^{2+}$, $Na^+$, and $K^+$ associated with $SO_4^{2-}$, $Cl^-$, and $NO_3^-$ dominate the soil solution together with Si. Colonization by nitrogen-fixing lupine (*Lupinus lepidus*) has introduced a new regimen in the weathering and soil genesis. Soils under lupine plants show a pH higher than the near-barren sites. The increase in pH is due to the neutralizing effect of the acid deposition by the lupine foliage, which exchanges base cations with the $H^+$ of the rain. In the barren areas, the consumption of protons occurs in the first 5 cm; in the presence of lupines, a redistribution of the proton load occupies the entire rooting zone. Hydrogen ions are released by the roots in exchange for cations that buffer the acidity in the foliage. As a result of lupine invasion, the soil solution under these plants contains more dissolved organic matter (DOM), $H_2CO_3/HCO_3^-$, $NH_4^+$, and $Cl^-$.

In summary, in the proximity of an active volcano, soil development on pyroclastic material 4 years old must be interpreted as a function of those events that have occurred prior to, during, and after the deposition of volcanoclastic material. On the barren sites, strong mineral acids are the major proton donors, but they have been neutralized in the upper part of the soil (**Figure 2**). Under this weathering

**Figure 1** Soil profile (Entisol) developed on Mount Saint Helens pyroclastic flow. The surface of the flow is visible in the background. The measuring tape is 1.0 m, with 10-cm intervals.

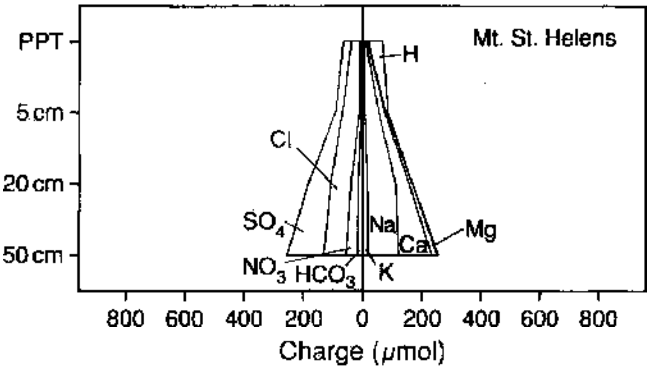


**Figure 2** Charge balance for the precipitation (PPT) and for the soil solution of the Entisol, Mount Saint Helens. Concentrations of the different charged species as they exit the depths marked on the left. (Reproduced from Ugolini FC and Dahlgren RA (2002) Soil development in volcanic ash. *Global Environmental Research* 6: 73.)

regimen, there is a loss of Si, Ca, Mg, Na, and K; Fe and Al are not mobilized. With the colonization of nitrogen-fixing plants, the acidity of the incoming precipitation is neutralized by the foliage of the lupine rather than by the soil. Here, there is a shift in proton donors compared with the barren sites. Under the lupines, DOM and $H_2CO_3/HCO_3^-$ are the major proton donors. The concentration of cations, anions, and Si under the lupines is higher than in the barren sites.

## Migration of Particles in the Process of Podzolization

The unique morphology of Spodosols, supported by analytical data, has been attributed to the movement of DOM and metal cations in solution or suspension. Migration of clay particles has not been considered of importance in this process, although it has been detected in the early evolutionary stages of Spodosols. Lysimeters emplaced in well-developed Spodosols formed under Pacific silver fir (*Abies amabilis* (Dougl.) ex J. Forbes) forest, in the Washington Cascades (USA), have shown the existence of two distinct chemical compartments: the upper one starting from the top of the canopy and ending at the boundaries of the Bhs/Bs horizons, and the lower one below this boundary that reaches into the parent material. Weathering and ion transport in the upper compartment are governed by DOM that acidifies the soil (pH 3.8) and prevents the dissociation of the $H_2CO_3$. In the lower compartment, the pH is higher (approx. 5) than in the higher compartment, owing to the liberation of $OH^-$ during the ligand exchange reactions between the DOM and the surface of amorphous materials. This compartment is dominated by $HCO_3^-$ (**Figure 3**). Fulvic acid (FA) has been detected in the upper compartment, with 70% reduction across the Bhs/Bs. Electrodialysis, adsorption, molecular sieving, and other techniques indicate that FA is transported in solution in particulate form. However, direct evidence of their migration in soil is lacking. Lysimeters placed at the boundaries of the E, Bhs, and Bs of a Spodosol collect soil solution and suspended material, including humic substances. Upon filtration through a 0.1-$\mu$m Nucleopore filter, the suspended material is collected, lyophilized, and observed on a scanning electron microscope (SEM) equipped with energy-dispersive X-ray analysis (EDAX). Through the SEM it is possible to observe the presence of filaments 2–12 $\mu$m long extending from oblong bodies 2–3 $\mu$m wide. These bodies contain minor quantities of Al, Fe, P, S, and Si (**Figure 4**). Leachates collected below the Bs, on the other hand, show almost exclusively the presence of mineral particles ranging in size approximately from 2.0 to 22 $\mu$m. EDAX spectra of these particles show that they are made of Si, Al with some Fe, Ca, and K, reflecting the composition of silicate minerals. This study indicates that the suspended material in the soil solution reflects the trend of the solution itself. The reduction of DOM across the Bhs/Bs is also verified by the presence of organic particles in the upper profile (**Figure 4**) and their absence below the Bhs/Bs (**Figure 5**). This study provides direct evidence for the mechanism of podzolization and the migration of particulate matter.

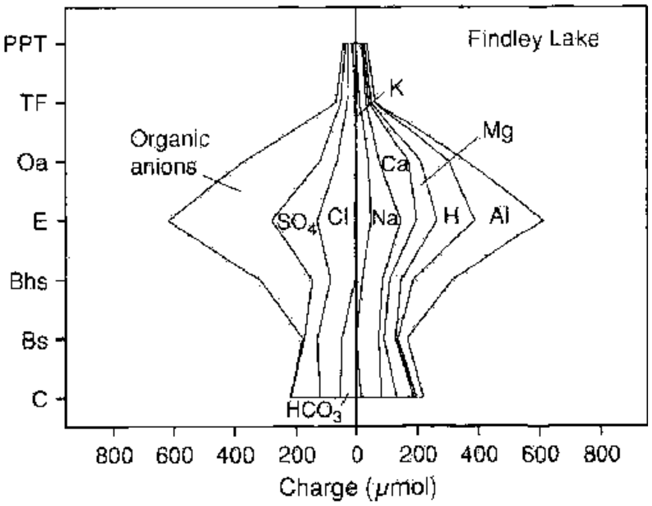


**Figure 3** Charge balance for the soil solution of the Spodosol, Washington Cascades. Concentrations of the charged species for precipitation (PPT), canopy throughfall (TF), and the soil horizon solutions. The anion deficit was assumed to be due to organic anions. (Reproduced from Ugolini FC and Dahlgren RA (2002) Soil development in volcanic ash. *Global Environmental Research* 6: 73.)

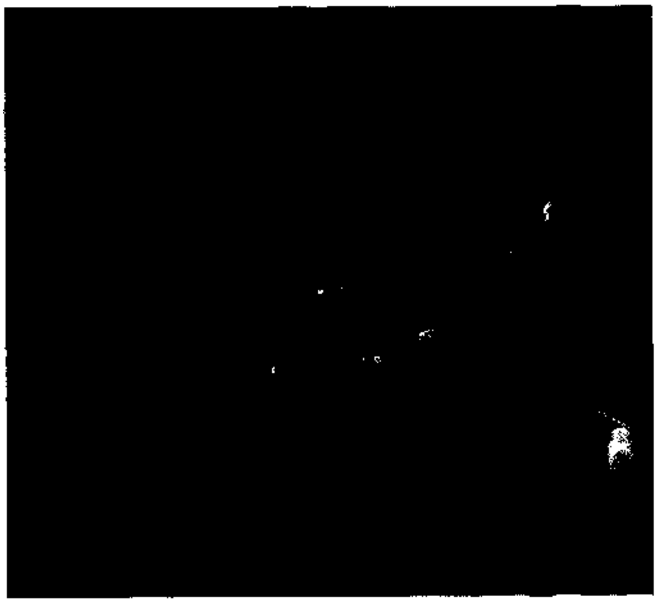


**Figure 4** Scanning electron micrograph of an organic particle collected on 0.1-$\mu$m Nucleopore filter from soil solution obtained below the Oa horizon of a Spodosol, Washington Cascades. The filaments are 5–12 $\mu$m long, the oblong bodies are 2–3 $\mu$m wide (×5000).

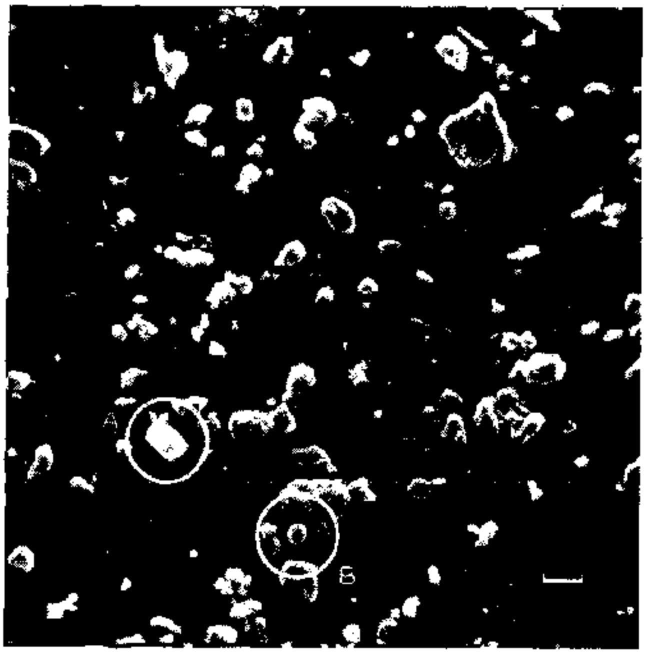


**Figure 5** Scanning electron micrograph of silicate particles collected on 0.1-μm Nucleopore filter from soil solution obtained below the Bs horizon of a Spodosol. Washington Cascades. Circle A encloses a diatom, circle B, a spherical siliceous particle. Scale bar 10 μm (×450). (Reproduced with permission from Ugolini FC, Dawson HJ, and Zachara J (1977) Direct evidence of particle migration in the soil solution of a Podzol. *Science* 198: 604.)

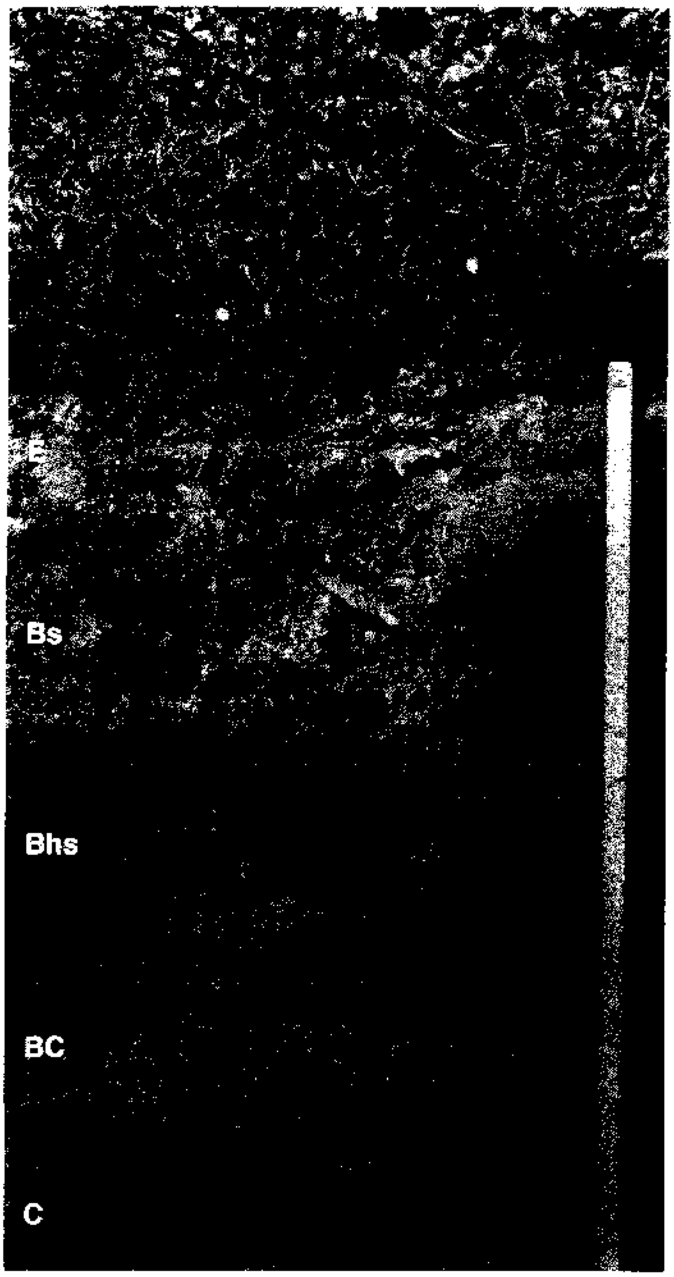


**Figure 6** (see color plate 48) Boreal forest Spodosol showing an inverted sequence: a Bs over a Bhs; Walker Lake, Alaska. The yardstick is 50 cm long.

## Evidence of Catastrophic Events and Horizon Inversion

Traditionally, soil formation is considered to be a relatively slow process, representing a continuum in time and space. In the absence of adequate techniques, short-term or catastrophic events in soil development have not been appreciated. However, improvement in soil-solution collection and analytical procedures has made the investigation of isolated events possible. At Walker Lake, Alaska (67° 09′ N and 154° 22′ W), the use of lysimeters has provided new evidence of threshold-controlled episodes associated with rapid solution-transport and particle migration. These events result in the formation of horizon inversion. At Walker Lake, in the boreal forest, the Spodosols have formed in the boreal forest on glacial drift deposited c. 13 000 BP. The horizon sequence shows a perplexing inversion of the Bhs below the Bs (Figure 6). From 25–28 July 1981, an intense storm followed by a prolonged leaching episode caused a radical departure from the average trend. The sequential change in solution composition shows a distinct flashing effect associated with normal leaching episodes. The flashing effect is a well-known phenomenon associated with antecedent moisture conditions, and it is typical of semiarid and subhumid climates. In the absence of complete weather information for this remote site, sodium can be used as an indicator of antecedent moisture conditions. High initial concentrations of this ion indicate that a long period of drought has occurred. The high-intensity, low-frequency rainstorm delivered over 150% of the average total precipitation anticipated for the entire month. The pedogenic response was a progressive and dramatic surge in subsurface translocation processes (**Figure 7**). The successive stages in these processes, using iron as an indicator parameter, are shown in **Figure 7**. The first stage displays a normal leaching trend: depletion of Fe from the E horizon and accumulation into the Bs. After continuous leaching, phase 2 was initiated. In

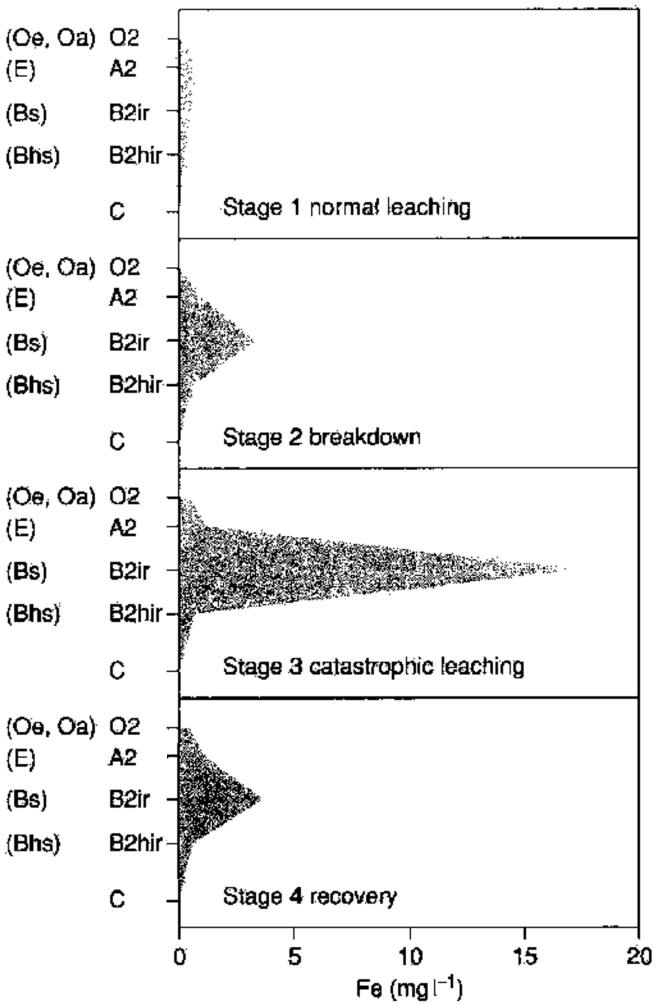


**Figure 7** Dissolved iron (milligrams per liter) used as an indicator to separate the stages of a catastrophic leaching episode in a boreal forest Spodosol, Walker Lake, Alaska. (Adapted from Stoner MG and Ugolini FC (1988) Arctic pedogenesis. 2. Threshold-controlled subsurface leaching episodes. *Soil Science* 145: 49 © 1988, Lippincott Williams & Wilkins, with permission.)

this stage there is a depletion of metal cations and DOM from the Bs and translocation into the Bhs. Since no abnormal concentration of metal cations and DOM appears to leave the O and E horizon leachates, this surge indicates that the source is the Bs itself. Evidently, this horizon had been undergoing a degradation of the microstructure with the liberation of metal–organic (metal-organo) particles that had become suspended in the soil solution. Stage 3 shows a maximum liberation of particulate matter that leaves the Bs and is deposited in the Bhs. Filtered and unfiltered solutions show that 80% of Fe and Al is in particle form. This influx of suspended material tends to build up the Bhs horizon below the Bs. Stage 4 represents a recovery. This monitored response to a major rainstorm helps to explain the observed horizon inversion – the Bs over the Bhs – a feature observed in other Spodosols of the Arctic. Storms of less intensity are able to moisten only the upper part of the profile. During these episodes, the penetrating wetting front leaves mobile metal-organo complexes stranded in the upper Bs. These complexes are only temporarily detained, because their metal-to-C ratio is too low to induce their precipitation. This example serves to demonstrate the usefulness of lysimetry for capturing the impact of individual rainstorms. Further, it appears that in subhumid climates soil development occurs through episodic pulses corresponding to rainstorms of moderate magnitude. This trend is punctuated by catastrophic events that are responsible for the unusual migration of Fe, Al, and organic matter in particulate form. The catastrophic events are responsible for the observed B-horizon inversion. A similar horizon inversion has been observed for a Podzol located in the Massif Central (France). However, this phenomenon is explained by two separate pedogenic phases involving truncation and burial.

## Effects of Species Substitution

Dynamic pedology is particularly useful when applied to study the impacts of replacement of species either naturally or anthropically induced. In northern Idaho, disturbed areas such as clearcuttings in the Grand Fir Mosaic Ecosystem are poorly revegetated by the native firs. Bracken (*Pteridium aquilinum* (L.) Kuhn) rapidly occupies the disturbed sites preventing the regeneration of the Grand Fir (*Abies grandis* (Dougl. ex D. Don) Lindley). Solutions collected under the bracken show a lower pH (minimum 4.8), a higher Al concentration (maximum 1.6 mg l$^{-1}$), and higher DOM than the undisturbed forest during 1994 and 1995.

Another study on the effects of species substitution has been conducted in northeastern Japan, in the Miyagi prefecture, to examine, through soil solution, the effects of Japanese oak trees (*Quercus serrata* Thumb.) on Alic Melanudands (nonallophanic Andisols). The oaks were established 50 years ago on a site originally occupied by Japanese pampas grass (*Miscanthus sinensis* Anders). The soil-solution composition collected under the Japanese oaks has been compared with that of the pampas grass growing in a plot next to the oaks (**Figure 8**). The acid precipitation (pH 4.8) caused by the presence of $HNO_3$ and $H_2SO_4$, of volcanic origin, and paddy fields is partially neutralized by the pampas grass canopy from pH 4.8 to 6.0, while the oak canopy increases the pH only to 5.4. As the canopy drip penetrates the soil, the soil solution tends to acquire a lower value (pH 4.6) in the upper 50 cm of the soil under oak than in the pampas grass (pH 5.5). Al concentrations, in the horizon under the oaks are significantly higher (38 $\mu$mol l$^{-1}$) than in the pampas

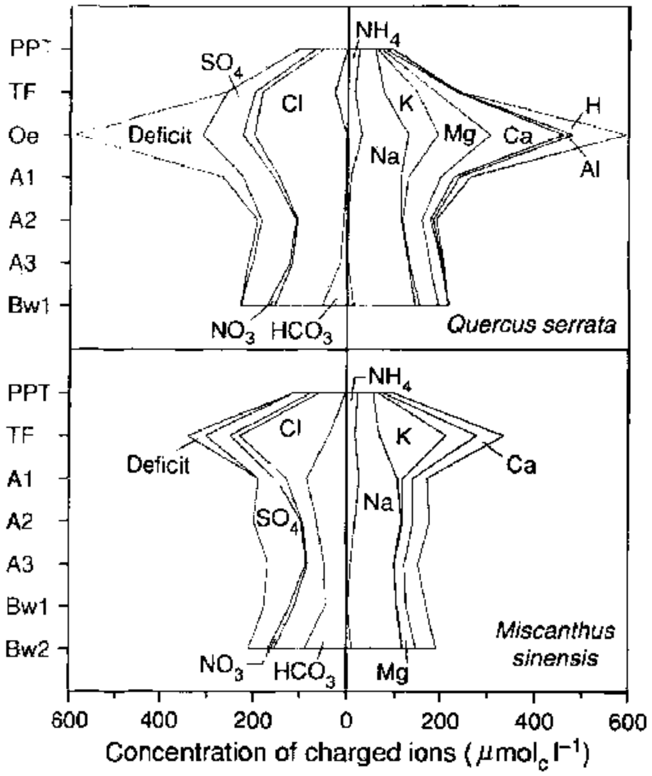


**Figure 8** Charge balance for the soil solution of the Andisols developed under *Quercus serrata* and *Miscanthus sinensis*, Japan. The graph shows the concentrations of the charged species for precipitation (PPT), canopy throughfall (TF), and the soil horizon solutions. The anion deficit was assumed to be due to organic anions. (Reproduced with permission from Dahlgren RA, Ugolini FC, Shoji S, Ito T, and Slatten RS (1991) Soil-forming processes in Alic Melanudands under Japanese pampas grass and oak. (1991) *Soil Science Society of America Journal* 53: 1055.)

grass (4 $\mu$mol l$^{-1}$). However, in both cases there has been no significant movement of Al into the B horizon. Iron concentrations are less than 1 $\mu$mol l$^{-1}$ in all horizons, except the Oe under the oaks, where a mean value of 0.4 $\mu$mol l$^{-1}$ has been recorded. Silicon concentrations under the oaks are high (231 $\mu$mol l$^{-1}$) in the Oe due to biocycling. In the pampas grass plot, the highest levels are approximately 122 $\mu$mol l$^{-1}$ in the A horizon. Also, here the large amount of Si is due to the biocycling of silicon by the pampas grass. DOM concentrations reach a maximum of 3000 $\mu$mol l$^{-1}$ in the Oa of the oak site, but they decrease to less than 500 $\mu$mol l$^{-1}$ in the Al horizon. In the pampas grass site, the values are lower, approximately 1000 $\mu$mol l$^{-1}$, than under the oak. Under both species the mobility of DOM is limited to the surficial horizons. Uptake processes, in addition to the task of neutralizing the acid precipitation, mostly affect base cations (Na, K, Ca, and Mg). Calcium shows a higher concentration in the pampas grass, while Mg and Na are more abundant in the oaks. Anions show both differences and similarities between the two species. Chloride and sulfate show the highest concentrations. Sulfate reaches a minimum value in the B horizons, where it is retained by amorphous material.

Bicarbonate levels are higher under the pampas grass than under the oaks. In conclusion, the invasion of Japanese oak trees on a soil previously occupied by pampas grass has resulted in acidification of the soil solution in the upper 50 cm of the profile. This acidification has resulted in an increase in Al concentration and in a decrease in bicarbonate. Silicon tends to accumulate in the B horizon of the pampas grass, but with a net loss under the oaks. Also the oaks induce a greater cycling of Ca, Mg, K, and Cl. In spite of these differences, the process of andosolization, as characterized by *in situ* weathering in the B horizon and accumulation of organic carbon in the A, is still active under both species. The proton donors in the Oa of the oak site are the organic acids, while these acids have a lesser role in the pampas grass site, where carbonic acid prevails. Protons released by cation uptake dominate in the rooting zone of the oaks.

## Impact of Burial by Volcanic Ash on Preeruption Soils

Regions in the proximity of active or dormant volcanoes are subjected to intermittent deposition of airfall tephra. These additions tend to rejuvenate the soil surface of the existing soils, bringing a high level of fertility and making these soils among the most productive in the world. The added volcanoclastic material varies in thickness, depending on many factors: volume of the ejecta, wind direction and velocity, and proximity to the source. Areas throughout the world that have received increments of airfall tephra display buried evidence associated with lithologic discontinuities. Although this phenomenon is widespread, there is a lack of information on the processes occurring when a new layer of fresh material is added to an existing soil. Dynamic pedology is the most suitable approach, because it provides an instantaneous record of the events and processes induced by the addition of tephra. Soils of the Pacific Northwest, USA, have been subjected, throughout the Holocene, to deposition of tephra produced by the volcanoes that dot the western part of the states of Washington, Oregon, and northern California. As recently as May 18, 1980, the eruption of Mount Saint Helens, western Washington, USA, provided an estimated 1.1 km$^3$ of airfall tephra that reached the soils along the fallout plume (**Figure 9**). In connection with this eruption, an experiment has been conducted by adding fresh tephra to soils that had received airfall tephra during the Holocene but had escaped the eruption of 18 May. Three metric tons of clean tephra have been

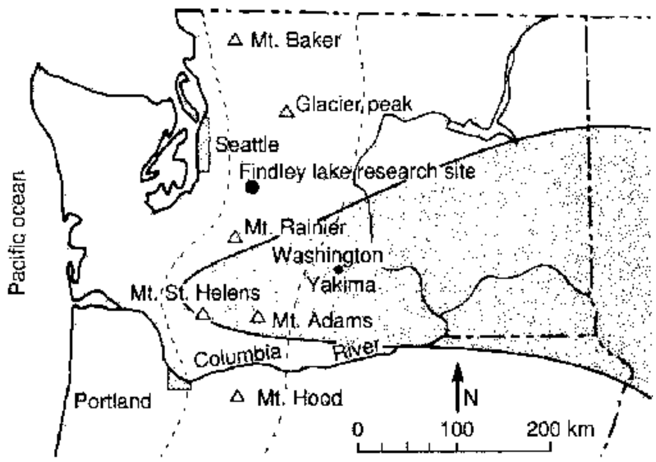


**Figure 9** The shaded area indicates the approximate distribution of the 1980 Mt. St. Helens tephra. The location of the study, Findley Lake, Washington Cascades, is marked. (Adapted from Ugolini FC, Dahlgren RA, LaManna J, Nuhn W, and Zachara J (1991) Mineralogy and weathering processes in Recent and Holocene tephra deposits of the Pacific Northwest, USA. *Geoderma* 51: 279, with permission.)

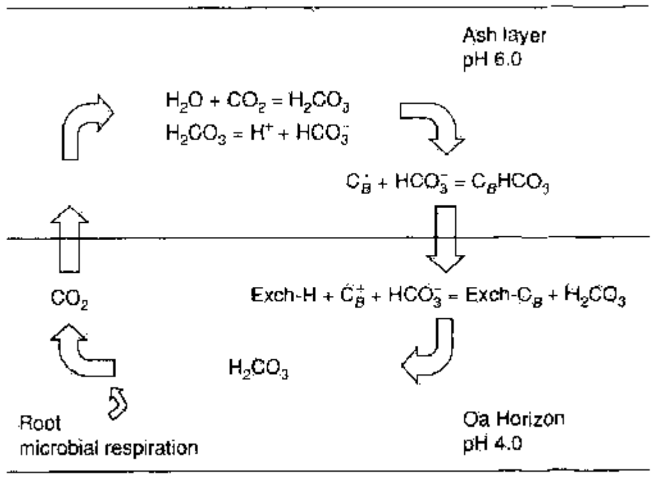


**Figure 10** The $CO_2$–$H_2CO_3$–$HCO_3$ weathering and transport cycle between the added ash layer and the buried organic horizon (Oa). $C_B$, base cation. (Adapted from Dahlgren RA, Ugolini FC, and Casey WH (1999) Field weathering rates of Mt. St. Helens tephra. *Geochimica et Cosmochimica Acta* 63: 596, with permission.)

collected and applied in 5- and 15-cm increments to plots to be compared with those that had not received tephra. Hence a modern analog has been created of past burial episodes. The applied tephra contains soluble salts, mostly sulfates of Ca and Na formed while the ash was still airborne. These salts are readily leached once the tephra is exposed to the rainy and cool climate of the central Cascades, where the experiments have been carried out. The area receives 2300 mm of precipitation, two-thirds in the form of snow, and experiences a mean annual temperature of 5.5°C. The basic cations present in the leachates upon entering the buried forest floor displace $H^+$ and $Al^{3+}$. Consequently, solute entering the mineral soil (E horizon) have a low pH and a high concentration of soluble Al. Biological uptake, adsorption, and exchange reactions attenuate the pulse of basic cations. Calcium and $K^+$ are preferentially retained in the E horizon; $Mg^{2+}$, $Na^+$, and $Cl^-$ are quickly removed from the solum. Sulfate is retained in the B horizon owing to ligand exchange reactions with imogolite/allophane material enhanced by the lowering of the pH. As the input from the fresh tephra of $SO_4$ decreases, most of the adsorbed $SO_4$ is desorbed from the B horizon. Al–bearing secondary minerals apparently control Al concentrations in the leachates collected below the B horizon. This *in vivo* experiment has revealed the mechanism of weathering in the newly added tephra layer. Carbonic acid is the major proton donor that liberates basic cations and Si and provides the conjugate base for their transport. Upon percolating through the acidic forest floor, the $HCO_3$ gains a proton and is converted to $H_2CO_3$. At the low pH of the forest floor (approx. 4.0), volatilization of $HCO_3$ occurs, with the consequent diffusion of $CO_2$ upward into the tephra layer, where the pH (approx. 7.0) favors the formation of $H_2CO_3$, thus initiating a new weathering/transport cycle (**Figure 10**). This weathering scenario was observed during the first 4 months of monitoring the soil solution, but the process continued for 4 years. The information collected has allowed an estimation of field weathering rates, as well as of the weathering of the solid phase, after 10 years of exposure to the cool humid climate of the Cascades Mountains.

## Testing New Theories of Soil Formation

From 1980 the classic theory of podzolization, also commonly known as the 'fulvate theory,' has been challenged by Farmer and his coworkers (**Figure 11a**). The fulvate theory holds that Al and Fe mobilization and transport occur through complexation with soluble organic acids. The new theory proposes that the migration of Al from the E horizon of Podzols occurs as positively charged, soluble hydroxyaluminosilicate sols called 'protoimogolite.' They are adsorbed/precipitated in the Bs and eventually transformed into imogolite (**Figure 11b**). Support for this process is the repeated presence of imogolite in the Bs horizons of Podzols/Spodosols derived from a variety of parent materials. However, in the absence of laboratory or field experiments that show the migration of protoimogolite from the E horizon, its presence in the Bs does not insure that it is derived from the E. It could have formed *in situ* in the Bs. Furthermore, previous experiments with aqueous solutions containing organic matter have shown the ability of

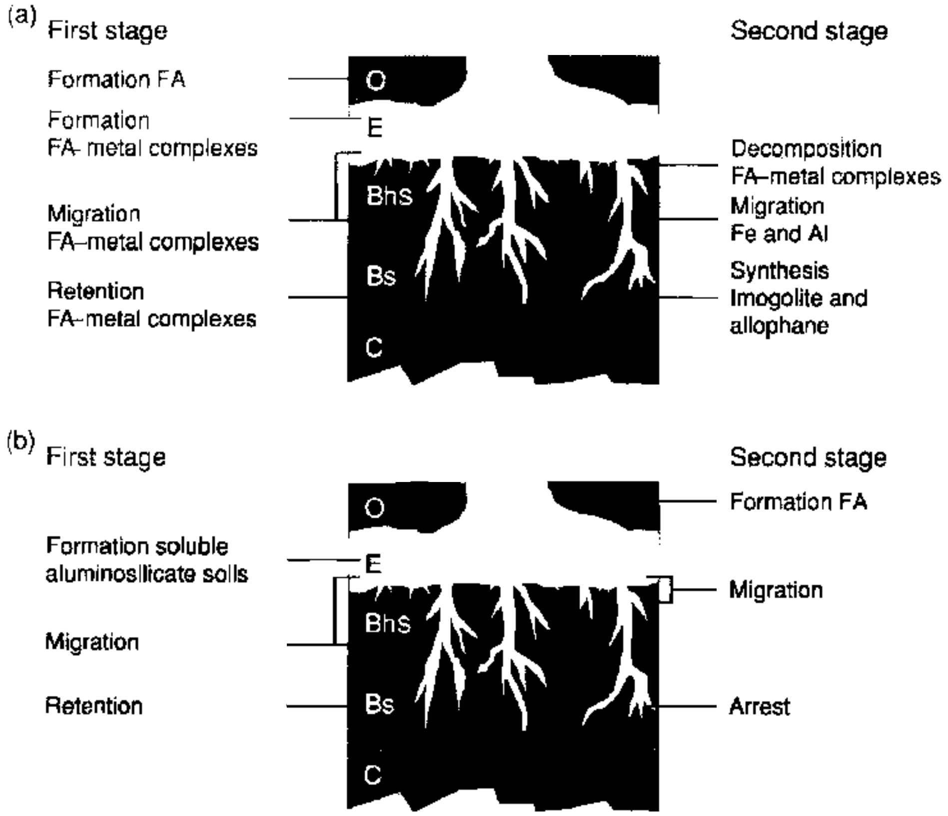


**Figure 11** Schematic representation of the theories of podzolization: (a) fulvate theory; (b) protoimogolite theory. (Adapted from Ugolini FC and Dahlgren RA (1987) Mechanism of podzolization as revealed by soil solution studies. In: Righi D and Chauvel A (eds) *Podzols et Podzolization*. INRA, p. 196, with permission.)

these compounds to complex and mobilize Al and Fe. Also, soil solutions collected by lysimeters from a subalpine Podzol/Spodosol have evidenced the role of soluble organic matter in the process of podzolization. Nevertheless, to establish the validity of the proposed 'protoimogolite theory,' at least for the Pacific Northwest Podzols/Spodosols, it is crucial to answer a number of questions, such as: Does the Al migration occur in organic or inorganically complexed forms? Does protoimogolite migrate in the soil solution? Where is imogolite found in a subalpine Podzol/Spodosol? Is imogolite thermodynamically stable within the soil profile under the current weathering regimen?

To answer these questions it is necessary to collect soil solutions from the genetic horizons of a Podzol/Spodosol and to have available a method to fractionate the Al so as to establish the forms of this metal. Podzol/Spodosol profiles, located in the Cascades Mountains of Washington (USA), have previously been equipped with lysimeters and intensively studied. Soil solutions have been collected seasonally from this subalpine Podzol/Spodosol except in the summer because of low precipitation. Al fractionation is performed according to the technique proposed by Driscoll. This procedure partitions Al into three operationally defined fractions: strongly bound Al complexes, nonlabile monomeric Al, and labile monomeric Al. The labile monomeric Al includes aqueous Al as well as inorganic complexes. The $Al^{3+}$ is estimated by speciating the labile monomeric fraction using GEOCHEM. The nonlabile monomeric fraction represents weakly held Al–organic complexes, while the strongly bound Al complexes appear to be a combination of colloidal Al or strong Al–organic complexes.

The first question is: In what form(s) does Al migrate from the E horizon? The answer is that between 73 and 100% of Al is organically complexed, mainly with the nonlabile monomeric form (**Figure 12**).

The second question asks whether protoimogolite migrates in the profile. The finding that Al in solution is organically complexed diminishes the role of inorganic Al forms and therefore of protoimogolite in the Al migration. However, an additional test has been carried out. Assuming, according to Farmer, that protoimogolite sols are present in the soil solution leaving the E horizon, being positively charged, they should have been retained on a cation exchange resin and hence decreased the Si concentrations of the solutions that had passed through the resin. However, statistical analysis of Si before and after passage through the cation exchange resin shows no change in Si concentration. This indicates the absence of protoimogolite sols in the soil solution. Additionally, if the soil solution exiting the E horizon does contain protoimogolite, the value of the total Si, which includes also the Si in the aluminum silicate soils, should be higher than monomeric Si. However, the two values are the same; consequently, Si moves only in monomeric form and is not associated with Al. From these findings, apparently in these Podzols there is neither appreciable movement of inorganic Al nor migration of aluminum silicate soils.

To answer the question of whether imogolite is present in the Podzol under study, a search for this mineral has been made in all the inorganic horizons: E, Bhs, Bs, BC, and C. The methods used include selective dissolution, IR spectroscopy, scanning and transmission electron microscopy, and potentiometric titration. In the clay fraction, the minerals of short-range order represent less than 5% in the E horizon, while they range from 20 to 50% in the B and C horizons.

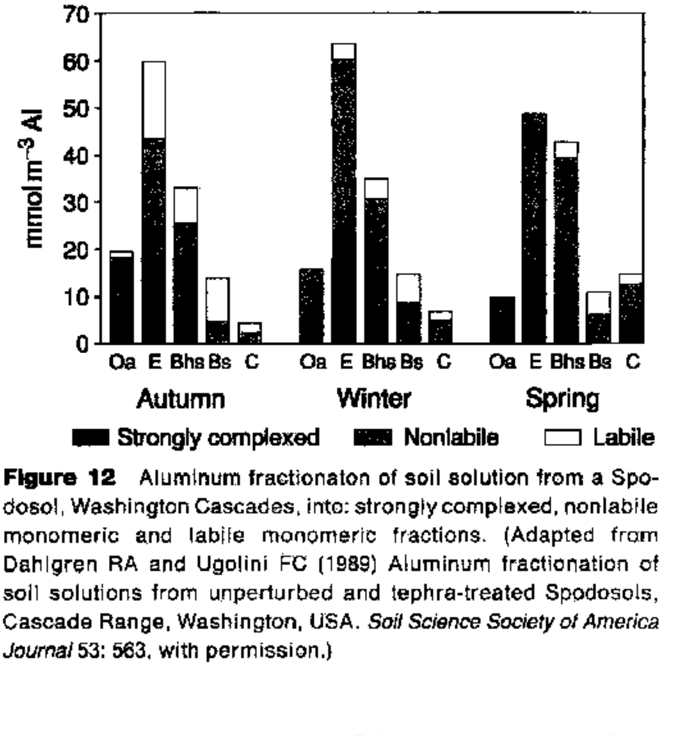


**Figure 12** Aluminum fractionaton of soil solution from a Spodosol, Washington Cascades, into: strongly complexed, nonlabile monomeric and labile monomeric fractions. (Adapted from Dahlgren RA and Ugolini FC (1989) Aluminum fractionation of soil solutions from unperturbed and tephra-treated Spodosols, Cascade Range, Washington, USA. *Soil Science Society of America Journal* 53: 563, with permission.)

The final question is: Is imogolite thermodynamically stable in the soil studied? More specifically, is imogolite stable in the E horizon, where it should form according to the protoimogolite theory of podzolization? Chemical equilibria are used to determine the stability of imogolite and other minerals (**Figure 13**). The calculated saturation indices show that soil solution in the E horizon is undersaturated with respect to imogolite; consequently, the imogolite sols cannot be stable in this horizon. Soil solution from the Bs and C horizons appears to be in equilibrium with the hydroxyaluminum interlayer of 2:1 clay minerals and imogolite. The hydroxyaluminum interlayer seems to control Al activities, while imogolite controls $H_4SiO_4$ activities. From this study it appears that Farmer's proposed theory of protoimogolite to explain the process of podzolization is not applicable in Podzols/Spodosols developed in cool to cold forested environments, where organic acids tend to dominate the upper part of the profile where protoimogolite should form. Al is effectively complexed by the DOM and therefore not available for reacting with Si to form aluminum silicate sols.

## Summary

Dynamic pedology, in the context of soil science, occupies a particular role. Based on collection, analysis, and interpretation of soil solution, dynamic pedology attempts to provide answers that traditional solid-phase investigations cannot give. It obtains an instantaneous picture of the ongoing processes and the impact of perturbations. However, a complete understanding of soil can be acquired by studying the liquid, the solid, and the gaseous phases. The adoption of dynamic pedology is restricted to those soils that experience percolative events, although collection of solution has been secured by irrigating the soil. Inherited limitations of dynamic pedology are obviously dependent on the collection of representative soil solution, which in turn is conditioned by the residence

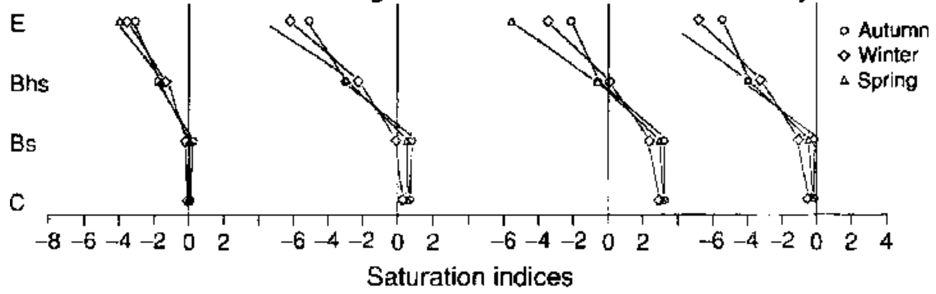


**Figure 13** Saturation index of selected minerals calculated for the autumn, winter, and spring soil solutions collected from a Spodosol, Washington Cascades. (Adapted from Dahlgren RA and Ugolini FC (1989) Formation and stability of imogolite in a tephritic spodosol, Cascade Range, Washington, USA. *Geochimica et Cosmochimica Acta* 53, 1901.)

time of the water in the soil. Solution collected by gravitational means, zero-tension lysimeters, or tension lysimeters may differ. Seasonal variations in solution with tension lysimeters held at a constant of 10-kPa tension are few compared with solutions obtained with low (0–30 kPA)- and high-speed centrifuge (30–3000 kPa). Also the seasonal changes in solution composition determine the thermodynamic stability of minerals. Limitations in the use of dynamic pedology are therefore related to the sampling methodology and the time of sampling of the soil solution.

## List of Technical Nomenclature

| | |
|---|---|
| **Concentrations:** | |
| **μmol** | Micromoles |
| **$\mu mol_c\, l^{-1}$** | Micromoles of charge per liter |
| **$mg\, l^{-1}$** | Milligram per liter |
| **$mmol\, m^{-3}$** | Millimoles per cubic meter |
| **Length:** | |
| **μm** | Micrometer |
| **cm** | Centimeter |
| **km** | Kilometer |
| **m** | Meter |
| **mm** | Millimeter |
| **Pressure:** | |
| **kPa** | Kilopascal |
| **Temperature:** | |
| **°C** | Degrees Celsius |
| **Volume:** | |
| **l** | Liter |

## Further Reading

Dahlgren RA and Ugolini FC (1991) Distribution and characterization of short-range-order minerals in Spodosols from the Washington Cascades. *Geoderma* 48: 391–413.

Dawson HJ, Ugolini FC, Hrutfiord BF, and Zachara J (1978) Role of soluble organics in soil processes of a Podzol, central Cascades, Washington. Soil Science 126: 290–296.

Driscoll CT (1984) A procedure for the fractionation of aqueous aluminum in dilute acid waters. *International Journal of Environmental Analytical Chemistry* 16: 267–284.

Duchaufour P (1982) *Pedology*. (Translated by Paton TR.) London, UK: Allen and Unwin.

Johnson-Maynard JL, McDanial PA, Ferguson DE, and Falen AL (1998) Changes in soil solution chemistry of Andisols following invasion of bracken fern. *Soil Science* 163: 814–821.

Likens GE, Bormann FH, Pierce RS, Eaton JS, and Johnson NM (1977) *Biogeochemistry of a Forest Ecosystem*. New York: Springer-Verlag.

Malcom RL (1990) Variations between humic substances isolated from soils, streams waters, and groundwaters as revealed by $^{13}$C-NMR spectroscopy. In: MacCarthy P *et al.* (eds) *Humic Substances in Soil and Crop Sciences: Selected Readings*, pp. 13–35. Madison, WI: American Society of Agronomy Soil Science Society of America.

Mattigod SV and Esposito G (1979) Chemical modeling of trace metal equilibria in contaminated soil solutions using a computer program GEOCHEM. In: Jenne EA (ed.) *Chemical modeling of aqueous systems*, pp. 837–856. Washington, DC: American Chemical Society Publications.

Nuhn WW (1987) *Soil Genesis on the 1980 Pyroclastic Flow of Mt. St. Helens*. MS thesis, College of Forest Resources. Seattle, WA: University of Washington.

Pearson RW (1971) Introduction to symposium – the soil solution. *Soil Science Society of America Proceedings* 35: 417–420.

Schnitzer M and Kodama H (1977) Reactions of minerals with soil humic substances. In: Dixon JB and Weed SB (eds) *Minerals in the Soil Environment*, pp. 741–770. Madison, WI: Soil Society of America.

Schwab AP (2000) The soil solution. In: Sumner ME (ed.) *Handbook of Soil Science, sect B, Soil Chemistry*, vol. 3, pp. 85–122. New York: CRC Press.

Singer M, Ugolini FC, and Zachara J (1978) In situ study of podzolization on tephra and bedrock. *Soil Science Society of America Journal* 42: 105–111.

Stevenson FJ (1994) *Humus Chemistry, Genesis, Composition, Reactions*, 2nd edn. New York: Wiley.

Ugolini FC and Sletten RS (1991) The role of proton donors in pedogenesis as revealed by the soil solution studies. *Soil Science* 151: 59–75.

Ugolini FC, Zachara JM, and Reanier RE (1982) Dynamics of soil-forming processes in the Arctic. *Permafrost*. pp. 103–115. The 4th International Conference Proceedings of National Research Council of Canada, Ottawa. Ottawa, ON: NRC Research Press.

Ugolini FC, Dahlgren RA, Shoji S, and Ito T (1988) An example of andosolization and podzolization as revealed by soil solution studies, S. Hakkoda, NE Japan. *Soil Science* 145: 111–125.

Zabowski D and Ugolini FC (1992) Seasonality in the mineral stability of a subalpine Spodosol. *Soil Science* 154: 497–507.

# PEDOMETRICS

**I O A Odeh and A B McBratney**, University of Sydney, Sydney, NSW, Australia

## Introduction

The term 'pedometrics,' first coined by McBratney in the late 1980s, 'is a neologism derived from the Greek roots pedos or *πεδοσ* (soil) and metron or *μετρον* (measurement).' It is used in a similar fashion to other words such as biometrics, psychometrics, econometrics, chemometrics, and, the oldest of all, geometry. Furthermore the term covers two main ideas – the 'soil' or 'pedo' part, which corresponds to that branch of soil science we call pedology, and the 'metrics' part, which is restricted to mathematical and statistical methods. Simply defined, pedometrics is the use of quantitative methods for the study of soil distribution and genesis, and as a sustainable resource. Another problem-oriented definition is 'soil science under uncertainty.' In this sense pedometrics deals with uncertainty in soil models that describe deterministic or stochastic variation, vagueness, and lack of knowledge of soil properties and processes (**Table 1**). Thus mathematical, statistical and numerical methods could be applied to resolve the uncertainty and complexity inherent in a soil system model, including numerical approaches to classification, which deals with supposedly deterministic variation.

Pedometrics is not new, as mathematical and statistical methods have been applied to soil studies since the 1960s and 1970s. However, it is now emerging as a technical branch of soil science complementing traditional pedology. Over time the use of computers has increased in both fields, and the difference between the two has decreased and, in some cases, overlapped. Due to new demands for quantitative soil information required for global-scale models, regional environmental planning, and field-scale agricultural land management, traditional pedology has become more quantitative through the increased use of computerized soil information systems. Concurrently, pedometrics has emerged as a collection of quantitative tools, which are increasingly being used to account for conceptual pedological models of soil variation. As of 2003, there is a strong and growing overlap and synthesis between 'traditional' pedology and quantitative pedology or pedometrics (as shown in **Figure 1**).

## Methods and Problems

### Pedometrics: Principles and Hypotheses

As the definition implies, pedometrics encompasses the whole quantitative approach to the study and description of soil, especially soil in the field. It also involves the use of mechanistic, stochastic, deterministic, and empirical models (**Table 1**). A more restricted form of pedometrics is one that involves applications of statistical and probabilistic modeling to soil variation. Pedometrics principles are therefore embodied in (geo-)statistical modeling of pedological processes and the associated phenomena as they influence soil variation, and the associated model uncertainty.

Pedometrics is therefore aimed at resolving a number of questions or hypotheses:

1. What is the pattern of soil distribution in multivariate space, i.e., soil taxonomy?
2. How does the soil vary spatially and temporally?
3. What are the utility and quality of soil?
4. What are the causal factors of soil, i.e., soil genesis?

**Table 1** Types of quantitative models used for soil studies with examples and sources of uncertainty

| *Causality/ uncertainty* | *Model* | | |
|---|---|---|---|
| | *Deterministic* | *Stochastic* | *Nonstatistical* |
| Empirical | Generalized linear models, numerical taxonomy, Jenny functional relations and canonical ordination | Time series, spatial processes, temporal and spatial variation of soil properties | Fuzzy systems |
| Mechanistic | Flow and diffusion of soil plasmic materials, profile and landscape development | Indeterminate | Indeterminate |
| Uncertainty | Imprecise measurements, model uncertainty | Random process, probability | Vagueness, ambiguity and fuzzy geostatistics |

Source: McBratney AB and Moran CJ (1986) *Pedometric Research. CSIRO Australia Division of Soils Research Report*. Canberra, ACT: CSIRO Division of Soils

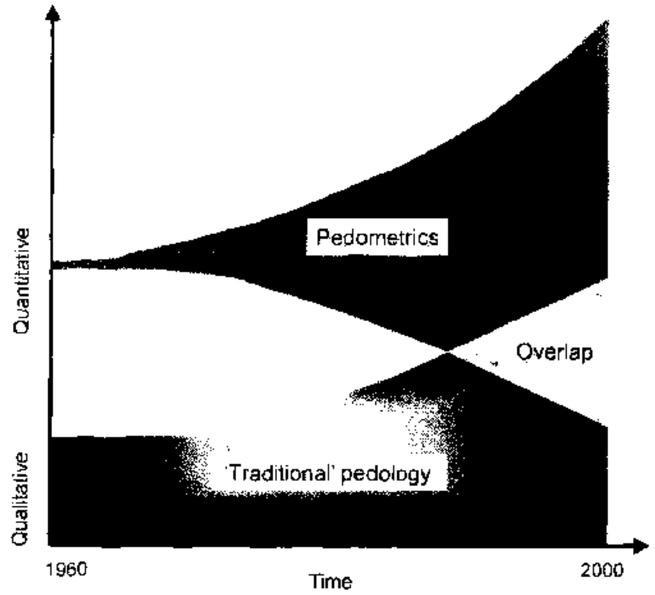


**Figure 1** A time line of the growth of pedology and pedometrics. (Source: McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* 97: 293–327.)

5. What is the uncertainty of modeling the spatial pattern of soil phenomena, i.e., is a model a true representation of reality?

With the exception of (5), these hypotheses dovetail well with the requirements of conventional pedology. Although in the past developments in pedometrics have been restricted to statistics and probability, increasingly mechanistic models are being applied to model soil processes.

### Soil Measurements and Properties

In order to model soil quantitatively, we need to perceive it, i.e., sense it or measure some quality or quantity of it. In other words, how do we numerically represent soil properties in order mathematically and/or statistically to model their variation? Ever since the advent of modern environmental science, technological advances have improved immensely the way we characterize or measure the quality and/or quantity attributes of soil. Soil measurement techniques include direct measurements, visual observations, and remote sensing (including proximal, airborne, and space-based modes of sensing).

**Soil color and spectroradiometry** Measurement of soil color has gained tremendously from advances in technology, particularly in spectroradiometry. The human eye is capable of seeing colors made up of combinations of reflected wavelengths throughout the visible portion of the electromagnetic spectrum. The Munsell System of Color Notation, a color order system invented by Albert H. Munsell in 1905, is the most-used soil color system prior to the development of a quantitative means of measuring soil color. It specifies a limited number of 'standard' colors, within an asymmetrical, 'cylindrical' color space. While the hue (H) describes the similarity to a dominant color (red, yellow, and blue) or a combination of any two, the value (V) is the color intensity (lightness) and chroma (C) is the relative purity of the dominant wavelength (saturation). The Munsell color notations are semiquantitative, requiring further transformation for quantitative data analysis.

In 1931 the CIE, or Commission Internationale de l'Eclairage, standardized color order systems by specifying the light source, the observer, and the methodology used to derive the values for describing color. The CIE system defines how to map a spectral power distribution to a set of tristimulus coordinates which describe the color space: a luminance component $Y$ and two other 'imaginary' components of the primary spectra, $X$ and $Z$ (**Figure 2a**). These in turn could be mapped to $xy$ coordinates of chromaticity (**Figure 2b**). More recently, in 1978 the CIE has provided additional transformations (CIELUV and CIELAB color spaces) to make color representation on a continuous numeric color space more amenable for pedometric analysis.

Optical properties of the soil are primarily a function of its mineral composition, since the soil is formed from the transformation of rocks. A soil reflectance spectrum is therefore superimposition of spectra of the soil mineral components. Thus a soil type has its unique spectral reflectance curve. *In situ* or reference data are often collected at the time of image acquisition. One form of reference data is the ground-based measurement of the reflectance of surface features to determine their spectral response patterns. This might be done in the laboratory or in the field using a spectroradiometer. This device measures, as a function of wavelength, the energy coming from an object within its view. It could be used to prepare spectral reflectance curves of many soil types for classification or comparison purposes (e.g., **Figure 3**). Pedometricians are currently investigating the predictive power of such spectra. Such analysis is well known in chemometrics.

**Field properties: hard and fuzzy descriptors?** Other than color, many soil morphological properties, required in routine soil surveys, are described in somewhat vague terms. For example, soil structure grade is described as 'structureless' or 'weak' or 'moderate' or 'strong' in many national soil survey handbooks. To make the grades more amenable to quantitative

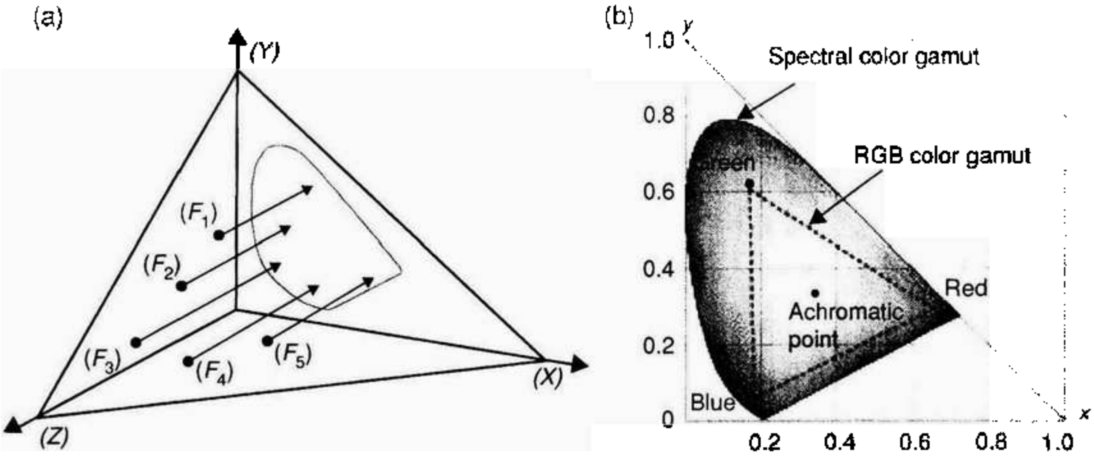


**Figure 2** (see color plate 49) (a) Mapping of *XYZ* tristimulus values to (b) an *xy* chromaticity.

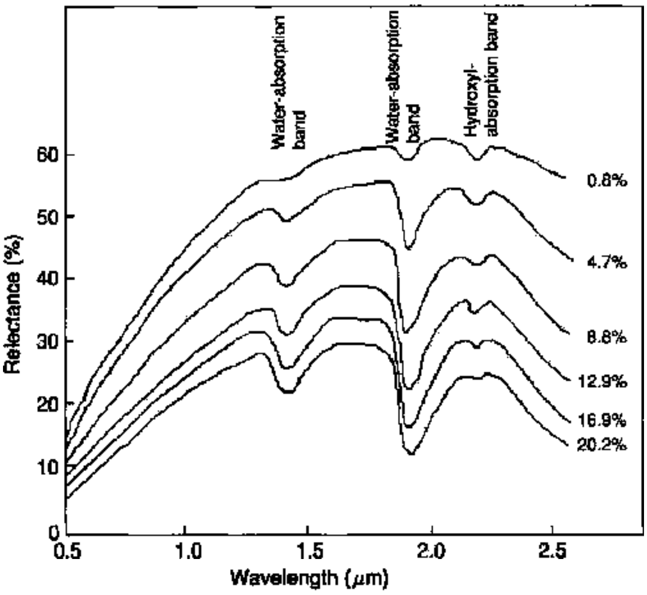


**Figure 3** Spectral reflectance curves for Newtonia silt loam at various moisture contents. (Adapted from Bowers SA and Hanks RJ Reflection of radiant energy from soil. *Soil Science* (1965) 100: 130–138.)

**Table 2** Fuzzy coding of structural type

| *Type* | *Horizontality* | *Verticality* | *Flatness* | *Accommodation* |
|---|---|---|---|---|
| Platy | 1.0 | 0.1 | 1.0 | 1.0 |
| Lenticular | 1.0 | 0.3 | 0.3 | 1.0 |
| Prismatic | 0.2 | 1.0 | 1.0 | 1.0 |
| Columnar | 0.2 | 1.0 | 0.9 | 0.9 |
| Angular blocky | 1.0 | 1.0 | 1.0 | 1.0 |
| Subangular blocky | 0.7 | 0.7 | 0.5 | 0.5 |
| Granular | 0.2 | 0.2 | 0.0 | 0.1 |
| Massive | 0.0 | 0.0 | 0.0 | 1.0 |
| Single grain | 0.0 | 0.0 | 0.0 | 0.0 |

Reproduced with permission from Odeh IOA, McBratney AB, and Chittleborough DJ (1991) Elucidation of soil–landform interrelationships by canonical ordination analysis. *Geoderma* 49: 1–32.

analysis (e.g., multivariate statistical analysis) these terms are usually coded as 0, 1, 2, and 3, respectively. But because of the vagueness in the linguistic characterization, fuzzy coding, which accounts for uncertainty or lack of clear boundary between the grades, has been used. Other morphological descriptors such as soil aggregate (structure) size classes, e.g., 'fine,' 'medium,' and 'coarse,' may be given linguistic variables which could be fuzzified to determine the degree of truth that a given soil layer is characterized by. Simple examples of fuzzy coding are presented in Table 2.

**Pedotransfer functions** The term 'pedotransfer function' (PTF) is generally defined as translating the raw soil data into more useful information. It can also be defined as predictive functions of certain soil properties from other easily, routinely, or cheaply measured properties. The most readily available data come from soil survey, such as field morphology, texture, structure, and pH. PTFs add value to the basic soil data by translating them into predicates of other more laborious and expensively determined soil properties. These functions fill the gap between the available soil data and the properties, which are more useful or required for a particular model or quality assessment. A simple illustration of how to apply PTFs to practical problems is shown in **Figure 4**.

## Quantitative Soil Geography – Classification Models

### Soil Classification and Soil Map Units

The pattern of soil in the landscape is as a result of effects of soil-forming factors, namely: climate, organisms, parent material, topography, and time

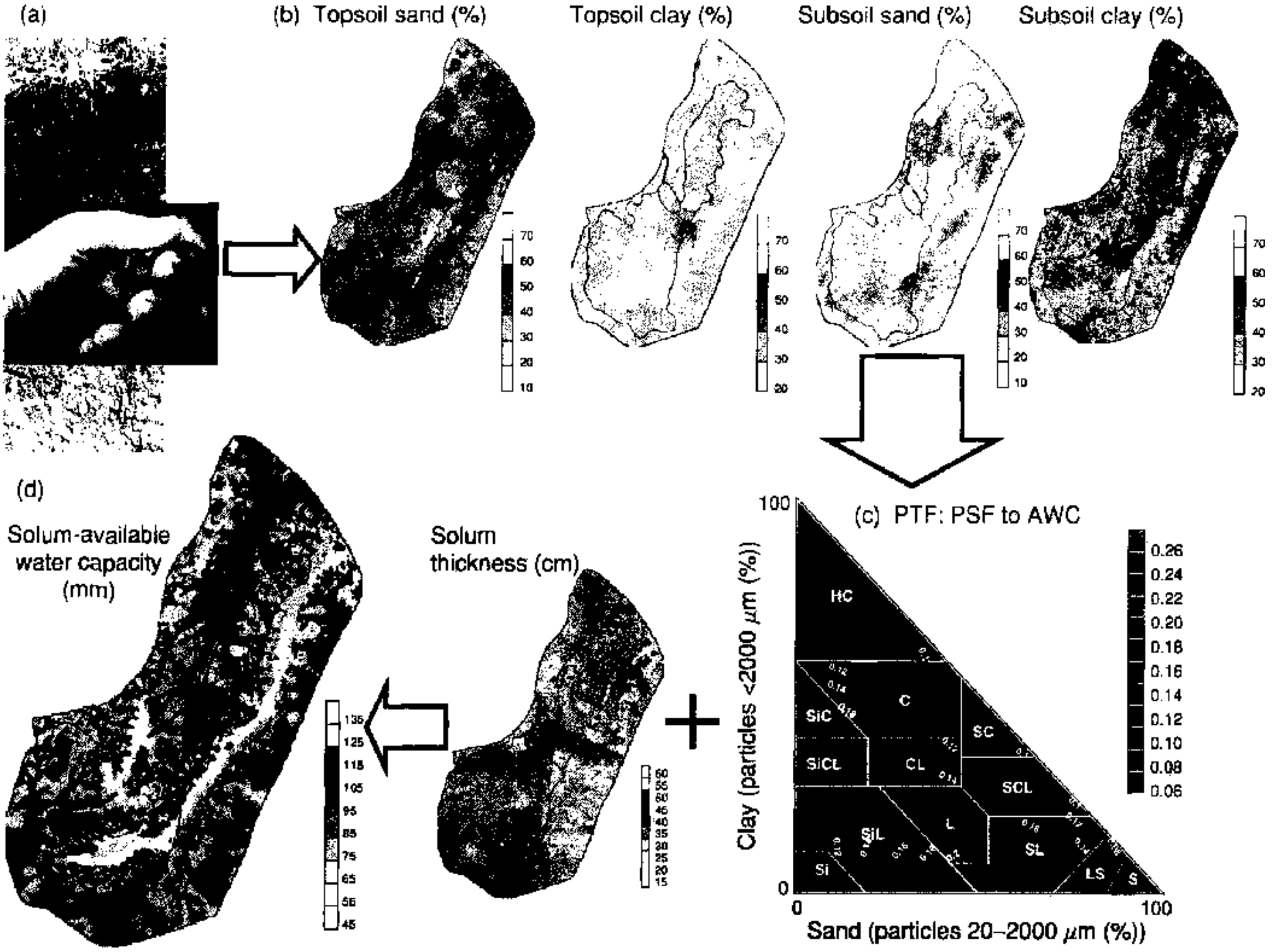


**Figure 4** (see color plate 50) A simple example of pedotransfer functions (PTFs): (a) field texture was first determined by 'feel' method for topsoil and subsoil at each sample location; (b) particle-size fractions were then estimated from the textural classes and interpolated on to a fine grid; (c) horizon-available water capacity was in turn estimated through a PTF expressed as a function of particles sizes; (d) a combination of both horizons produces the solum-available water capacity. AWC, available water capacity.

(usually referred to as 'corpt,' but recently extended to 'scorpan' – see Eqn [2], below). These factors, particularly climate and parent material, produce broad patterns of soil in geographic space. The complex combination of these factors is what causes repetitive patterns of soil in the landscape, which form the basis for soil classification, identification, and mapping. In traditional soil classification, both the conceptual and real classes are qualitative and have no clear-cut boundary. Pedologists have, in the past, attempted to avoid these by applying numerical classification to complex soil data. Soil classification, in both geographical and taxonomic space, is a simple representation of complex, and sometimes repetitive patterns of soil in the landscape. Class identification, on the other hand, is aimed at matching classes that are conceptually described (usually in a classification system) with reality – represented by data obtained by soil morphological description and measurements.

A soil map unit is a collection of areas that are relatively homogeneous in soil constitution or miscellaneous areas that constitute different soil components or both. In a given survey region, a map unit differs in some respect from all others and can be uniquely identified on a soil map. A soil map delineation, on the other hand, is an individual area on the map, bounded by other delineations or the map boundary. Usually delineation contains the dominant components in the map unit name, but it may not always contain a representative of each kind of inclusion. Soil boundaries can seldom be shown with complete accuracy on soil maps; hence parts and pieces of adjacent polypedons are inadvertently included or excluded from delineations. This problem has been the stimulus for paradigm shifts in soil classification in the latter part of the 1990s; hence the applications of numerical hard and fuzzy classification systems in soil science.

**Fuzzy classes and the breakdown of the classification paradigm** Traditionally most classification systems are composed of mutually exclusive classes in order to conform to the discontinuous soil variation embedded in the traditional concept of soil map units. But soil variation is more continuous than discrete. The pioneer work in pedometrics, involving computer-based numerical classification, was designed to address this limitation among others. While the applications of numerical soil classification to soil studies are, to some extent, based on continuous representation of soil in space, their results are still interpreted in terms of 'hard' classes. There is also lack of any spatial coherence for the classes to be mapped. Recent advances are based on fuzzy sets for optimized prediction quality of the resulting classification, and which take cognizance of the continuous nature of soil variation and nonlinearity in the interattribute relationships.

The first application of fuzzy set theory in soil survey was principally for classification. Two different but complementary approaches to grouping individuals into fuzzy classes in soil science are: (1) fuzzy $c$-means (FCM, also known as 'fuzzy $k$-means'), and semantic import model (SI).

The FCM algorithm for improved predictive classification by providing for membership to an extragrade class is based on the objective function, defining the within-class sum-of-square errors, $J_{\mathrm{E}}$, expressed as:

$$J_{\mathrm{E}}(M,c) = \alpha \sum_{i=1}^{n} \sum_{j=1}^{c} m_{ij}^{\phi} d_{ij}^{2} + (1-\alpha) \sum_{i=1}^{c} m_{i*} \sum_{j=1}^{c} d_{ij}^{-2} \quad [1]$$

where $c$ is the number of classes, $n$ is the number of individuals or pedons; $m_{ij}$ is the membership of an individual $i$ in class $j$; $\varphi$ is the fuzziness exponent $(1 < \varphi < \infty)$; $d_{ij}$ is the character space between the feature value of an individual, $i$, and the feature centroidal value for class $j$; $\alpha$ is the parameter that determines the mean value of $m_{i\cdot}$, which is the membership value of an individual, $i$, in the extragrade class.

The SI model was developed primarily for land evaluation. The need for using fuzzy set theory in land evaluation, such as defined in the Food and Agriculture Organization of the UN (FAO) framework, arises because basic soil information used for land evaluation is mostly described by seemingly vague terms such as 'poorly drained,' 'slightly susceptible to soil erosion,' and 'moderate nutrient availability.' Not even when these terms are defined precisely is the qualitative ambiguity removed. Usually, the land evaluator's aim is to produce a set of clearly defined classes of land qualities based on specified land use requirements. These subsequently provide the means of transferring information about the soil and its use. As land qualities are complex attributes that are derived from land characteristics such as topography, soil, water, or biological and human activity, subsequent Boolean logical operations in the process of land evaluation tend to throw away much useful information.

### Spatial Prediction of Soil Classes

**The corpt approach** As previously stated, the 'corpt' methods are based on the empirical-deterministic models that originated from Han Jenny's *Factors of Soil Formation*. Jenny's state-factor equation, in its extended form, is expressed as:

$$S = f(s, c, o, r, p, a, n) \quad [2]$$

where $S$ is some soil properties or soil type as a function of state factors: $s$ as some other soil property at a point, $c$ as climate, $o$ as organisms, $r$ as relief, $p$ as parent material, $a$ as time or age, and $n$ as space or spatial position. Soil spatial variability is therefore considered as being causative realizations of the complex combinations of soil-forming processes as influenced by the soil-forming factors. The scorpan function (Eqn [2]) – in its original form, the corpt – earlier in the nineteenth century stimulated numerous studies, mainly quantitative prediction of soil attributes, which we shall treat in a later section. Until recently, because data on soil classes were nonquantitative, the scorpan methods were only restricted to predicting soil attributes measured on a continuous scale.

Models such as expressed in Eqn [2] can be derived for predicting soil classes or soil types using two different techniques: logistic regression and classification trees. Logistic regression is designed specifically for situations in which we have a dichotomous (nominal or ordinal) dependent variable (e.g., soil classes), in comparison to classical linear regression typically used for continuous dependent variables (e.g., soil pH value). Several problems arise when a dependent variable is binary: the error terms are nonnormal and their variance is nonconstant. Although the errors are not normal, this method still provides unbiased regression estimators that are approximately normal if the sample size is large. Classification tree algorithms, on the other hand, search for combinations of values of independent variables (e.g., Jenny's state factors or attributes derived from them) that best predict the value of the dependent variable (e.g., unordered factors such as soil classes). Prediction quality is based on criteria such as the within-group noises, variance, or statistical significance of the conditional frequency distribution of values for the dependent variable, conditioned on the answers

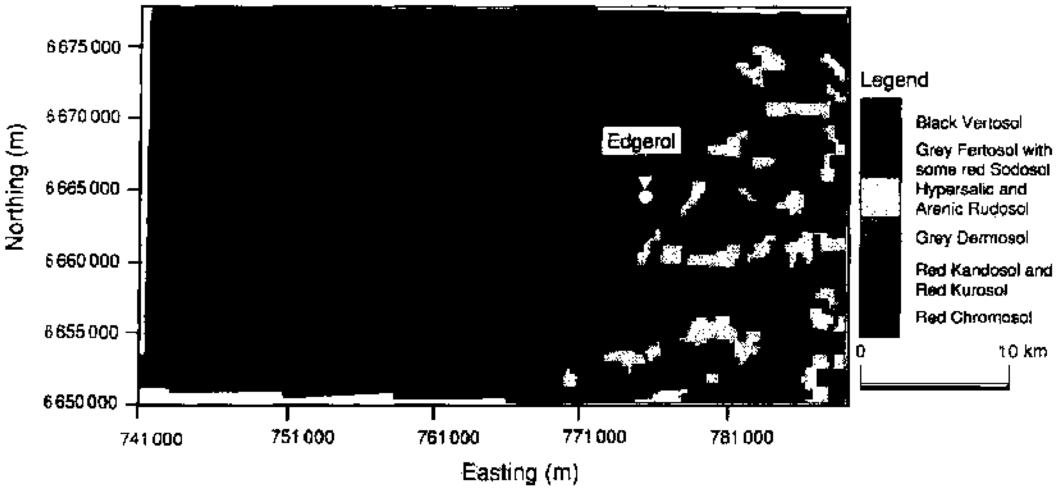


**Figure 5** (see color plate 51) Predicted soil classes for the Edgeroi area in accordance with Australian Soil Classification. (Reproduced with permission from McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* 97: 293–327.)

to questions asked. A digital map of soil suborder classes, as produced by classification tree model, is shown in **Figure 5**.

### Spatial Prediction of Soil Attributes

**The scorpan approach** The 'scorpan' approach (Eqn [2]) earlier in the nineteenth century stimulated numerous studies in quantitatively predicting soil attributes measured on a continuous scale. Much of the earlier studies, and indeed some recent examples, was based on general and bivariate-simple linear regression, although multiple polynomial regression models were occasionally applied. But due to nonlinearity in the correlations among many soil variables, and indeed of many of the soil variables with ancillary variables, robust methods such as generalized linear models (GLMs), generalized additive models (GAMs), and regression trees (RTs) have been developed and applied. Another development is the artificial neural network models, which are nonparametric modeling techniques that mimic the neural networks of the brain. The networks are composed of processing units, or neurons, which are organized into layers, i.e., input, hidden, and output layers.

But the problem is, while the classic models or the more robust methods may take care of the deterministic relations, they do not account for spatial autocorrelations of the soil properties, especially at the local level. To solve this problem, the pioneer pedometricians initiated the application of geostatistics (which was primarily developed for the mining industry).

**Geostatistics** Matheronian geostatistics is based on the theory of regionalized variables, which allows us to consider spatial variability of a soil property or even soil types, if quantified, as a realization of a random function represented by a stochastic model. The generic geostatistical method of spatial interpolation is termed 'kriging' in its various forms: simple, ordinary, lognormal, and disjunctive kriging.

In geostatistics, the semivariogram is a primary requirement for spatial prediction or kriging of a target geographical feature. The semivariogram can be obtained by processing a data set to produce empirical semivariances through the computation of the spatial correlation or covariance between pairs of samples at certain distances apart. The plot of the semivariance and the corresponding distance or lag at which the pairs are separated produce the semivariogram. The latter describes the magnitude, spatial scale, and the general form of variation of a given variable. An example of semivariogram is shown in **Figure 6**, with several spherical models based on different fitting methods.

The first major applications of kriging, in its ordinary and univariate form, for soil studies were in the early 1980s. Since then ordinary kriging has been widely used in various subfields of soil science, including soil reclamation, in soil classification, and soil pollution. Major limitations of the univariate geostatistics technique of kriging are due to the assumptions of stationarity, which are not often met by the field-sampled data sets and, of course, the often-cited requirement of large amounts of data to define the spatial autocorrelation. However, with increasing availability of ancillary information, the lack of adequate samples has been partially solved. The univariate usage of kriging is also limiting in situations of complex terrain where the soil-forming processes are themselves complex. In such situations there is the need to model both

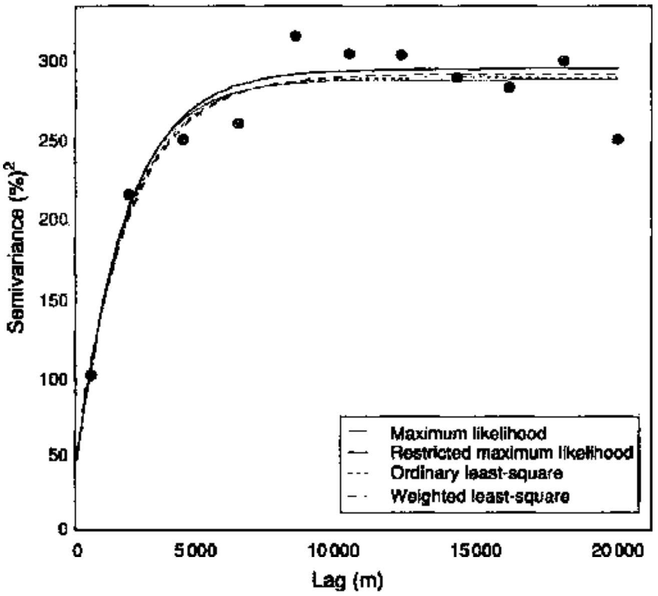


**Figure 6** An experimental semivariogram fitted with spherical models based on different fitting techniques.

the structured and the spatially dependent components of the soil variable. Also there are economic and logistic reasons for including the ancillary variables influencing soil variability, especially if the latter are more readily and cheaply available. As both the soil and the exogenous factors are multivariate, the most obvious choices are appropriate combinations of multivariate or univariate analysis using the scorpan factors and the geostatistical methods. These combinations constitute the hybrid techniques.

**The hybrid of scorpan and geostatistics** The hybrid techniques for soil survey and mapping are based on various combinations of the geostatistical and multivariate or univariate scorpan methods. Let us suppose that a data vector describing a soil property is a random variable Z, determined at locations in a region, $X = x_1 \ldots, x_N$, and consisting of three components as:

$$Z = m + Z_1 + \epsilon \qquad [3]$$

where $m$ is the local mean for the region, $Z_1$ is the spatially dependent component, and $\epsilon$ the residual error term, spatially independent. Now there may be situations where $m$ is varying and dependent on some exogenous factors such as the scorpan factors. In other words it is deterministically related to some causative factors (in geostatistics parlance, the variable is said to exhibit a trend). Wherever trend exists, the ordinary univariate kriging is inappropriate. Several methods have been designed to accommodate the trend.

Universal kriging has been the commonly used method to accommodate the trend or 'changing drift,' as it is sometimes known, in a soil variable. Universal kriging is a combination of the standard model of multiple-linear regression and the geostatistical method of ordinary kriging, which is also analogous to combining scorpan methods with the univariate kriging, but only using the geographic coordinates as determining the drift (the $n$ factor of the scorpan model). More recently, a more advanced approach, the intrinsic random function of order $k$ (IRF-k), has been used to accommodate the varying nature of the trend in a regionalized soil variable. The term '$k$' represents the order of polynomial trends: $k = 0$ means constant drift, and the IRF-k is equivalent to ordinary kriging system of equations. If $k = 1$, we have linear drift; $k = 2$ yields quadratic drift; but where there is no trend but deterministic relationships are with some known or readily available and inexpensive covariates (scorpan factors) or other easy-to-measure soil variables, co-kriging has played a major role in efficiently predicting the target soil variable. Universal co-kriging is also possible when considering the trend and covariation with one or more secondary variable.

Co-kriging is the multivariate extension of kriging that allows the inclusion of more readily available and inexpensive attributes in the prediction process. There are many instances in soil survey where the scorpan factors such as topography, time, and variable parent material, are easily discernible or are either readily available and/or are cheap to obtain. The most efficient way to predict the expensive-to-measure target soil variable, the variation of which is affected by the scorpan factors, is to use the factors in co-kriging the target soil variable, sampled at fewer locations, into dense grid nodes.

Regression-kriging (RK) is another hybrid method that combines either a simple or multiple-linear regression model (or a variant of GLM, GAM, and RTs) with ordinary, or simple, kriging of the regression residuals. The assumption here is that the deterministic component ($m$ in Eqn [3]) of the target (soil) variable is accounted for by the regression model, while the model residuals represent the spatially varying but dependent component $Z_1$ in Eqn [3]). If the exogenous variables used in the regression equation are available at more dense locations than the target variable, the equation can then be used to predict the $m$ on to those locations. The $Z_1$ can also be predicted to the same locations by simple kriging system of equations, and then added to the $m$ to obtain $Z^*$. A variant of RK is kriging with uncertainty, which introduces regression residuals (as representing model uncertainty) into the kriging system used to predict the target soil variable. This reduces the extrema of the target soil variable and therefore produces a

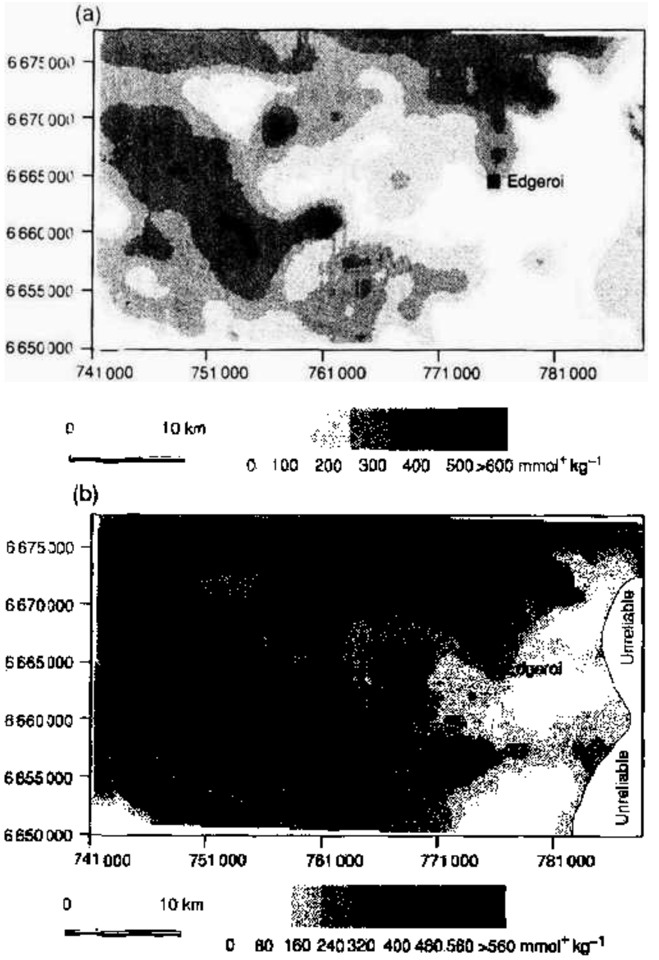


**Figure 7** Predicted cation exchange capacity (CEC; millimoles per kilogram) for the Edgeroi area using (a) kriging with elevation as external drift and (b) regression-kriging. (Reproduced with permission from McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* 97: 293–327.)

smoother function of the predicted values. Another variant of RK, kriging with external drift (KED), is a hybrid technique that integrates the universality conditions into the kriging system using one or more of the ancillary drift variables. It is similar to universal kriging, but using an ancillary variable to represent the trend. As shown in **Figure 7a**, a regional digital map of cation exchange capacity (CEC) has been produced using KED with elevation as the external drift. Comparing this map with one produced using RK (**Figure 7b**) indicates a slightly more smoothed map produced by KED than the RK one.

## Soil Utility and Quality

### Land Evaluation, Soil Quality, and Soil Utility

Pedometrics has played a major role in advancing the role of land evaluation and, more recently, the quantification of soil quality for land management and sustainable use of land resources. Although 'soil quality assessment' is often used as a misnomer for 'land evaluation,' both could be regarded as the interpretative phase of soil survey. While land evaluation is concerned with the assessment of land performance when used for specified purposes, soil quality is defined as 'the capacity of a specific kind of soil to function, within natural or managed ecosystem boundaries, to sustain plant and animal productivity, maintain or enhance water and air quality, and support human health and habitation.' In considering productivity, environmental quality, and health as major functions of soil, this definition requires that values be placed on specific soil functions as they relate to the overall sustainability of alternate land-use decisions. Pedometric techniques (based on 'hard' and/or fuzzy techniques) are being used to quantify soil quality and associated uncertainty. For example soil quality can be assessed in terms of requirement for different specific uses such as shown in **Figure 8**. Many of the soil quality indicators are required at various scales and spatial extents in which pedometrics is playing a major role in spatially modeling them for incorporation into regional, catchment, and field-scale modeling. Another approach to soil quality analysis may be based on the concept of pedodiversity.

## Recent Advances: Quantitative Pedogenesis

As stated above, soil variability is a function of factors of soil formation. Since the publication of Jenny's state factors, pedologists have attempted to develop pedogenetic (mathematical) functions that could explain soil variability and, in many instances, predict or even simulate the soil. A few examples are given here.

### Process Models

The soil system is dynamic and regenerative under the influence of soil-forming processes. Since the soil is important for the survival of human civilization, it is imperative that researchers' interest in soil processes is ever-increasing. This has resulted in models of pedogenesis based on chemical and physical processes. The formulation of the process models and their applications are dependent on the scales in both the temporal and spatial dimensions. This modeling space–time continuum is determined by the complexity of the modeling process, which can be characterized in several ways. The first is based on computational complexity of the model, which ranges

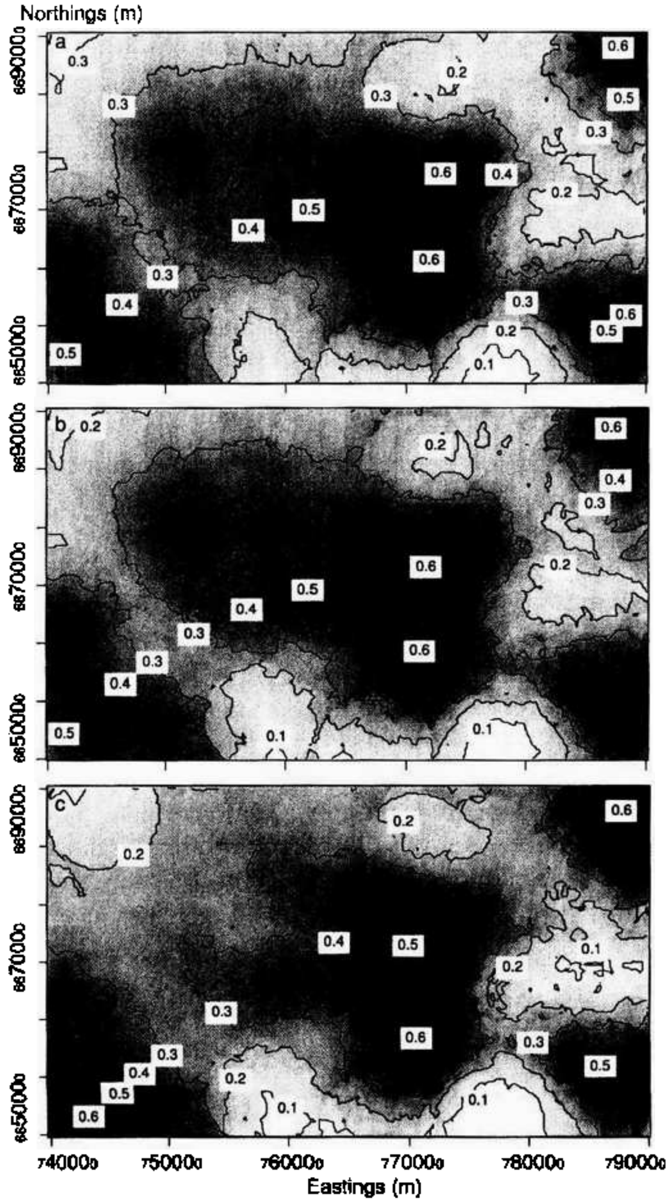


**Figure 8** Interpolated suitability of (a) wheat, (b) sorghum and oats, and (c) dryland cotton. (Reproduced from Triantafilis J, Ward WT, and McBratney AB (2001) Land suitability assessment in the lower Namoi Valley of Australia using a continuous model. *Australian Journal of Soil Research* 39: 273–290.)

from purely qualitative (mental models) to highly quantitative, with the latter involving complex computer coding; the second is based on the complexity of the model structure, which distinguishes between mechanistic (highly complex) and empirical (simplified functional) models; and the last, but not least, is based on organizational hierarchy, which determines at which level a model is used to simulate the soil system. The hierarchical levels range from $i-4$ (molecular-level processes) to $i+6$ (global level). Each level is a subsystem of the level above it, therefore allowing room for upscaling.

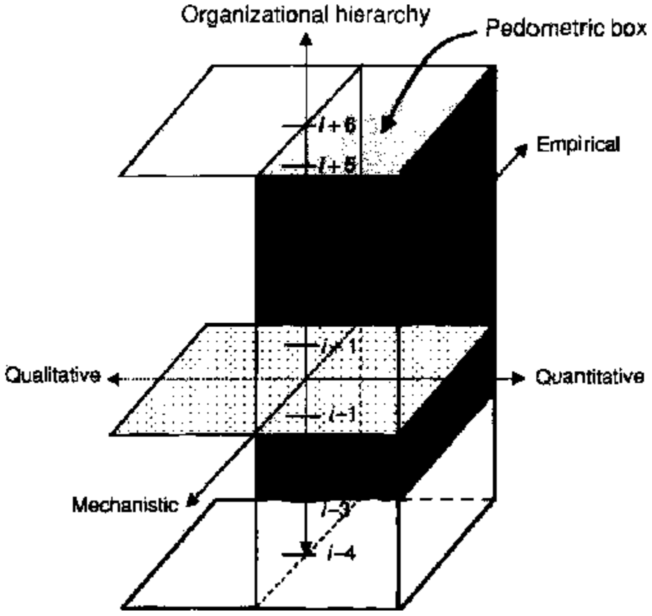


**Figure 9** Pedometric box in the organizational hierarchy of pedogenetic modeling approach. (Adapted from Hoosbeek MR and Bryant RB (1992) Towards quantitative modeling of pedogenesis – a review. *Geoderma* 55: 183–210, with permission.)

The three characterizations above are well illustrated in **Figure 9**. It should be noted that pedometrics has played a dominant role in the quantitative half of the organizational hierarchy, particularly from and above the $i-3$ level, which is governed by material fluxes in the pores between the primary particles.

Recent developments in process modeling focus on mechanistic stochastic simulations, particularly at levels $i-3$ to $i$ (**Figure 9**) and on how the latter can be upscaled to any of the higher hierarchies. For example: a mechanistic pedogenetic model, based on a digital elevation model (DEM), was used to simulate pedogenesis by a combination of several submodels: (1) physical weathering rate as exponential decline of soil thickness; (2) chemical weathering rate represented as a negative exponential function of both soil thickness and time; and (3) the movement of material as characterized by diffusion transport model. The upscaling result of such an analysis is illustrated in **Figure 10**, whereby, after 10 000 years, soil accumulation is predominant in the gullies compared with the ridges, where soil erodes.

Other emerging methods are geostatistical simulation models, e.g., indicator simulation, Gaussian simulation, etc. However, with increasing environmental concern, the development and applications of landscape dynamic models under different land-use scenarios may well dominate pedometric research in the first few decades of the twenty-first century.

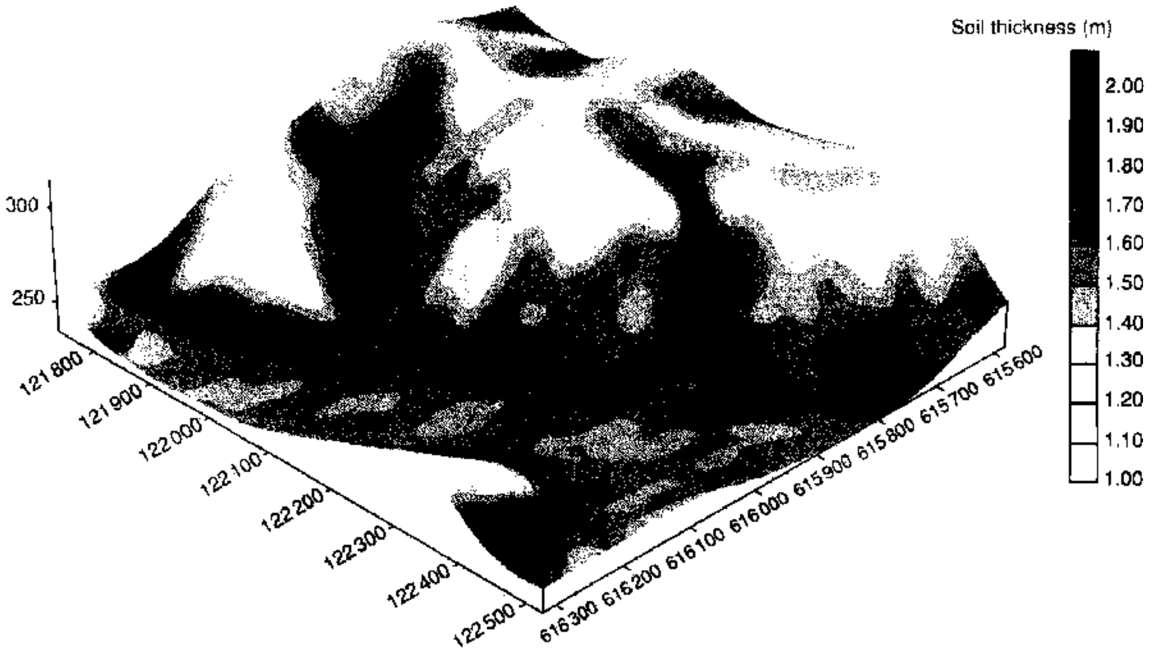


**Figure 10** Soil formation in a landscape after 10 000 years. (Adapted from Minasny B and McBratney AB (2001) A rudimentary mechanistic model for soil formation and landscape development: II. A two-dimensional model incorporating chemical weathering. *Geoderma* 103: 161–179, with permission.)

*See also:* **Classification of Soils**; **Pedology**: Basic Principles

## Further Reading

Addiscott TM (1998) Modeling concepts and their relation to the scale of the problem. In: Finke AF, Bouma J, and Hoosbeek RM (eds) *Soil and Water Quality at Different Scales*, pp. 239–246. Proceedings of the workshop *Soil and Water Quality at Different Scales*, 7–9 August 1996, Wageningen, The Netherlands. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Arnold RW and Wilding LP (1991) The need to quantify spatial variability. In: Mausbach MJ and Wilding LP (eds) *Spatial Variability of Soils and Landform*, pp. 1–8. SSSA Special Publication No. 28, Madison, WI: Soil Science Society of America.

Burrough AB and McDonnell RA (2000) *Principles of Geographical Information Systems*. New York: Oxford University Press.

Collin M, McBratney A, Voltz M, and Walter C (eds) (2000) Developments in quantitative soil resource assessment (Pedometrics 1998). *Geoderma* 97(3–4): 147–424.

de Gruijter JJ (ed.) (1999) Special issue. (Pedometrics 1997). *Geoderma* 89(1–2): 1–175.

de Gruijter JJ, McBratney AB, and McSweeney K (eds) (1997) Fuzzy sets in soil science (Pedometrics 1994). *Geoderma* 77(2–4): 1–82.

de Gruijter JJ, Webster R, and Myers DE (eds) (1994) Development in spatial statistics for soil science (Pedometrics 1992). *Geoderma* 62(1–3): 1–326.

McBratney AB, Odeh IOA, Bishop TFA, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* 97: 293–327.

McBratney AB, Mendonça Santos ML, and Minasny B (2003) On digital soil mapping. *Geoderma* 117: 3–52.

Odeh IOA and McBratney AB (eds) (2001) Estimating uncertainty in soil models (Pedometrics 1999). *Geoderma* 103(1–2): 1–229.

Wagenet RJ, Hutson JL, and Bouma J (1994) Modeling water and chemical fluxes as driving forces of pedogenesis. In: Bryant RB and Arnold RW (eds) *Quantitative Modeling of Soil Forming Processes*, pp. 17–35. SSSA Special Publication No. 39. Madison, WI: Soil Science Society of America.

Webster R (1994) The developments of pedometrics. *Geoderma* 62: 1–15.

Webster R and Oliver AO (2001) *Geostatistics for Environmental Scientists*. Chichester, UK: John Wiley.

# PENMAN, HOWARD LATIMER

**J L Monteith**, Edinburgh, UK

The amount of water needed to ensure that a crop will provide the maximum yield possible in a given environment depends in a complex way on three main factors: prevailing weather, soil type, and the nature of physiological responses to the environment in general. Throughout the twentieth century, steady progress was made in measuring, quantifying, and interpreting these responses in terms of mechanisms. Towards the end of the century, developments in computing encouraged the building of 'models' that were intended to predict how the growth and yield of crops would respond to environmental factors in general and to weather in particular. Unfortunately, many of these models grew so rapidly that they shared the fate of the dinosaurs! More compact models have lasted longer because they follow classical lines, employing standard procedures based on minimal assumptions. A notable example in agricultural science is the equation for estimating the water use of arable crops, published by H.L. Penman of Rothamsted Experimental Station in 1948.

Howard Latimer Penman was born in 1909 in Dunston-on-Tyne, a small town in the north-east corner of England where his father worked as a master carpenter in a shipyard. At Blaydon County Secondary School, Howard passed the Oxford Higher School Certificate in 1927 and was accepted by the University of Durham for a Joint Honours course in mathematics and physics. Years later, a fellow-student wrote: "He was very much a northerner – a Tynesider – and on a suitable occasion could amuse by lapsing into the vernacular or singing a Tyneside song. His general approach to physics was stimulating; he accepted little on authority and thought things out for himself."

In 1930, Penman was awarded a First Class Honours degree in physics and for a short time taught at a boys' school until he decided to make research his career. At this stage, he was fortunate to encounter E.G. Richardson who offered him a studentship to work on the absorption of sound in porous materials. The work was successful and original enough to be published in 1933 with Richardson as a co-author; it was the first of Penman's 104 scientific papers written over the following 43 years. A Senior Exhibition offered by the County Education Committee then enabled him to embark on a research project managed by Richardson to investigate the dependence on temperature of the dispersion of sound in a gas. When completed, this study gained him a PhD from Durham University. Outside the laboratory, Penman was an active member of the Student Representative Council, played soccer and cricket, and sang in two choirs, an activity that he enjoyed for most of his life.

In 1934, Penman successfully applied for a post in the photochemistry section of the Rayon Department at the Shirley Institute in Manchester, administered by the British Cotton Research Association. The reports he presumably wrote on this work did not provide material for general publication. In 1935, however, a short paper on 'The effect of temperature on supersonic dispersion in gases,' based on his PhD, was published in the *Proceedings of the Physical Society.*

In 1937, Penman took a crucial step in his career when he successfully applied for a post in the Soil Physics Department at Rothamsted Experimental Station in Harpenden, a small town about 30 miles (50 km) north of London. Established by Sir John Lawes in 1843, Rothamsted is one of the world's oldest and most distinguished centers of research on agricultural crops.

Three years later, Penman published two papers on 'Gas and vapour movement in the soil' in the *Journal of Agricultural Science*, followed in the same year and journal by an analysis of 'Meteorological and soil factors affecting evaporation from fallow soil.' The latter paper was Penman's first excursion into what he often referred to as 'muddy-boot physics.' In 1941, he collaborated with his colleague R.K. Schofield to produce a major analysis of many measurements of evaporation from fallow soil that predecessors had installed in tanks surrounding Rothamsted's 'Bear Pit.' This was a deep, circular brick-lined enclosure built many years previously by Rothamsted's founders, Sir John Bennet Lawes and Henry Gilbert. (It survived well into the twentieth century but was eventually filled in to satisfy safety regulations imposed by vandals ignorant of the historical importance of the site!) A formidable archive of rainfall and drainage records existed but no previous analysis had been successful because the physics of natural evaporation was not properly understood. To get a better feel for the scale and interaction of the processes involved, Penman conducted a set of laboratory experiments that he described in the *Journal of Agricultural Science*, also in 1941.

In 1942, Penman produced a paper that marked an important extension of his research horizon from soil to the complex ways in which crops respond to climate as well as to soil conditions. However, it was not until 1948 that he published, in the *Proceedings of the*

*Royal Society*, the classic paper that brought him world renown – 'Natural evaporation from open water, bare soil, and grass.' This was based on a highly original combination of two fundamental equations, one describing the heat balance of a crop stand in terms of micrometeorology and the other describing its water balance in terms of hydrology. The variables involved (as daily averages) were:

1. the net amount of heat available from radiation, both solar and terrestrial
2. air temperature at screen height
3. vapor pressure at screen height
4. windspeed.

One variable was common to both equations – the effective surface temperature of foliage. The core of Penman's contribution to agricultural science was his realization that this unmeasured and, at the time, virtually unmeasurable quantity, could be eliminated by solving the two equations simultaneously to derive a new third equation that achieved world renown as the Penman formula.

To subject his equation to extensive tests in the field, Penman set up parallel irrigation experiments at Rothamsted, which has a clay soil, and at its Woburn outstation, where the soil is sandy. Each spring, he mustered a working party from his staff to install irrigation pipes – hard manual labor but a welcome escape from the laboratory provided the weather was fine. In due course, two major series of papers describing the analysis of Woburn measurements were published in the *Journal of Agricultural Science*. When a working group of the Agricultural Research Council described such a study as 'speculative,' Penman responded in characteristically trenchant language: "The term is unfortunate in its association with gambles and guesses.... Why the Research Council for the country's chief industry should seem to need special pleading for this kind of work is puzzling."

Following successful local tests of his equation, Penman applied it first to the whole of the UK and eventually, when checks against hydrological measurements appeared successful, to the mainland of Europe. His work on crop evaporation is an outstanding example of how a relatively simple but rigorous general model can be applied to many types of mixed vegetation growing in a wide range of climates.

(When Penman's seminal paper was published, he had worked at Rothamsted for 11 years during which he published 14 papers – a fairly leisurely rate by the norms of today. By 1974, however, when he retired, he had published over 100 papers and articles in a variety of journals ranging from the *Proceedings of the Royal Society of London* to the *Transactions of the Hertfordshire Natural History Society Field Club*.)

In due course, the Penman equation was adopted by several government agencies and used both on relatively small scales to estimate the irrigation need of agricultural and horticultural crops and on much larger scales to assess the hydrological efficiency of catchments. As the scientific importance and practical value of Penman's work became recognized internationally, he was increasingly invited to attend conferences abroad and to serve on committees such as the Coordinating Council for the International Hydrological Decade.

In 1954, Penman succeeded R.K. Schofield as Head of the Physics Department at Rothamsted and continued the tradition of allowing his staff the freedom to identify and explore lines of work that challenged them. This treatment followed his own experience when, during his first interview with his head of department he was told to "find something interesting to do and get on with it!" Most, but not all of his staff, made good use of this invitation to tackle challenging problems of their own choice. The benefits were personal while the work was in hand and universal once it was published.

Penman did not live to know that, in 1990, an extended form of his equation was adopted by the Food and Agriculture Organization of the United Nations (FAO) as the international standard for estimating the water need of crops. This would certainly have given him great satisfaction, but he was a modest man and might well have concealed this!

Although the Bear Pit site was not ideal, Penman was confident that records, carefully preserved for many years, could be used to test formulae relating to evaporation to weather.

There are at least two ways in which such testing can be done: by empiricism based on the correlation of estimates with rates of evaporation measured at one or more sites over long periods; and by deriving appropriate formulae to describe the mechanisms involved. The first route was taken by an American, Warren Thornthwaite, who explored the correlation of evaporation rates with temperature as reported from standard weather stations. This statistical approach to the problem did not appeal to Penman, nor did Thornthwaite's suggestion that the total amount of water lost by evaporation from plants and directly from the soil on which they grew, should be referred to as 'evapotranspiration' – a clumsy hybrid banned from Rothamsted publications! Penman coined the term 'potential transpiration' to describe the rate at which water was lost from a stand of vegetation with an abundant water supply and he provided an unambiguous definition: "The loss of water from short, green vegetation, completely covering the ground and never short of water."

Penman's mechanistic formula for evaporation was founded on two major physical principles: the conservation of mass and the conservation of energy. Conservation of mass implied that the rate of evaporation from a uniform area of vegetation should be proportional to the change of vapor pressure with height and to a function of windspeed. Conservation of energy implied that the net input of heat from radiation must be balanced by losses of energy in the form of convection and evaporation.

The two equations describing these processes contain one common term: the surface temperature of the source for evaporation, but this could not be accurately measured in Penman's time, let alone recorded over long periods. Penman overcame this obstacle in an ingenious but essentially simple way: he solved the mass and energy balance equations simultaneously by eliminating surface temperature, the term common to both.

The success of this exercise led to the publication in the *Proceedings of the Royal Society of London* of 'Natural evaporation from open water, bare soil, and grass.' This paper rapidly achieved recognition as a major step advance with great potential benefits for both agronomists and for hydrologists. By a curious coincidence, the American geographer C.W. Thornthwaite, also in 1948, published, in the *Geographical Review*, a paper titled 'A rational approach to the classification of climate.' This included the proposal that rates of evaporation from both vegetation and water could be assessed from empirical correlations with air temperature. Once, at a meeting of meteorologists in the USA, Thornthwaite was asked to say what he thought of the Penman formula. He thought for a moment before replying, with the faint vestige of a smile: "What I really dislike about Penman's formula is that it's so empirical!" The non-American members of the audience were convulsed in laughter, leaving their American friends somewhat baffled!

The soil at Rothamsted is classified as 'clay with flints' in marked contrast to the light sandy soil at an outstation on sandy soil at Woburn. Penman was anxious to establish the extent to which crop growth and yield could be increased by irrigation but recognized that soil type would determine optimal schedules. Every spring, the younger members of his department were deployed to install irrigation spray-lines at both sites and every autumn they were housed for the winter. Practical conclusions from this work were brought together in a UK Ministry of Agriculture, Fisheries and Food Technical Bulletin (No. 4) which had several external co-authors, including L.P. Smith from the Meteorological Office, with whom Penman enjoyed a very fruitful working relationship. Details of measurements and their scientific implications were discussed in a major series of papers, published in the *Journal of Agricultural Science* in 1962.

Building on the success of local trials, Penman began to extend his expertise to the hydrology of catchments, particularly those that were forested, both in the UK and elsewhere in Europe. At the outset, he was convinced that because evaporation from vegetation was mainly limited by the supply of radiant energy, there should be little difference between annual totals for farmland and for forests. However, some of the early field work by staff at the Institute of Hydrology at Wallingford, Berkshire (whose Director, James McCulloch, had worked as a postgraduate with Penman) demonstrated unequivocally that forests lose significantly more water than crops exposed to the same climate, partly because they intercept more water and partly because they are aerodynamically much rougher. It was therefore inappropriate to use the same wind function for both types of vegetation, as Penman initially assumed.

Penman was very fortunate to have Ian Long among his staff, an experimental officer with unique skills in designing, building, and maintaining equipment. One of his most impressive achievements was the construction of a set of neutron probes for measuring the water content of soil as a function of depth; shortly after, this device became available commercially. In 1973, a report on extensive tests of this device occupied about 80 pages of the Rothamsted Annual Report for 1972. Under Penman's direction, Long acquired thermistors soon after they came on the market and used them for many years to provide long-term records of dry- and wet-bulb temperature.

Penman's success in estimating evaporation on a field scale encouraged him to move up to the catchment scale and eventually to the national scale by integrating catchment estimates. In 1950, a major paper on 'Evaporation over the British Isles' was published in the *Quarterly Journal of the Royal Meteorological Society*. This was followed in 1954 by 'Evaporation over parts of Europe,' presented to the General Assembly of the International Association for Scientific Hydrology.

Also in 1954, Penman collaborated with staff at the Ministry of Agriculture, Fisheries and Food and at the Meteorological Office to produce a Technical Bulletin on 'The calculation of irrigation need,' applicable to both agriculture and horticulture. Further afield, he contributed to a government publication on evaporation from Lake Volta.

In 1963, Penman initiated irrigation trials on the Rothamsted farm after getting permission for the drilling of a borehole. Using a neutron probe, changes of water content were monitored at regular intervals as a function of depth and records for 13 seasons were eventually analysed. One major conclusion was that the soil water deficit at which growth became inhibited by drought was about 2.5 times larger than for the same crops grown on sandy soil at the Woburn farm.

The field study at Woburn, referred to above, ran for many years and major conclusions from this work were eventually published in a series of papers. The first three were published in the *Journal of Agricultural Science* in 1962, followed by a further group in 1970. Major sets of measurements made with the neutron probe were presented in Annual Reports of Rothamsted Experimental Station between 1969 and 1972 and a set of three major papers describing this work was published in the 1973 Report. Unfortunately, this valuable material did not find its way into the open literature.

Also in the 1960s, Penman began to use his formula to explore the hydrology of forested catchments in the UK and attended an International Symposium on Forest Hydrology at Pennsylvania State University to read a paper on the subject. This was followed by a review of the literature on evaporation from forests comparing field measurements with predictions from theory. On a much smaller scale, he became interested in the role that the stomata play in controlling the loss of water from leaves and collaborated with F.L. Milthorpe at the University of Nottingham to publish a paper on the diffusive conductivity of stomata on wheat leaves.

By this stage in his career Penman was much in demand as a speaker, partly because his work was of widespread interest to agronomists, ecologists, and meteorologists, but also because he lectured with admirable clarity and with occasional flashes of north-country humour. He also became involved in the management of scientific societies as a Councillor and Editorial Board member of the British Society of Soil Science, as Secretary of the London Branch of the Institute of Physics, as Secretary and eventually as President of the Royal Meteorological Society which elected him to Honorary Fellowship.

His help was sought by many national bodies: the Ministry of Agriculture for whom he helped advise the group that prepared a major report on 'The calculation of irrigation need'; the Institute of Water Engineers, the Meteorological Office, the Institute of Hydrology as it grew from a small unit to a major international centre. Further afield, he led the British delegation to meetings of the International Hydrological Decade and, in 1972, was host at the World Water Balance Symposium in Reading.

Away from his office, Penman's main interest was in music. During most of his time at Rothamsted, he was a member of the BBC Choral Society; but because this involved a fair amount of travelling and occasional nights away from home, he eventually joined the Bach Choir in nearby St Albans. However, he finally withdrew from this too because, as he explained "I almost sang a wrong note one day!" His repertoire included popular music, particularly folk-songs in the strong Geordie accent that he enjoyed reverting to long after leaving Newcastle. Over the years, he accumulated an impressive collection of classical records, played on an early hi-fi system. This was a rare extravagance, for the habit of thrift acquired early in life remained with him even after he became a relatively prosperous Head of Department. He never owned a car, for example, and, twice a day, walked back and forth across Harpenden Common on the route between his very modest home in Rowan Way to his office at Rothamsted, which looked eastward over Harpenden Common. The fine view was sometimes partly obscured by smoke from Penman's pipe that became particularly dense when a new manuscript was on the stocks!

Penman's international reputation ensured that he had many opportunities to travel abroad, to lecture, and to discuss with colleagues a wide range of hydrological problems. These visits included Australia, Canada, Kenya, Lebanon, the Netherlands, the USA, and West Africa. These and many other countries have benefited immeasurably from his work. Throughout the world, water is an essential commodity but there are many regions where it is scarce so that prudent management and use are vital. Penman's formula made it possible to explore and provide answers to a wide range of major environmental problems such as the optimal application of water to crops in rain-scarce areas; the rate of loss of surface water from reservoirs; and, most recently, the hydrological implications of global warming. The economic value of the Penman formula for evaporation is therefore immeasurable.

*See also:* **Evaporation of Water from Bare Soil; Evapotranspiration; Penman–Monteith Equation**

## Further Reading

Howard Latimer Penman (1909–1984) *Biographical Memoirs of Fellows of the Royal Society* 32: 379–404.

Penman HL (1940) Gas and vapour movement in the soil. *Journal of Agricultural Science.*

Penman HL (1940) Meteorological and soil factors affecting evaporation from fallow soil. *Journal of Agricultural Science.*

Penman HL (1948) Natural evaporation from open water, base soil and grass. *Proceedings of the Royal Society of London* 143: 120–145.

Penman HL (1950) Evaporation over the British Isles. *Quarterly Journal of the Royal Meteorological Society.*

# PENMAN–MONTEITH EQUATION

**R Allen**, University of Idaho, Kimberly, ID, USA

The Penman–Monteith (PM) evapotranspiration (ET) equation predicts the rate of total evaporation and transpiration from the earth's surface using commonly measured weather data (solar radiation, air temperature, vapor content, and wind speed). The PM equation follows a single-layer or 'big leaf' approach, where single-surface resistance and single aerodynamic resistance terms represent the transport properties of the cropped surface. The PM procedure provides scientists, engineers, and managers with the ability to evaluate effects of modifications to cultural practices and crop variety improvements on ET such as increased plant height, increased or reduced leaf size or density, reduced row spacing, and reduced stomatal conductance etc., by varying these parameters in the calculation algorithms. The PM method is useful in providing direction in determining changes in crop coefficients under new cropping conditions. Determination of crop coefficients is an expensive and time-consuming process, as it requires additional lysimeter or micrometeorological studies.

The PM equation has the form:

$$\text{E or ET} = \frac{\Delta(R_n - G) + \rho_a c_p (e_s - e_a)/r_a}{\left(\Delta + \gamma\left(1 + \frac{r_s}{r_a}\right)\right)\lambda\rho_w} \qquad [1]$$

where: $\Delta$ is the slope of the saturation vapor pressure vs. temperature curve; $R_n$ is the net radiation flux at the surface; $G$ is the sensible heat exchange from the surface to the soil (positive if the soil is warming); $\rho_a$ is air density; $c_p$ is specific heat of dry air; $e_s$ is the saturation vapor pressure of the air at some height above the surface; $e_a$ is the actual vapor pressure of the air; $r_a$ is aerodynamic resistance to turbulent heat and/or vapor transfer from the surface to some height $z$ above the surface; $\gamma$ is the psychrometric constant (defined later); $r_s$ is a bulk surface resistance that describes the resistance to flow of water vapor from inside the leaf, vegetation canopy, or soil to outside the surface; $\lambda$ is the latent heat of vaporization, defined as the energy required to convert a mass of liquid water into vapor (having typical units of joules per kilogram); $\rho_w$ is density of liquid water. All parameter units in eqn [1] must cancel so that the remaining units for evaporation ($E$) or ET (depending on application to evaporation from soil or ET from soil and vegetation) are presented as $L\,T^{-1}$, for example, in millimeters per hour or millimeters per day.

Surface resistance $r_s$ may represent the resistance to vapor flow through plant leaf stomates and/or resistance to vapor flow within soil to the soil surface. For a wet (saturated) surface, $r_s$ is by definition zero. When $r_s$ is zero, eqn [1] reverts to the original Penman equation.

## Derivation of the Penman–Monteith Equation

It is useful to review the derivation of the PM equation in order to understand how and why it functions. There are two fundamental approaches to predict $E$ (or ET). These are the surface energy balance equation and the aerodynamic equation. Each equation uses a different approach and each has difficulties for solution when using typical weather data comprised of solar radiation, air temperature, vapor pressure, and wind speed. Penman combined these two approaches into the so-called combination or Penman equation that could be solved using standard weather data only. Monteith later improved on the Penman equation by including a surface resistance term and a more rigorous term for aerodynamic transfer. It is useful to review the combination of the two approaches.

The surface energy balance equation relates to the various ways in which net radiation energy from the sun and atmosphere is balanced at the surface by inputs or outputs of energy from nonradiative parameters. The vertical energy balance at the surface is the sum of sensible heat flux to or from the air and to or from the soil, along with latent heat flux, net radiation, and other miscellaneous fluxes. The major components of the vertical energy balance are expressed as:

$$R_n = G + \lambda E + H \qquad [2]$$

where $\lambda E$ is the latent heat flux (positive during evaporation) and $H$ is the sensible heat exchange from the surface to air (positive if the air is warming). In eqn [2], $E$ (or ET) is expressed in terms of the energy required to convert the amount of liquid evaporated into vapor by multipling by $\lambda$. The miscellaneous terms involved in the surface energy balance, such as heat storage within foliage, photosynthesis, and respiration, are generally insignificant relative to magnitudes of $R_n$, $\lambda E$, $H$, and $G$, and are commonly neglected. Under some conditions, change in heat storage in a canopy may need attention, especially in forest canopies. If $R_n$ and sensible heat flux densities $H$ and $G$ can be measured or estimated reliably, then the latent heat flux density, $\lambda E$, can be computed from eqn [2].

$$\lambda E = R_n - G - H \qquad [3]$$

The challenge to the solution of eqn [3] using standard weather data is in solving for $H$. The traditional aerodynamic expression $H$ is:

$$H = \frac{\rho c_p (T_o - T_a)}{r_{ah}} \qquad [4]$$

where $T_o$ is surface temperature, $T_a$ is air temperature, and $r_{ah}$ is the aerodynamic resistance to heat transfer between some mean height within the surface to some height a few meters above the surface. Generally, $r_{ah}$ is assumed equal to $r_a$ of eqn [1].

Equation [4] is analogous to Ohm's law for electrical current flow, where $T_o$ and $T_a$ are equivalent to voltages or potentials, $r_{ah}$ is equivalent to an electrical resistor, and $H$ is analogous to current. Thus, eqn [4] behaves as a linear electrical circuit, so that if the difference between $T_o$ and $T_a$ is doubled, $H$ will double. If resistance $r_{ah}$ is halved (through a doubling of wind speed), then $H$ will double.

The challenge of eqn [4] and thus eqn [3] is in the parameter $T_o$. $T_o$ can be measured using infrared thermometers, if surface emissivity is known, or it can be measured using a fine thermocouple or other temperature sensor placed just beneath the surface skin. However, these measurements are not common due to the high operational and maintenance requirements of the instruments and difficulty in transfer of measured values to other locations and surface conditions, even those nearby.

The traditional equation for $r_{ah}$ (and $r_a$) is:

$$r_{ah} = \frac{\left(\ln\left(\frac{z_u - d}{z_{om}}\right) - \psi_m\right)\left(\ln\left(\frac{z_T - d}{z_{oh}}\right) - \psi_h\right)}{k^2 u_z} \qquad [5]$$

where: $z_u$ is the height above the ground surface for the wind speed measurement, $d$ is the zero plane displacement of the logarithmic wind profile, $z_{om}$ is a roughness length governing the transfer of momentum from the surface, $\psi_m$ is a correction factor for momentum transfer to account for buoyant instability or stability of the boundary layer, $z_T$ is the height of the air temperature measurement above the ground surface, $z_{oh}$ is an assumed roughness length governing the transfer of sensible heat from the surface, $\psi_h$ is a correction factor for sensible heat transfer to account for buoyant instability or stability of the boundary layer, $k$ is the von-Karman constant (0.41), and $u_z$ is the wind speed measured at the $z_u$ height. Equation [5] is sometimes criticized for its use of $z_{oh}$, commonly expressed as some constant fraction of $z_{om}$, whereas in reality the ratio of $z_{oh} : z_{om}$ can vary substantially, especially for dry, sparse vegetation. Other forms of eqn [5] can be written that rely on differences in $T_a$ and wind speed at two different heights above the surface. These forms eliminate the need for $z_{oh}$; however, measurement of $T_a$ and wind speed at two heights is not common and it is challenging to produce an unbiased pair of measurements. A variety of studies have investigated the effects of foliage density and geometry on apparent bulk roughness of vegetation.

The aerodynamic-based expression for $\lambda E$ is:

$$\lambda E = \frac{\rho c_p (e_o - e_a)}{\gamma (r_{av} + r_s)} \qquad [6]$$

where $e_o$ is vapor pressure 'inside' the surface, for example inside a leaf stomatal cavity or beneath the soil skin, $e_a$ is vapor pressure of the air at some height $z$ above the ground surface, $\gamma$ is the pyschrometric constant, $r_{av}$ is aerodynamic resistance to turbulent vapor transfer from the surface to some height $z$ above the surface, and $r_s$ represents any surface resistance to flow of vapor (defined after eqn [1]).

Aerodynamic resistance $r_{av}$ is computed exactly as for $r_{ah}$, i.e., using eqn [5], except that $z_T$ becomes $z_v$ (height of vapor pressure measurement above the ground), $z_{oh}$ becomes $z_{ov}$, which is an assumed roughness length governing the transfer of vapor from the surface, and $\psi_h$ becomes $\psi_v$, a correction factor to account for the effect of instability or stability of the boundary layer on vapor flow. Generally, $z_T = z_v$ (air temperature and humidity sensors are usually collocated at the same height), and $z_{ov}$ is assumed equal to $z_{oh}$, and $\psi_v$ is assumed equal to $\psi_h$. Brutsaert and Allen have shown that the stability parameters $\psi_v$, $\psi_h$, and $\psi_m$ can generally be set to zero when predicting ET from well-watered (moist) surfaces because of nearly neutral stability (buoyancy) for those conditions.

Again, the aerodynamic vapor transfer equation is analogous to Ohm's law, where $e_o$ and $e_a$ are equivalent to voltages or potentials, $r_{av}$ and $r_s$ are equivalent to electrical resistors in series, and $\lambda E$ (vapor flux) is analogous to electrical current.

Vapor pressure of air can be determined from paired measurements of relative humidity and air temperature, which are commonly available, or from measurement of dew-point temperature, also commonly available. Vapor pressure inside the soil

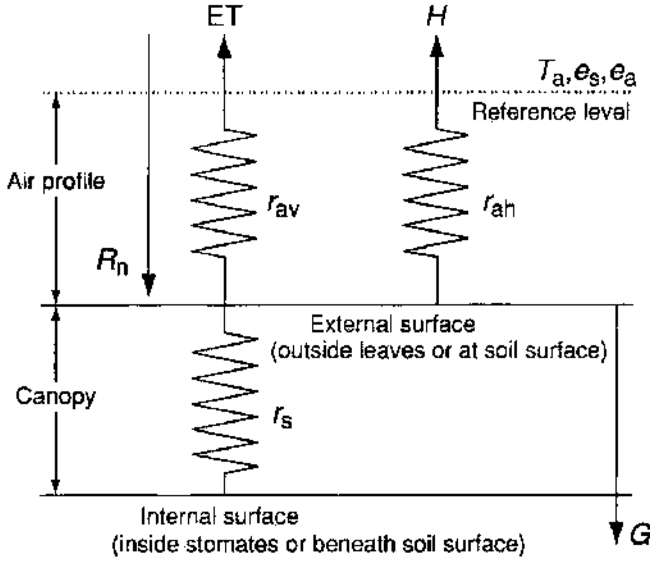


**Figure 1** Schematic showing linkage between resistance terms in the Penman–Monteith equation relative to the surface and elevation of temperature and humidity measurements. ET, evapotranspiration; $e_s$, saturation vapor pressure at mean air temperature; $e_a$, actual vapor pressure of the air; *G*, heat exchange from surface to soil; *H*, heat exchange from surface to air; $r_a$, aerodynamic resistance; $r_s$, bulk surface resistance; $R_n$, net radiation flux at the surface; $T_a$, air temperature; $r_{av}$, aerodynamic resistance to vapor transfer; $r_{ah}$, aerodynamic resistance to heat transfer.

or plant surface is rarely measured, however, so that, as with eqn [3], eqn [6] is challenging to apply in routine practice. Hence the need for the Penman or PM combination equation.

**Figure 1** is a schematic showing the relative locations of resistances $r_a$ (or $r_{ah}$ and $r_{av}$) and $r_s$, where $r_s$ represents the internal resistance of the soil (if no vegetation) or of the soil, leaves, and canopy in parallel.

To create the PM equation, one needs to define what is known as the Bowen ratio, which is simply the ratio between *H* and $\lambda E$. Expanded into the equations for *H* and $\lambda E$ and written for transfer from inside the surface to the air above:

$$\beta = \frac{H}{\lambda E} = \frac{\rho c_p (T_o - T_a)\,\gamma\,(r_{av} + r_s)}{\rho c_p (e_o - e_a)(r_{ah})} \qquad [7]$$

Cancelling similar terms:

$$\beta = \frac{H}{\lambda E} = \gamma \frac{(T_o - T_a)(r_{av} + r_s)}{(e_o - e_a)(r_{ah})} \qquad [8]$$

The energy balance equation can be rearranged and written as:

$$\lambda E = \frac{R_n - G}{1 + \beta} = \frac{R_n - G}{1 + \gamma \frac{(T_o - T_a)(r_{av} + r_s)}{(e_o - e_a)(r_{ah})}} \qquad [9]$$

The next step in the derivation of the PM equation is to define the term $\Delta$, which is defined as the slope of the saturation vapor pressure vs. temperature relation. Saturation vapor pressure $e^o(T)$ is a singular function of *T* and describes the ability of air to hold more vapor as temperature increases. The slope $\Delta$ of the $e^o(T)$ vs. *T* curve can be expressed as a linear approximation for a segment of the saturation curve between two temperatures $T_o$ and $T_a$:

$$\Delta = \frac{e_o - e_s}{T_o - T_a} \qquad [10]$$

where $e_o$, $e_s$, $T_o$ and $T_a$ were defined previously. Vapor pressure inside the surface, $e_o$, is assumed to be saturated and is therefore equivalent to $e^o\ (T_o)$. $e_s$ is the saturation vapor pressure associated with $T_a$, so that $e_s = e^o\ (T_a)$. In practice, $\Delta$ is often calculated as the slope of $e^o\ (T)$ vs. *T* at $T_a$ only, since $T_o$ is usually not known.

Using the relationship from eqn [10] to define $(e_o - e_s)/\Delta = T_o - T_a$ and parsing $(e_o - e_s)$ into an equivalent expression of $(e_o - e_s) = (e_o - e_a) - (e_s - e_a)$ so that $T_o - T_a = [(e_o - e_a) - (e_s - e_a)]/\Delta$, this expression is substituted into eqn [9] so that:

$$\begin{aligned} \lambda E &= \frac{R_n - G}{1 + \beta} \\ &= \frac{R_n - G}{1 + \frac{\gamma}{\Delta} \frac{[(e_o - e_a) - (e_s - e_a)](r_{av} + r_s)}{(e_o - e_a)(r_{ah})}} \\ &= \frac{R_n - G}{1 + \frac{\gamma}{\Delta}\left[1 - \frac{(e_s - e_a)}{(e_o - e_a)}\right] \frac{(r_{av} + r_s)}{r_{ah}}} \end{aligned} \qquad [11]$$

We define one more temporary term, $\lambda E_a$, after Penman:

$$\lambda E_a = \frac{\rho c_p (e_s - e_a)}{\gamma r_{ah}} \qquad [12]$$

where $\lambda E_a$ is similar to eqn [6], the aerodynamic expression for $\lambda E$, except that $r_s$ is omitted and $(e_o - e_a)$ is replaced with $(e_s - e_a)$. $E_a$ is a fictitious term that allows the elimination of surface parameters $T_o$ and $e_s$ in the final PM equation. Term $(e_s - e_a)$ is known as the vapor pressure deficit (VPD) of the air, because it represents the difference between the saturation vapor pressure at air temperature $(e_s = e^o\ (T_a))$ and actual vapor pressure of the air $(e_a)$. Both $e_s$ and $e_a$ can be determined from commonly available humidity and air temperature data measured at some height above the surface, generally 1.5–2 m. Because:

$$\frac{e_s - e_a}{e_o - e_a} = \frac{\lambda E_a}{\lambda E} \frac{r_{ah}}{r_{av} + r_s} \qquad [13]$$

Then eqn [11] becomes:

$$\lambda E = \frac{R_n - G}{1 + \frac{\gamma}{\Delta}\left[1 - \frac{\lambda E_a}{\lambda E}\frac{r_{ah}}{r_{av} + r_s}\right]\frac{r_{av} + r_s}{r_{ah}}} \qquad [14]$$

Rearranging eqn [14] and setting $r_{av} = r_{ah}$ results in the PM equation:

$$\lambda E = \frac{\Delta(R_n - G) + \gamma \lambda E_a}{\Delta + \gamma\left(1 + \frac{r_s}{r_{ah}}\right)} = \frac{\Delta(R_n - G) + \rho c_p (e_s - e_a)/r_{ah}}{\Delta + \gamma\left(1 + \frac{r_s}{r_{ah}}\right)} \qquad [15]$$

where, in eqn [15], $E$ is expressed in terms of energy, rather than in terms of depth of liquid water as in eqn [1]. The PM equation has the extremely valuable advantage of requiring only commonly available weather data, namely solar radiation (used to calculate $R_n$), air temperature (used to calculate $\rho$, $e_s$ and $\Delta$), air humidity (used to calculate $e_a$ and assist with $R_n$), and wind speed (for $r_{ah}$). Parameter $G$ is generally estimated as a function of $R_n$ or by direct measurement, and $r_s$ is generally estimated as a function of amount of vegetation or amount of surface wetness if soil. The PM (and Penman equation where $r_s$ is assumed zero and $1/r_{ah}$ is replaced by a more simple empirical 'wind function') are widely used around the world as a standard means to predict $E$ and ET.

## Other Parameters in the Penman–Monteith Equation

The psychrometric constant $\gamma$ is computed as:

$$\gamma = \frac{c_p P}{\lambda \epsilon} \qquad [16]$$

where $c_p = 1.013\,\mathrm{kJ\,kg^{-1}\,K^{-1}}$ for moist air, $P$ is atmospheric air pressure, $\lambda$ is the latent heat of vaporization in kilojoules per kilogram, and $\epsilon = 0.622$ is the ratio of molecular weights of water vapor to air. For $P$ in kilopascals, $\gamma$ will have units of kilopascals per degrees celsius. Mean atmospheric air pressure, $P$, can be computed routinely as a function of surface elevation using the ideal gas law as:

$$P = P_{os}\left(\frac{T_{os} - \alpha(\text{elev} - \text{elev}_{os})}{T_{os}}\right)^{g/(\alpha_a R)} \qquad [17]$$

where $P_{os}$ and $T_{os}$ are standard atmospheric pressure in kilopascals and absolute temperature in degrees Kelvin at elevation $\text{elev}_{os}$, and elev is elevation of the ground surface above mean sea level; $\alpha_a$ is the assumed constant adiabatic lapse rate, normally taken as $0.0065\,\mathrm{K\,m^{-1}}$ for saturated air or sometimes as $0.01\,\mathrm{K\,m^{-1}}$ for nonsaturated air; $g$ is gravitational acceleration, $9.807\,\mathrm{m\,s^{-2}}$, and $R$ is the specific gas constant for dry air, $287.0\,\mathrm{J\,kg^{-1}\,K^{-1}}$. Values for $P_{os}$, $T_{os}$, and $\text{elev}_{os}$ are commonly taken for the standard atmosphere at sea level, which are 101.3 kPa, 288 K, and 0 m, respectively. Equation [17] is relatively insensitive to the value of $\alpha_a$ for elevations up to 3000 m.

Air density, $\rho$, can be computed as:

$$\rho = \frac{1000\,P}{T_v R} \qquad [18]$$

where $P$ is in kilopascals, $R$ is $287.0\,\mathrm{J\,kg^{-1}\,K^{-1}}$, and $T_v$ is virtual temperature, in degrees Kelvin. $T_v$ can be computed as:

$$T_v = \frac{T}{1 - 0.378\frac{e_a}{P}} \qquad [19]$$

where $T$ is air temperature in Kelvin and $e_a$ is mean vapor pressure of the air (kilopascals). Latent heat of vaporization, $\lambda$, in megajoules per kilogram can be computed as:

$$\lambda = 2.501 - 2.361 \times 10^{-3}\,T \qquad [20]$$

where $T$ is mean air temperature in degrees Celsius. If available, mean surface temperature or wet bulb temperature can be used to compute the value of $\lambda$, which better represents conditions at the evaporating surface.

The PM equation can be applied for short periods of 1 h or less, and is intended to represent nearly instantaneous fluxes of ET. However, the equation has traditionally been applied for daily and longer (up to 1 month) calculation time-steps owing to paucity of weather data. For daily and longer time-steps, $e_s$ in the PM equation should be computed as:

$$e_s = \frac{e^o(T_{max}) + e^o(T_{min})}{2} \qquad [21]$$

where $T_{max}$ and $T_{min}$ are maximum and minimum daily air temperature for the period and $e^o$ is the saturation vapor pressure function. Actual vapor pressure of the air, $e_a$, can be computed from relative humidity data, from wet-bulb/dry-bulb pyschrometer data, or from dew-point temperature measurements:

$$e_a = e^o(T_d) \qquad [22]$$

where $T_d$ is mean daily or early morning dew-point temperature.

### Aerodynamic and Surface Parameters

When used to predict ET from dense vegetation, aerodynamic resistance in the PM equation can be calculated using parameters for $d$, $z_{om}$, and $z_{oh}$ determined as a function of vegetation height:

$$d = 0.67H \qquad [23]$$

$$z_{om} = 0.12H \qquad [24]$$

$$z_{oh} \approx 0.1z_{om} \qquad [25]$$

where $z_{om}$, $z_{oh}$, $d$, and mean plant height, $H$, are in meters. For bare soils, common values for $z_{om}$ range from 0.00 001 m for mud flats to 0.005 m for smooth soil to 0.01 m or more for very roughly ploughed soil.

## Parameterization of Surface Resistance

The bulk surface resistance $r_s$ for a crop is an integration of effects of transpiration from various locations within the vegetation and contributions from soil evaporation, along with transfers of radiation and sensible heat within the canopy. Simple approaches determine soil evaporation indirectly by estimating the distribution of net radiation below and between a plant canopy having partial ground cover. For crops that completely shade the ground, total surface resistance can be computed, where surface resistances from soil and vegetation are added in parallel according to leaf area and radiation extinction coefficients. For an immature row crop having exposed bare soil between plant rows and incomplete wetting of the soil surface during irrigation, $r_s$ can be expressed through a combination of various surface resistances acting in parallel. The combination is made according to the fractions of surface C represented by each parallel resistance component (**Figure 2**):

$$\frac{1}{r_s} = \frac{C(1 - K_r^{\mathrm{LAI}/C})}{r_{sc}} + K_r^{\mathrm{LAI}/C}\left(\frac{C_{sW}}{r_{sssW}} + \frac{C_{sD}}{r_{sssD}}\right) + \frac{C_{eW}}{r_{sseW}} + \frac{C_{eD}}{r_{sseD}} \qquad [26]$$

where C is the effective fraction of ground shaded by vegetation and is computed as a function of sun angle, row height and width, leaf density, and row orientation; and $C_{sW}$, $C_{sD}$, $C_{eW}$, and $C_{eD}$ are the fractions of the soil surface that are shaded and wet, shaded and dry, sunlit and wet, and sunlit and dry ($C_{sW} + C_{sD} + C_{eW} + C_{eD} = 1$ and $C_{sW} + C_{sD} = C$). Resistance $r_{sc}$ is the surface resistance of the canopy, $r_{sssW}$ is the surface resistance of wet, shaded soil; $r_{sssD}$ is the surface resistance of dry, shaded soil; $r_{sseW}$ is the surface resistance of wet, sunlit soil; and $r_{sseD}$ is the surface resistance of dry, sunlit soil. Parameter $K_r$ is a radiation extinction coefficient for the vegetation canopy, having a value of approximately 0.7; and LAI is leaf area index.

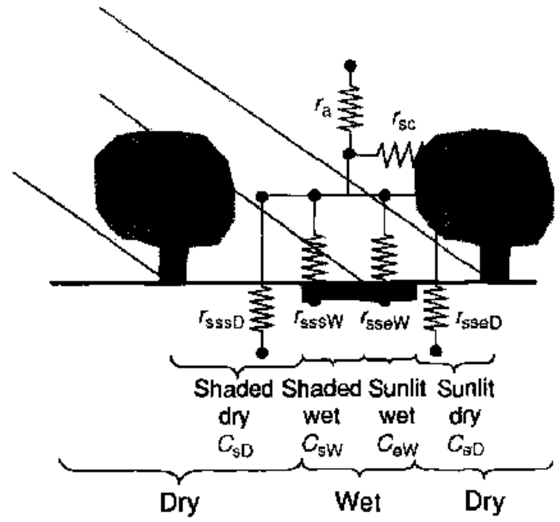


**Figure 2** Schematic showing breakdown and definition of surface resistances for soil when vegetation is at partial cover. $r_a$, aerodynamic resistance; $r_{sc}$, surface resistance for the vegetation cover; $r_{sssW}$, surface resistance of wet, shaded soil; $r_{sssD}$, surface resistance of dry, shaded soil; $r_{sseW}$, surface resistance of wet, sunlit soil; $r_{sseD}$, surface resistance of dry, sunlit soil; $C_{sD}$, fraction of the soil surface that is shaded and dry; $C_{sW}$, fraction of the soil surface that is shaded and wet; $C_{eW}$, fraction of the soil surface that is sunlit and wet; $C_{eD}$, fraction of the soil surface that is sunlit and dry.

C has also been computed as LAI/4. Separate PM equations have been applied for evaporation and transpiration and then recoupled, similar to the procedure for multilayer ET models. For sun angles other than directly overhead, C can be estimated for a row crop as:

$$C = \frac{W + H\,|\sin(\eta + \beta)|\tan\theta_z}{S} \qquad [27]$$

where $H$ is mean plant height, $\theta_z$ is the solar zenith angle (zero = nadir (overhead)), $\eta$ is the sun azimuth angle ($\eta = 0$ when sun is due south, $\eta$ is negative when sun is east of south, and $\eta$ is positive when sun is west of south), and $\beta$ is the row orientation relative to north ($\beta = 0$ for north–south orientation, $\beta = 90°$ for east–west row orientation), $W$ is row width, and $S$ is row spacing. Equation [27] presumes a nearly rectangular row shape of height $H$ and width $W$ and has an upper limit of $C = 1$. In some situations, $W$ may not be measured or available, but an estimate of LAI may be available. In this case, $C$ may be approximated during incomplete row closure as:

$$C = \frac{\mathrm{LAI}}{\mathrm{LAI}_C} + \frac{H}{S}\,|\sin(\eta + \beta)|\tan\theta_z \qquad [28]$$

where $\mathrm{LAI}_c$ is the LAI at complete row closure and shading of the ground. The value for $C$ and, subsequently, for $r_s$ changes with time of day and interacts with $R_n$, $G$, wind speed, and temperature. For 24-h computation time-steps, an average daily value for $C$ is determined for computing soil moisture balances for shaded and unshaded soil. An average daily value for $C$ for east–west rows can be determined by weighting eqn [26] for sunlight intensity and integrating between the limits of sunrise and sunset. For north–south row orientation, all soil is sunlit and shaded at various times of the day.

Surface resistance, $r_{sc}$, for densely growing vegetation is often computed in the 'single-layer' PM as a function of leaf area by assuming all leaves function as resistors in parallel:

$$r_{sc} = \frac{r_l}{\mathrm{LAI}_{\mathrm{eff}}} \qquad [29]$$

where $r_l$ is the bulk stomatal (or surface) resistance of the vegetation per unit LAI (seconds per meter) and $\mathrm{LAI}_{\mathrm{eff}}$ is the effective LAI contributing to ET. Parameter $r_l$ is the inverse of $g_l$, the stomatal conductance per unit leaf area. The value of $r_l$ has been taken as $100\,\mathrm{s\,m^{-1}}$ for many well-watered agricultural crops when calculations are made on a 24-h basis. When ET calculations are made on an hourly or shorter basis, $r_l$ for agricultural crops has a value of approximately $70–80\,\mathrm{s\,m^{-1}}$, assuming that stability corrections are made or that buoyancy conditions are nearly neutral. The value for $r_l$ increases with environmental stresses, including soil moisture deficit.

Surface resistance, $r_{sc}$, plays an important role in controlling the flux of transpiration from plants into the atmosphere and must be accurately predicted for routine application of the PM and multilayer ET models. $r_s$, represents the ET-averaged daily resistance for 24-h calculation time steps or the mean canopy resistance for hourly time steps for a fully developed canopy. Algorithms are used to modify $r_{sc}$ as leaf area (LAI) and environmental parameters change. General, simple environmental stomatal regulation algorithms have been proposed, where bulk surface resistance is expressed in terms of multiplicative products of linear functions of solar radiation, air temperature, vapor pressure deficit, and soil water deficit.

Soil resistance is generally assumed to increase with drying of soil. Exponential functions describe the process relatively well. A moisture-based soil resistance function describes experimental data from a Portneuf silt loam soil near Kimberly, Idaho:

$$r_{ss} = 2000\mathrm{e}^{-5x_s} \qquad [30]$$

where $x_s$ is the remaining proportion of water in the evaporating layer of the soil surface and e is the natural number. $x_s$ is computed as $x_s = W_{ei}/W_{emax}$, where $W_{ei}$ is the total evaporable moisture near the soil surface on day $i$ and $W_{emax}$ represents the upper limit of evaporable water. A daily or more frequent water balance of the $C_{sW}$, $C_{sD}$, $C_{eW}$, and $C_{eD}$ fractions of the soil surface layer need to be conducted to predict $x_s$, and thus $r_{ss}$, for each surface fraction. The evaporation from each fraction is predicted in some proportion to ratio of the inverse of $r_{ss}$ for that fraction to the total bulk $r_s$ for the soil–canopy complex. The soil surface layer is typically presumed to be 0.1–0.15 m thick. Equation [30] predicts values for $r_{ss}$ ranging from approximately $50\,\mathrm{s\,m^{-1}}$ directly after wetting to approximately $2000\,\mathrm{s\,m^{-1}}$ for a relatively dry surface soil layer.

## Multilayer ET Models

The PM equation is basically a single-layer model where the resistances for vegetation and soil are assumed to reside in parallel, i.e., there are no intermediate aerodynamic or convective resistances separating them or within them. More complicated, multilayer ET models have been employed that contain sophisticated parameter algorithms, intermediate resistances, and relatively complex calculation mechanisms than the PM. Multilayer models decompose the computation of $r_s$, $r_{sc}$, and $r_a$ for discrete layers within the vegetation canopy. These models are generally capable of making more accurate predictions of ET than the PM, especially for short time-steps (hourly). Because of their increased complexity, multilayer models can provide more capability and flexibility in predicting ET under variable cropping and weather conditions without the need for calibration, depending on the physical basis of internal mechanisms. On the other hand, these models sometimes require substantial calibration of empirical coefficients internal to the model.

## Net Radiation

Net radiation, $R_n$, can be measured using a net radiometer that measures downward and upward fluxes of shortwave and longwave radiation. The surface beneath the net radiometer must be very similar in wetness and vegetation type and amount to that being predicted in the PM equation. In many situations, $R_n$ data are not available for a specific surface or condition, and therefore must be estimated from solar radiation data, air temperature, and possibly vapor pressure. $R_n$ for row crops is often calculated by separately considering the reflectances (albedo) for soil and vegetation. When calculated, the general formula for $R_n$ is:

$$R_n = (1 - \alpha)R_s + R_{nl} \qquad [31]$$

where $\alpha$ is the albedo of the surface, representing the crop-soil mixture, $R_s$ is incoming shortwave radiation (measured by a pyranometer), and $R_{nl}$ is net (incoming – outgoing) longwave radiation. $R_{nl}$ is commonly predicted from air temperature and humidity data from a local weather station.

A general formula for $\alpha$ of a crop-soil mixture is:

$$\alpha = C\alpha_c + (1 - C)\alpha_s \qquad [32]$$

where $\alpha_c$ is plant albedo, $\alpha_s$ is soil albedo, and $C$ is the relative fraction of vegetation cover (defined previously) that considers effects of sun angle. Plant albedo can be computed by choosing typical values at the key growth stages. For example, values of $\alpha_c$ have been found to be 0.25, 0.23, 0.23, and 0.30 for beans at Kimberly, Idaho, at emergence, effective cover, ripening, and harvest. Values are 0.25, 0.15, 0.15, and 0.35 for wheat for the same stages. For overcast conditions the value for $\alpha_c$ increases, often to approximately 0.35.

Soil albedo in eqn [32] can be computed as a composite of wet soil and dry soil albedo:

$$\alpha_s = \frac{\alpha_{sM} C_{eW} + \alpha_{sD} C_{eD}}{C_{eW} + C_{eD}} \qquad [33]$$

where $\alpha_{sM}$ is albedo of moist soil and $\alpha_{sD}$ is albedo of dry soil. The albedo of a moist, drying soil can be assumed to vary linearly with water content between the value for a fully wet soil and a fully dry soil over a given drying period:

$$\alpha_{sM} = \alpha_{sW} + (\alpha_{sD} - \alpha_{sW})(1 - x_{skin}) \qquad [34]$$

where $\alpha_{sW}$ is albedo of wet soil and $x_{skin}$ is the fraction of remaining water in the soil skin (top approx. 1 cm). Example values for $\alpha_{sW}$ and $\alpha_{sD}$ for the Portneuf silt loam near Kimberly are $\alpha_{sW} = 0.10$–$0.20$ and $\alpha_{sD} = 0.35$.

## Soil Heat Flux

The magnitude of soil heat storage or release can be significant over a few hours, but is usually small from day to day because heat stored early in the day as soil warms is lost late in the day and at night as the soil cools. For application of PM to short periods, e.g., using hourly data, measurements or estimates for $G$ are required.

One method commonly used in research is the use of heat flux plates installed at shallow depths below the soil surface (e.g., at 0.01 m). Measurements by soil heat flux plates, $G_{z_s}$ at some depth $z_s$ are adjusted to represent $G$ at the soil surface by incorporating the change in heat storage between the soil surface and heat flux plate:

$$G = G_{z_s} + C_s \int_0^{z_s} \frac{\partial T_s}{\partial t} dz \qquad [35]$$

where $\partial T_s/\partial t$ is the change in temperature with time within the $z_s$ thick layer above the heat flux plate. Temperature is usually measured with thermocouples or buriable thermistors.

Soil heat flux density for dense, short vegetation such as for grass, wheat, or alfalfa can be approximated during daylight periods as:

$$G = 0.4e^{-0.5\,\mathrm{LAI}} R_n \qquad [36]$$

where $e$ is the natural number and $G$ has the same units as $R_n$. Equation [36] predicts $G = 0.1\,R_n$ for an $\mathrm{LAI} = 2.8$, which is typical for clipped grass. A low fraction of $G/R_n$ (0.036) has been used during daylight for a deciduous forest. The American Society of Civil Engineers (ASCE) recommends $G/R_n = 0.04$ for full-cover alfalfa during daytime. For $\mathrm{LAI} = 0$ (i.e., bare soil), eqn [36] predicts $G/R_n = 0.4$.

For nighttime under grass forage, the recommendation is $G = 0.5\,R_n$ for grass and 0.2 for full-cover alfalfa. In these reported ratios, the effects of seasonal variations are not addressed. The seasonal lag of soil temperature behind solar radiation can be expected to produce higher ratios.

## Application of the Penman–Monteith Equation

Besides direct application of the PM equation to a complex surface of soil and vegetation or to soil only, the equation has been widely used to represent ET from a reference vegetation crop for use in irrigation management. The reference crop ET ($\mathrm{ET_o}$) represents ET from either a cool-season clipped grass or a tall, full-cover alfalfa crop, depending on the parameters specified for aerodynamic and surface resistances.

When the PM equation is used as an ET reference, actual ET for a soil or soil–vegetation surface is computed by multiplying $\mathrm{ET_o}$ by a crop coefficient, $K_c$, where $K_c$ represents the ratio of actual ET to $\mathrm{ET_o}$. The value for $K_c$ of soil is varied according to the moisture level of the soil. Figure 3 shows a comparison of hourly ET predicted by the ASCE version of the PM equation and ET measured by a precision-weighing lysimeter (essentially a huge flowerpot suspended

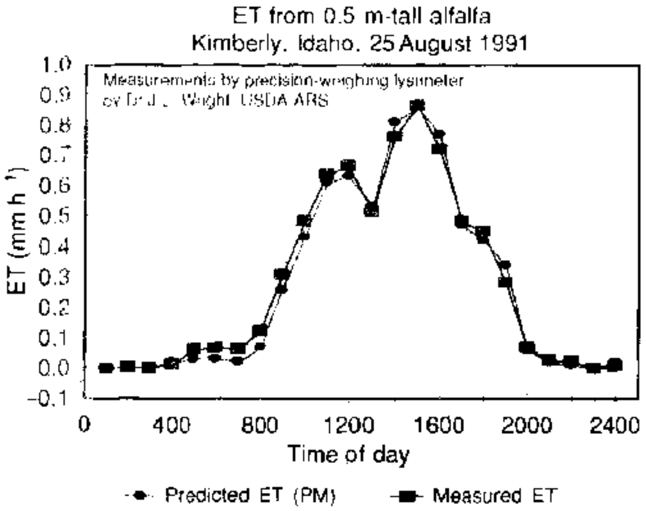


**Figure 3** Comparison between evapotranspiration (ET) during hourly periods predicted by the Penman–Monteith (PM) equation (using parameters from the American Society of Civil Engineers) and ET measured by a precision-weighing lysimeter.

below ground via mechanical balances). The PM estimates, based on measured solar radiation, air temperature, humidity, and wind speed, agree very well with field measurements. Results are not always this good, however, and depend on the representativeness of values for roughness and for surface resistance.

*See also:* **Energy Balance; Evapotranspiration; Penman, Howard Latimer**

## Further Reading

Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*. Irrigation and Drainage Paper 56. Rome, Italy: United Nations FAO.

ASCE (2002) *The ASCE Standardized Reference Evapotranspiration Equation*. Report of the Task Committee on Standardization of Reference Evapotranspiration, Environmental and Water Resources Institute of the American Society of Civil Engineers. Reston, VT: ASCE Publications.

Brisson N (1996) On the importance of soil evaporation and rainfall interception in irrigation scheduling. In: *Evapotranspiration and Irrigation Scheduling*, pp. 281–288. Proceedings of the International Conference of ASAE, San Antonio, TX.

Brutsaert W (1982) *Evaporation into the Atmosphere*. Dordrecht, the Netherlands: D. Reidel.

Choudhury BJ (1989) Estimating evaporation and carbon assimilation using infrared temperature data: vistas in modeling. In: Ghassem A (ed.) *Theory and Applications of Optical Remote Sensing*, pp. 628–690. New York: John Wiley.

Choudhury BJ, Idso SB, and Reginato RJ (1987) Analysis of an empirical model for soil heat flux under a growing wheat crop for estimating evaporation by an infrared temperature-based energy balance equation. *Agricultural and Forest Meteorology* 39: 283–297.

Clothier BE, Clawson KL, Pinter PJ, Moran MS, Reginato RJ, and Jackson RD (1986) Estimates of soil heat flux from net radiation during the growth of alfalfa. *Agricultural and Forest Meteorology* 37: 319–329.

Grant DR (1975) Comparison of evaporation from barley with Penman estimates. *Agricultural Meteorology* 15: 49–60.

Jensen ME, Burman RD, and Allen RG (1990) *Evapotranspiration and Irrigation Water Requirements*. ASCE Manuals and Reports on Engineering Practice No. 70. Reston, VA: ASCE Publications.

Lafleur PM and Rouse WR (1990) Application of an energy combination model for evaporation from sparse canopies. *Agricultural and Forest Meteorology* 49: 135–153.

Leuning R, Condon AG, Dunin FX, Zegelin S, and Denmead OT (1994) Rainfall interception and evaporation from soil below a wheat canopy. *Agricultural and Forest Meteorology* 67(3–4): 221–238.

Lhomme JP (1988) A generalized combination equation derived from a multi-layer micrometeorological model. *Boundary-Layer Meteorology* 45: 103–115.

Liu S, Riekerk H, and Gholz HL (1996) ETM evapotranspiration model: an introduction. In: *Evapotranspiration and Irrigation Scheduling*, pp. 329–334. Proceedings of the International Conference of ASAE, San Antonio, TX.

Monteith JL (1965) Evaporation and the environment. In: *The State and Movement of Water in Living Organisms*, pp. 205–234. XIXth Symposium of the Society for Experimental Biology, Swansea. Cambridge, UK: Cambridge University Press.

Penman HL (1948) Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London*. Series A: Mathematical and Physical Sciences 193: 120–146.

Perrier A (1982) Land surface processes: vegetation. In: Eagleson P (ed.) *Land Surface Processes in Atmospheric General Circulation Models*, pp. 395–448. Cambridge, UK: Cambridge University Press.

Price DT and Black TA (1989) Estimation of forest transpiration and $CO_2$ uptake using the Penman–Monteith equation and a physiological photosynthesis model. In: Black TA, Spittlehouse DA, Novak MD, and Price DT (eds) *Estimation of Areal Evapotranspiration*, pp. 213–227. IAHS Publication No. 177.

Qualls R and Brutsaert W (1996) Effect of vegetation density on the parameterization of scalar roughness to estimate spatially distributed sensible heat fluxes. *Water Resources Research* 32(3): 645–652.

Seginer I (1973) Aerodynamic roughness of vegetated surfaces. *Boundary Layer Meteorology* 5: 383–393.

Shuttleworth WJ and Wallace JS (1985) Evaporation from sparse crops – an energy combination theory.

*Quarterly Journal of the Royal Meteorological Society* 111: 839–853.

Stewart JB (1989) On the use of the Penman–Monteith equation for determining areal evapotranspiration. In: Black TA, Spittlehouse DL, Novak MD, and Price DT (eds) *Estimation of Areal Evapotranspiration*, pp. 3–12. IAHS Publication No. 177.

Szeicz G and Long IF (1969) Surface resistance of crop canopies. *Water Resources Research* 5: 622–633.

Wallace JS, Roberts JM, and Sivakumar MVK (1990) The estimation of transpiration from sparse dryland millet using stomatal conductance and vegetation area indices. *Agricultural and Forest Meteorology* 51: 35–49.

## Percolation *See* Hydrodynamics in Soils

## Permafrost *See* Polar Soils

## Permeability *See* Hydrodynamics in Soils

## Persistent Organic Pollutants (POPs) *See* Pollutants: Persistent Organic (POPs)

# PESTICIDES

**R H Bromilow**, Rothamsted Research, Harpenden, UK

## Introduction

Pesticides are used worldwide and increase crop yield on average by 30% as well as improving crop quality. Modern organic (i.e., carbon-containing) pesticides have been used widely since the late 1940s, and so there is over 50 years of experience of their behavior in the environment and their impact upon it. The spectrum of compounds applied over this time has changed substantially, with, for example, the older organochlorine insecticides being phased out and with completely new types of chemical structures being regularly introduced. Though increasingly strict registration criteria, coupled with market forces, are leading to the discontinuation of the use of many older compounds, it is likely that several hundred active ingredients will remain available. In the European Union, for example, perhaps 300 of approximately 850 formerly approved compounds will still be marketed after the current review process.

The soil is the ultimate sink for much of the pesticide that is applied to crops or indeed to uncropped land such as railway tracks. Compounds such as nematicides (for control of plant-parasitic nematode worms) are usually incorporated into the topsoil. Many herbicides used in arable crops are applied either to bare soil prior to emergence of the sown crop or to the seed bed immediately postgermination, and so most of the compounds directly reach the soil surface. In contrast, insecticides and fungicides are usually applied later in the season when the crop may have closed its canopy, and so most compound will be intercepted by the crop foliage; even so, some spray will drift down, and wash-off by rain may subsequently move more compound to the soil. The rates at which pesticides are applied vary widely: a typical arable field in the UK in 2003 received approximately 4 kg ha$^{-1}$ of pesticide active ingredients in total per annum, though an intensively sprayed crop such as potatoes averages 15 kg ha$^{-1}$.

To understand the behavior of so many diverse compounds, it is necessary to develop an overarching framework of understanding based on the physicochemical properties and degradative behavior of each compound. The behavior of new pesticides can then be assessed by comparison or benchmarking with current compounds of similar properties. Important aspects of pesticide behavior in soil to be considered are persistence, sorption, movement with its potential for contamination of water, uptake by plants, and possible impacts on soil biota and microbial processes.

## Physicochemical Properties of Pesticides

The distribution of pesticides amongst the different environmental compartments, including the soil phases and the organisms living both in and on soil, is largely determined by the physicochemical properties of the compound. For example, vapor pressure determines the propensity of a pesticide to be lost by volatilization from a soil surface. If the soil is wet, then the Henry Law constant (the distribution coefficient between water and air) gives a better predictor:

$$H' = \frac{T}{273} \cdot \frac{\mathrm{VP}\,10^{-5}}{22.4\mathrm{WS}} \qquad [1]$$

where $H'$ is the dimensionless Henry Law constant, VP is the vapor pressure (in pascals), and WS is the water solubility (in moles per liter) measured at temperature $T$ (Kelvin). Organic compounds with high $H'$ will have high concentrations in the vapor phase compared with the soil water. As diffusion in air is approximately $10^4$ times faster than in water, then this region defines the boundary between movement predominantly via the air and via the water phase in soil (**Figure 1**).

Water solubility is an easily understood and oft-quoted property. However, lipophilicity, essentially a measure of the partitioning between water and lipid-like organic phases, is a better predictor of many processes, such as accumulation by organisms. To have comparable measurements, lipophilicity is

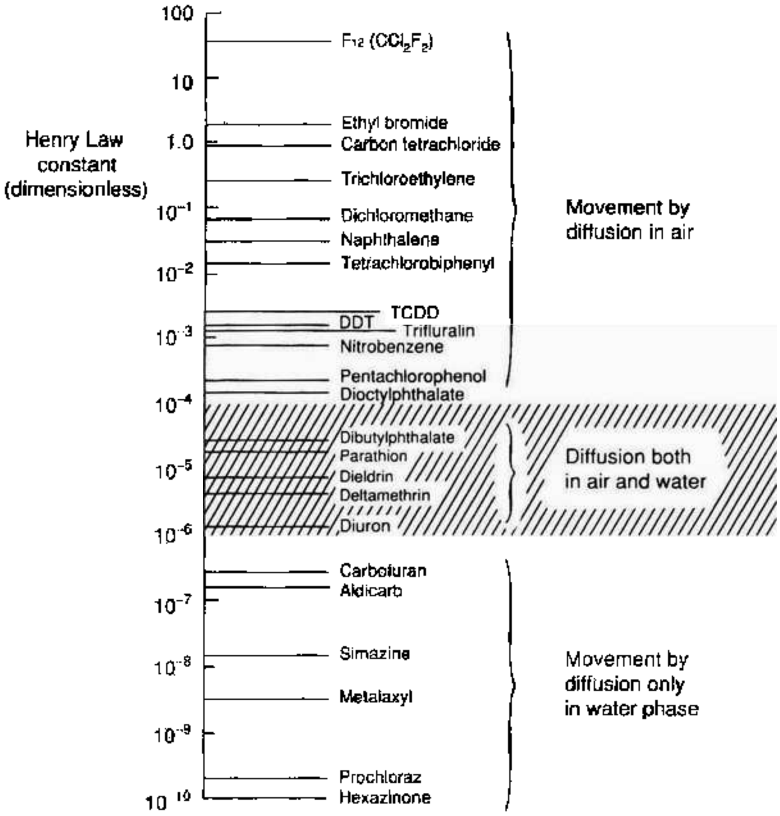


**Figure 1** Pathways of movement of pesticides and environmental contaminants by Henry Law constants. Reproduced with permission from Trapp S and Mc Farlane JC (eds) (1995) *Plant Contamination, Modeling and Simulation of Organic Chemical Processes*. Boca Raton, FL: Lewis Publishers. Copyright Lewis Publishers, an imprint of CRC Press, Boca Raton, FL.

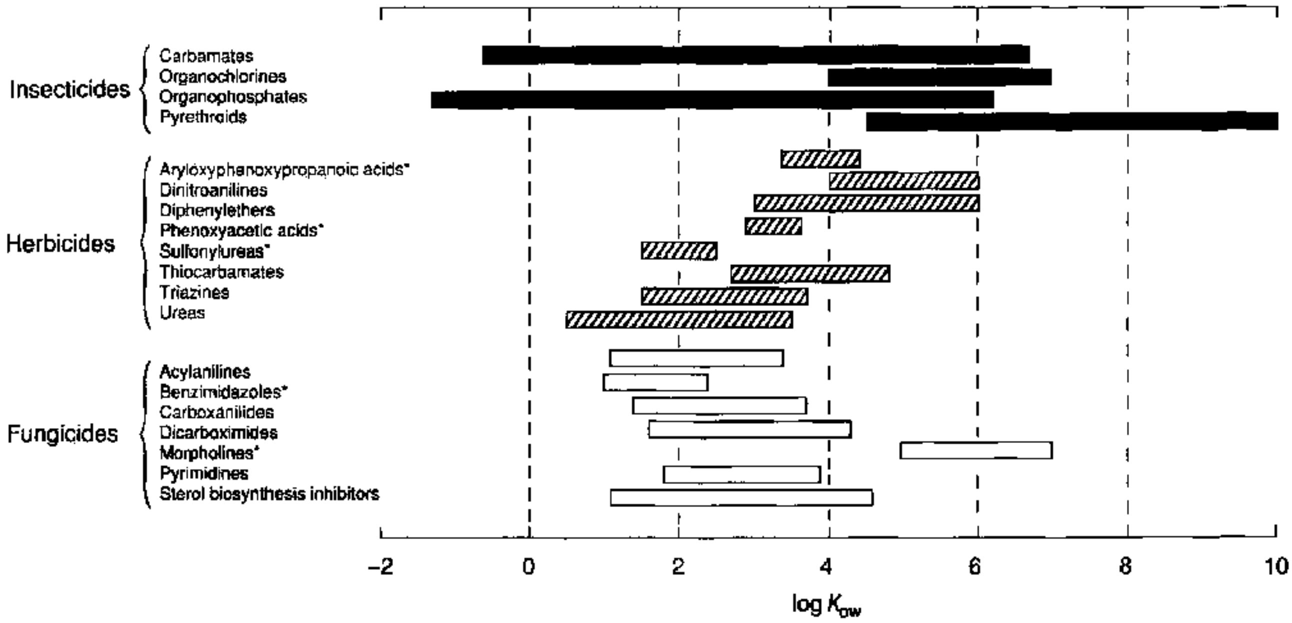


**Figure 2** Ranges of 1-octanol/water partition coefficients (as log $K_{ow}$) for commonly occurring compounds in various classes of pesticides. Asterisks indicate ionizable compounds, whose log $K_{ow}$ values are plotted for the undissociated molecule. Adapted from Trapp S and Mc Farlane JC (eds) (1995) *Plant Contamination, Modeling and Simulation of Organic Chemical Processes*. Boca Raton, FL: Lewis Publishers with permission. Copyright Lewis Publishers, an imprint of CRC Press, Boca Raton, FL.

normally assessed as the 1-octanol/water partition coefficient, $K_{ow}$ (usually expressed as $\log_{10} K_{ow}$). Some pesticide classes span quite a narrow range of lipophilicity, whereas others span up to 7 orders of magnitude of $K_{ow}$ (**Figure 2**). For compounds that are acids or bases, lipophilicity will depend on the balance of nonionized and ionized forms present; this is determined by the pH and the compound $pK_a$ (the $pK_a$ is the pH at which a particular acid or base function is 50% ionized). Ionized compounds are much less lipophilic (i.e., much more polar) than their undissociated forms.

Rates of degradation influence the depth of penetration into soil and degree of accumulation by biota; these rates generally need to be measured, for only approximate and often uncertain estimates can be made knowing the chemical structure.

## Persistence in Soil

### Loss Processes and Factors of Influence

Degradation is the main loss process for most pesticides in soil. Initial breakdown may be by biotic or abiotic processes, the latter including photolysis as well as chemical reactions such as hydrolysis and oxidation. However, the further extensive breakdown of complex fragments such as phenyl rings will be an essentially biotic process, with many species of microfungi, bacteria, and actinomycetes able to participate in the various steps. Thus a pesticide reaching soil will be subject to breakdown, leading to a sequence of metabolites or degradation fragments whose ultimate fate is mostly to be oxidized to $CO_2$ or incorporated into the polymeric matrix of the soil organic matter. This organic matter is itself constantly undergoing slow degradation, balanced by input of fresh organic matter such as crop residues such that an equilibrium level is reached over a number of years.

In soils, the rate of pesticide breakdown under constant conditions of temperature and moisture usually approximates to first-order kinetics, i.e., the rate of decomposition is proportional to the amount remaining:

$$\frac{dc}{dt} = -kc \qquad [2]$$

where $c$ is pesticide concentration, $t$ is the time, and $k$ the rate constant. This is akin to radioactive decay, and by analogy a half-life can be proposed. The half-life ($t_{1/2}$) is estimated by plotting the logarithm of the amount of compound remaining against time, which should approximate to a straight line (**Figure 3a**). Integrating eqn [1], the rate constant $k$ (units are per time) can then be calculated from:

$$k = \frac{\ln 2}{t_{1/2}} \qquad [3]$$

In the field, however, soil conditions of temperature and moisture can vary considerably over the lifetime of a pesticide, and this influences loss rates. Within

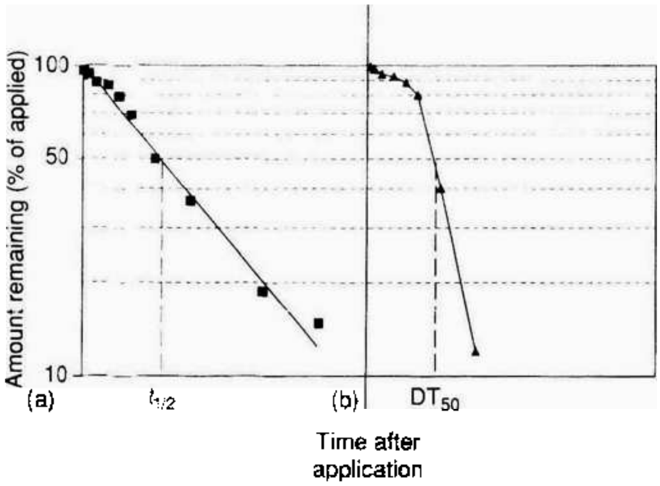


**Figure 3** Kinetics of pesticide degradation in soil under constant conditions (a) first-order kinetics showing the estimation of half-life; (b) microbial adaptation, showing a lag phase followed by rapid breakdown.

the normal range of ambient temperatures, a 10°C rise in temperature increases breakdown rates by two- to threefold both for abiotic and biotic degradation. In frozen soils, breakdown is usually very slow or negligible.

Most of the breakdown of pesticides by microorganisms is incidental to the general breakdown and turnover process of organic matter in soil, and so does not change the microbiota. The quantity of organic material present in the pesticide is normally a minute fraction of that being decomposed naturally in soil. However, occasionally certain pesticides or their metabolites, despite their low concentrations in soil, can be used as a food source by one or more species of microorganisms, and this is known as 'microbial adaptation.' The kinetics of breakdown then follow an initial lag phase, during which the population of the adapted organism increases until it reaches such levels that the pesticide is very rapidly degraded (Figure 3b). A second application would be immediately degraded. It should be noted that certain pesticides (e.g., preemergence herbicides and soil insecticides) need to persist in soil for a reasonable length of time if they are to be effective (e.g., several weeks), and such microbial adaptation with its concomitant rapid breakdown has thus limited the efficacy of some compounds, e.g., thiocarbamate herbicides and the carbamate insecticide carbofuran used to treat against corn rootworm (*Diabrotica* spp.).

Soil-moisture content influences both the availability of a pesticide for breakdown and the populations and activities of the microorganisms. It is thought that only compounds in the soil water are freely available to microorganisms for breakdown, and that sorbed compounds only become available as desorption into the soil water occurs to reestablish the sorption equilibrium after such breakdown. Thus, for example, the strongly sorbed herbicide paraquat is very persistent in soil, with a half-life of several years, but in culture medium without soil it can be rapidly degraded by microorganisms. As soils dry, the proportion of compound in the soil water will be decreased and also the activity of microorganisms will be increasingly limited, both of these processes causing a strong slowing of breakdown in drier soils.

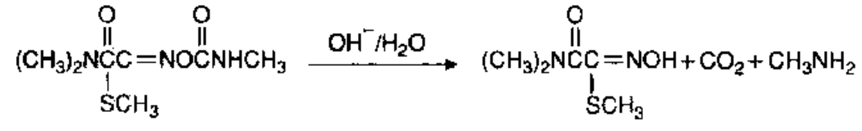
It should be noted that these two factors of temperature and moisture can often serve to cancel each other out in the field, because a sunny spell of weather causes both drying and warming of the surface layer of the soil, which often contains most of the pesticide. Thus breakdown in the field may approximate to first-order kinetics despite the changing environmental conditions caused by weather. However, losses of pesticide in the field can be biphasic, with a more rapid loss in the first few days of application than subsequently; this initial loss is usually attributed to surface processes such as volatilization or photolysis, and the first major rainfall event will move the compound sufficiently into the soil to block photolysis and reduce volatilization. Given the differing processes and variable conditions in the field, it is usual to describe pesticide persistence here by the $DT_{50}$ or $DT_{90}$ values, the disappearance time for 50% or 90%, respectively, of the parent compound; these are measured values with no implications as to the processes or kinetics involved.

Other factors can also influence breakdown rates of pesticides in soil. Increasing organic matter content encourages microbial activity but reduces pesticide availability in the soil water due to stronger sorption. Again these two factors are working in opposite directions, though, in poor sandy soils with little organic matter, degradation can be much slower than in more fertile soils.

Soil pH can also influence breakdown. For example, carbamoyloxime nematicides such as aldicarb and oxamyl are sensitive to hydrolysis at alkaline pH, and so these compounds are not very persistent in soils of pH >7.5 (**Figure 4a**). Conversely, the first generation of sulfonylurea herbicides such as chlorsulfuron (**Figure 4b**) are subject to acid hydrolysis and so do not persist in soils of low pH.

Microbial activity also usually diminishes with soil depth, resulting in breakdown in subsoil usually being slower than in topsoil. Nevertheless, microorganisms can survive to tens of meters depth, in numbers sufficient to cause slow degradation of pesticides in aquifers. Where aquifers are deep and the recharge takes decades, such degradation may remove pesticide traces before the water is abstracted. However,

(a) Oxamyl



$$(CH_3)_2N\overset{O}{\overset{\|}{C}}\underset{\underset{SCH_3}{|}}{C}=NO\overset{O}{\overset{\|}{C}}NHCH_3 \xrightarrow{OH^-/H_2O} (CH_3)_2N\overset{O}{\overset{\|}{C}}\underset{\underset{SCH_3}{|}}{C}=NOH+CO_2+CH_3NH_2$$

(b) Chlorsulfuron

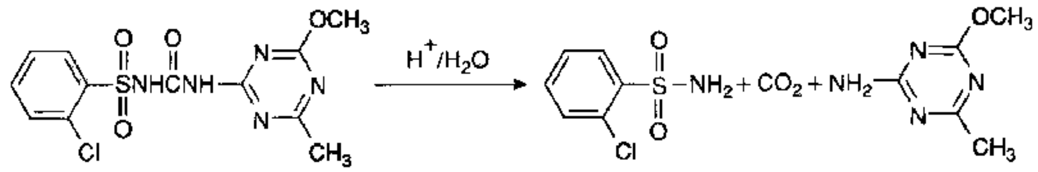


**Figure 4** Hydrolysis of pesticides in soil: (a) alkali-catalyzed hydrolysis of oxamyl; (b) acid-catalyzed hydrolysis of chlorsulfuron.

shallow aquifers are inevitably more susceptible to contamination, and much monitoring is done to ensure that widespread contamination of groundwater does not occur.

Another factor in the breakdown process is whether soils are aerobic or anaerobic (i.e., lacking in oxygen). Anaerobic conditions are brought about by prolonged flooding, in which breakdown of soil organic matter uses up the oxygen. An example occurs in the cultivation of rice in paddy fields. This leads to strong reducing conditions, which can cause new degradation pathways to become predominant. Examples are the reduction of aromatic nitro compounds to anilines and of ketones to secondary alcohols. A further aspect is that ferric ions are reduced to ferrous ions under such conditions, and the latter can catalyze the breakdown of some pesticides such as the carbamoyloxime nematicides.

### Methods for Studying Pesticide Breakdown

The breakdown pathways of pesticides are identified by incubating radiolabeled pesticide in soil under controlled laboratory conditions. The most usual radiolabel is $^{14}$C, the pesticide being synthesized from an appropriate $^{14}$C-labeled precursor such that the $^{14}$C is in the portion of the molecule likely to be the most recalcitrant to degradation. Such core labeling ensures that most of the metabolites will contain the $^{14}$C atom and so can be traced and isolated by thin-layer chromatography or high-pressure liquid chromatography. They can then be identified by techniques such as nuclear magnetic resonance spectroscopy and mass spectrometry. Usually such incubations are done in contained systems such that $^{14}CO_2$ or other volatile radiolabeled compounds produced can be trapped and estimated by liquid scintillation counting. After exhaustive extraction of the soil, typically with a succession of organic solvents increasing in polarity, $^{14}$C incorporated in the soil organic matter can be estimated by combustion of the extracted soil with trapping of the evolved $^{14}CO_2$. Adding up the $^{14}$C found in these various fractions at each time gives the mass balance which, if the experiments have been done satisfactorily, should be close to the dose applied to the soil at the beginning of the incubation.

Subsequent field studies are usually done with nonradiolabeled compound, with measurement of the parent compound and any major metabolites of interest by high-pressure liquid or gas-liquid chromatography. These techniques need to be very sensitive as some modern pesticides are applied at rates of only $10\,\mathrm{g\,ha^{-1}}$ or even less.

## Sorption of Pesticides by Soil

The sorption of pesticides to soil is a very important process in determining pesticide behavior in the environment. For example, strong sorption limits movement by leaching, and so such compounds would be unlikely to contaminate groundwater. It is not usually possible to distinguish between adsorption (a surface phenomenon) and absorption (whereby the pesticide penetrates the sorbing matrix such as soil organic matter), and hence use of the term 'sorption,' which covers both processes.

Sorption is typically measured by equilibrating sieved soil with an aqueous solution of the pesticide over a period of 4–24 h. The sorption coefficient ($K_d$ ml g$^{-1}$) is then calculated from:

$$K_d = \frac{\text{concentration in soil } (\mathrm{g\,g^{-1}})}{\text{concentration in water } (\mathrm{g\,ml^{-1}})} \qquad [4]$$

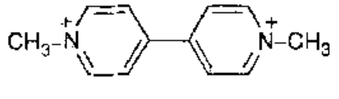
Sorption of nonionized and acidic pesticides is largely by partitioning into the soil organic matter. This process is viewed as essentially a reversible equilibrium, though detailed measurements usually reveal that desorption is somewhat slower than sorption. To allow for comparisons in sorption behavior over soils ranging in organic matter content, $K_d$ can be converted into $K_{om}$ (sorption coefficient on to soil organic matter; ml g$^{-1}$) or $K_{oc}$ (sorption coefficient on to soil organic carbon; ml g$^{-1}$) by:

$$K_d = \frac{\%\ \text{organic matter}}{100} K_{om} = \frac{\%\ \text{organic carbon}}{100} K_{oc} \qquad [5]$$

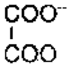
Such partitioning is largely controlled by the lipophilicity of a pesticide, and various relationships have been proposed to predict sorption from lipophilicity as assessed by the 1-octanol/water partition coefficient ($K_{ow}$) discussed above. Briggs proposed:

$$\log K_{om} = 0.52 \log K_{ow} + 0.62 \qquad [6]$$

Though simplistic, this approach works reasonably well for most pesticides. However, certain compounds have functional groups that confer additional sorption by specific interactions. For example, compounds carrying a positive charge can be sorbed by displacing metal ions in the soil matrix, a process known as cation exchange. A particularly striking example occurs with the bipyridinium herbicides such as paraquat, which is a planar molecular carrying two positive charges (**Figure 5a**). This is strongly sorbed by both organic matter and clay particles; in soils containing the expanding-lattice clay montmorillonite, paraquat can be intercalated between the layers comprising the crystal lattice and is then very tightly bound. A further mechanism of binding is chelation to metal ions such as iron and aluminum, which is observed with oxalic acid, a simple dicarboxylic acid that is dissociated to the oxalate salt at soil pH (**Figure 5b**). Another example is the herbicide glyphosate, which carries both negative and positive charges (**Figure 5c**). Though very polar and water-soluble, it is sorbed quite strongly by soil, owing to chelation by the charged groups.

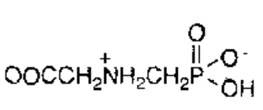


(a) Paraquat



$$\begin{array}{c}\mathrm{COO^-}\\ |\\ \mathrm{COO}\end{array}$$

(b) Oxalate



$$\mathrm{OOCCH_2\overset{+}{N}H_2CH_2P(=O)(O^-)OH}$$

(c) Glyphosate

**Figure 5** Charged compounds undergoing strong sorption by soil: (a) paraquat; (b) oxalate anion; and (c) glyphosate.

## Movement of Pesticides in Soil

In soil, a pesticide is distributed between the soil solids, soil water, and air phases. Most pesticides, with the notable exception of fumigants, are not very volatile, and so movement of most compounds will be via the soil water rather than via the air phase. Movement can be vertically down the soil profile or laterally. Vertical movement, usually referred to as leaching, can lead to contamination of groundwater if the pesticide is weakly sorbed and persistent. Such leaching can be through the bulk of the soil if the soil is not structured as in many sandy soils; in strongly structured clay soils, movement of pesticides in soil water can occur through cracks or pores such as worm or root channels (often termed 'macropore flow' or 'bypass flow'). As such structured soils are often drained, macropore flow can lead to rapid movement of small amounts of pesticide to drainage, which then contaminates surface waters such as streams and rivers. Lateral movement to surface waters can also occur by surface runoff in regions subject to intense rainfall events. Several general classifications of pesticide mobility in soil based on sorption $K_{oc}$ values have been proposed (and which can be related to $K_{om}$, $K_d$, and $K_{ow}$); one such example is given in **Table 1**. This is primarily of use for non-ionized pesticides, because sorption of ionizable pesticides may depend on soil pH, and shows that increasing lipophilicity decreases mobility.

**Table 1** Classification of pesticide mobility in soil

| *Mobility class* | *Log $K_{ow}$* | $K_{oc}$ *(ml g$^{-1}$)* | $K_{om}$ *(ml g$^{-1}$)* | $K_d$ *(ml g$^{-1}$)$^a$* |
|---|---|---|---|---|
| Nonmobile | >5.3 | >4000 | >2320 | >46 |
| Slightly mobile | 3.5–5.3 | 500–4000 | 290–2320 | 5.8–46 |
| Moderately mobile | 2.0–3.5 | 75–500 | 43–290 | 0.86–5.8 |
| Mobile | 0.6–2.0 | 15–75 | 9–43 | 0.18–0.86 |
| Very mobile | <0.6 | <15 | <9 | <0.18 |

[a] $K_d$ values given for a soil with 2% organic matter. (Source: Hollis JM (1991) Mapping the vulnerability of aquifers and surface waters to pesticide contamination at the national/regional scale. In: Walker A (ed.) *Pesticides in Soils and Water*. BCPC monograph 47: 165–174. Silsoe, UK. National Soil Resources Institute, Cranfield University.)

Leaching of pesticides is often studied using lysimeter systems. These are undisturbed blocks of soil, typically 1 m deep and $0.5-1.0\,\mathrm{m}^2$ in area, contained within a casing usually made of stainless steel or fiberglass. Such lysimeters can be taken from different soil types and then brought back to a central location for study. Radiolabeled pesticide is applied, and leachate from the lysimeters is monitored over a 2-year period. Any detected radioactivity is subjected to chromatography to identify it as parent compound or metabolite. In assessing the risk of pesticide leaching to groundwater, in the European Union, for example, the mean concentration of pesticide leached from such lysimeters would not be allowed to exceed $0.1\,\mu\mathrm{g\,l}^{-1}$, the drinking-water limit for every pesticide. This is a strict criterion that precludes the registration of many uses of pesticides. At the end of the experiment, soil cores can be taken to measure remaining amounts of pesticide and metabolites, as well as to assess the amount of pesticide incorporated into the soil organic matter.

The behavior of pesticides in soil can be simulated by modeling techniques, and a number of such computer programs are now available. These vary in their complexity, and it can be difficult to parameterize the more detailed models. At a minimum, simulation of pesticide behavior in soil requires as input a sorption coefficient, degradation constant, climate data (temperature, rainfall), and soil properties to allow description of water flow through the soil. Such models can simulate well the behavior of the bulk of pesticide in soil, but it has proved difficult to obtain reliable estimates of variable processes such as losses to drainage via macropore flow.

## Uptake of Pesticides from Soil by Plants

Plant roots and stem bases are in contact with the air and water in soil, and so can acquire pesticide present in these phases. Furthermore, crop plants typically transpire water amounting to several times their own weight each day, and so there is substantial mass flow to the roots of soil water and with it a potential burden of pesticide. In practice, the proportion of soil-applied pesticide found in plants is usually small, due in part to the pesticide being degraded and sorbed in soil and also often to it being in the uppermost layers of the soil above the main rooting zone. Nevertheless, pesticide uptake by plants can be important, examples being the carryover of herbicides from one crop to the next so causing phytotoxicity, and secondly the occurrence of pesticide residue in the edible portion of crops at harvest.

Only a few compounds move predominantly via soil air rather than via the water phase, these more volatile compounds being characterized as having the (dimensionless) Henry Law constant greater than $10^{-4}$. This discussion will thus focus on uptake into plants via the soil water, the predominant pathway for most pesticides. Uptake can be divided into two processes: (1) uptake into roots, and (2) transport to shoots via the transpiration stream flowing up the xylem vessels. Entry into roots for nonionized pesticides is primarily an equilibration from the soil water into the aqueous phase of the root cells, with partitioning on to the lipid-like constituents of roots. This latter process is determined by pesticide lipophilicity and is the main uptake process for compounds of $\log K_{\mathrm{ow}} > 2$. Compounds of $\log K_{\mathrm{ow}} > 4.5$ can be accumulated at concentrations greater than 100-fold (on a wet weight basis) compared with those of a bathing solution akin to soil water. As soils increase in organic matter content, stronger sorption reduces pesticide concentrations in the soil water and so uptake by plants is less. Translocation of such nonionized chemicals to shoots is, however, in no way related to uptake by roots. Such transport is a passive process, limited by the rate at which compounds can cross the root endodermis, which is a cylindrical sheath of cells surrounding the vascular tissues. Translocation efficiency is related to pesticide lipophilicity, being optimal at $\log K_{\mathrm{ow}} \sim 1.8$, at which compounds move with approximately 80% of the efficiency of water, with more polar and more lipophilic compounds becoming increasingly less well transported.

Uptake behavior of ionized compounds is more complex. The most important group is composed of the weak acid herbicides; for these, uptake by roots and transport to shoots are proportional, increasing strongly in more acidic soils because the undissociated forms are much better taken up than their respective anions.

## Effects on Soil Processes

The possible impact of pesticides on soil processes mediated by microorganisms, together with that on larger biota such as earthworms, has to be extensively studied before a pesticide is registered. Compounds that are applied to sterilize soil, known as 'soil sterilants' or 'fumigants,' are generally applied at a high rate and are volatile compounds such as methyl bromide (whose use will be phased out) or methyl isothiocyanate that are able to move in the soil air spaces throughout the top soil. These, as their name implies, kill most organisms in the treated zone, though recolonization can occur subsequently from the deeper soil layers that would be less affected.

Other pesticides, however, seem to have relatively little effect on microbially mediated soil processes such as nitrification or respiration. Although microfungi are an important constituent of the soil microorganisms, fungicides at normal rates seem to have little effect on soil processes, even when thoroughly mixed into soil; and, in the field, only a shallow layer of topsoil would normally contain most of the fungicide reaching the soil surface following foliar sprays. Larger organisms such as earthworms and ground-dwelling beetles can be affected by some insecticides and nematicides, but such effects at the population level are usually short-lived.

Long-term trials over 20 years or more with repeated applications of single or combinations of pesticides applied at normal agronomic rates have found no deleterious effects on soil fertility, as assessed by crop yield, or on soil processes such as respiration and levels of microbial biomass carbon. These findings reinforce the more general observation that problems of soil fertility in worldwide agriculture have not occurred despite several decades of use of pesticides. This shows that the soil is quite robust, no doubt because its biota have been decomposing large amounts of complex organic compounds from plants for millions of years. These residues, and those from organic pesticides, are so degraded by the time that they are incorporated into soil organic matter that any original biological activity has been lost. There is thus no reason to believe that use of pesticides is incompatible with sustainable agricultural systems.

## List of Technical Nomenclature

| | |
|---|---|
| $c$ | Concentration |
| $DT_{50}$, $DT_{90}$ | Disappearance time (50%, 90%) |
| $H'$ | Henry Law constant (dimensionless) |
| $k$ | First-order rate constant (per time) |
| $K_d$ | Sorption coefficient to soil (milliliters per gram) |
| $K_{oc}$ | Sorption coefficient to soil organic carbon (milliliters per gram) |
| $K_{om}$ | Sorption coefficient to soil organic matter (milliliters per gram) |
| $K_{ow}$ | 1-Octanol/water partition coefficient (dimensionless) |
| $T$ | Temperature (Kelvin) |
| $t$ | Time |
| $t_{1/2}$ | Half-life |
| VP | Vapor pressure (pascal) |
| WS | Water solubility (moles per liter) |

*See also:* **Pollutants:** Biodegradation; Effects on Microorganisms; Persistent Organic (POPs)

## Further Reading

Briggs GG (1981) Theoretical and experimental relationships between soil adsorption, octanol–water partition coefficients, water solubilities, bioconcentration factors, and the parachor. *Journal of Agricultural and Food Chemistry* 29: 1050–1059.

Bromilow RH, Evans AA, Nicholls PH, Todd AD, and Briggs GG (1996) The effect on soil fertility of repeated applications of pesticides over 20 years. *Pesticide Science* 48: 63–72.

Führ F and Hance RJ (eds) (1992) *Lysimeter Studies of the Fate of Pesticides in the Soil.* Monograph No. 53. Farnham, Surrey: British Crop Protection Council.

Goring CAI and Hamaker JW (eds) (1972) *Organic Chemicals in the Soil Environment*, vols 1 and 2. New York: Marcel Dekker.

Gustafson DJ (1989) Groundwater ubiquity score. *Environmental Toxicology and Chemistry* 8: 339–357.

Hartley GS and Graham-Bryce IJ (1980) *Physical Principles of Pesticide Behaviour,* vols 1 and 2. London, UK: Academic Press.

Howard PH (ed.) (1991) *Handbook of Environmental Fate and Exposure Data for Organic Chemicals,* vols 1–3. Boca Raton, FL: Lewis Publishers.

Kearney PC and Roberts T (eds) (1998) *Pesticide Remediation in Soil and Water.* Colchester, UK: John Wiley.

Plimmer J (ed.) (2002) *Encyclopedia of Agrochemicals.* New York: John Wiley.

Tomlin CDS (2000) *The Pesticide Manual,* 12th edn. Farnham, Surrey: British Crop Protection Council.

Trapp S and McFarlane JC (eds) (1995) *Plant Contamination: Modeling and Simulation of Organic Chemical Processes.* Boca Raton, FL: Lewis Publishers.

## Petroleum *See* Hydrocarbons

# pH

**N S Bolan and K Kandaswamy**, Massey University, Palmerston North, New Zealand

No other single chemical soil characteristic is more important in determining the chemical environment of higher plants and soil microbes than the pH. There are few reactions involving any component of the soil or of its biological inhabitants that are not sensitive to soil pH. This sensitivity must be recognised in any soil-management system (Brady NC and Weil RR (1999) *Nature and Properties of Soils*. Englewood Cliffs, NJ: Prentice-Hall).

## Introduction

The acidity–alkalinity status of a soil is referred to as 'soil reaction,' which is often quantified by measuring soil pH. pH is considered to be the 'master variable' of a soil, which determines many of the physical (e.g., flocculation or deflocculation), chemical (e.g., adsorption or desorption), and biological (e.g., mineralization or immobilization) properties of soils. In this chapter, the fundamentals of soil pH will be discussed in relation to its measurement and practical implications.

## Theoretical Basis of pH

Pure water dissociates into the hydronium ion $(H_3O)^+$, which is associated with acidity, and hydroxyl $(OH)^-$, which is connected with alkalinity:

$$H_2O + H_2O \rightarrow (H_3O)^+ + (OH)^- \qquad [1]$$

In aqueous media, the free acid cation is mostly present as the hydrated proton $(H_3O)^+$ but for convenience it is often referred to as a hydrogen ion or proton $(H^+)$, and thus the dissociation of water is often represented simply as:

$$H_2O \rightarrow H^+ + OH^- \qquad [2]$$

This is a reversible reaction which obeys the law of mass action:

$$[H^+] \times [OH^-]/[H_2O] = K \qquad [3]$$

where $K$ is a constant. Because the degree of dissociation of pure water is so small, the concentration of undissociated $H_2O$ molecules always remains constant, so that the product $[H^+] \times [OH^-]$ gives a constant value $(K_w)$, known as the dissociation constant or ionization constant of water. The $K_w$ value for pure water at 21°C is $1 \times 10^{-14}$, with equal concentration of $H^+$ and $OH^-$ ions ($1 \times 10^{-7}$ gEq):

$$K_w = [H^+][OH^-] = 1.0 \times 10^{-14} \qquad [4]$$

If a chemical substance is added to water, which either donates or consumes $H^+$ or $OH^-$ ions, then the concentration balance between $H^+$ and $OH^-$ ions is disturbed. The numerical product of the concentrations must, however, remain constant, so that if $[H^+]$ increases, $[OH^-]$ must decrease accordingly, and vice versa.

To simplify the arithmetical representation of eqn [4], Sörensen suggested that the equilibrium state could be defined simply by reference to the $[H^+]$ and, to avoid the use of cumbersome decimal fractions, he proposed the use of a term designated 'pH' (French *pouvoir hydrogène* or *puissance d'hydrogene*, or 'hydrogen power') which is defined as the $\log_{10}$ of the reciprocal of the $[H^+]$ measured in gEq:

$$\mathrm{pH} = \log_{10}\frac{1}{H^+} \text{ or } -\log_{10}(H^+) \qquad [5]$$

Expressing eqn [4] as a negative logarithm yields:

$$\mathrm{pH} + \mathrm{pOH} = 14 \qquad [6]$$

where pH and pOH are the negative logarithms of the $H^+$ and $OH^-$ ion activities, respectively.

Thus, the pH of an aqueous solution is defined as the negative logarithm of the $H^+$ ion activity in the solution, which is the same as the $[H^+]$ when the salt concentration in the solution is very low. This concept can be applied to a system in which the individual molecules and ions are uniformly dispersed throughout the solution within the limits set by variations due to molecular and ionic thermal movements.

The pH scale ranges from 0 to 14, with pH 7 as the neutral point. At pH 7, $[H^+]$ equals the $[OH^-]$. From pH 7 to 0, the solution is increasingly more acidic; from pH 7 to 14, the solution is increasingly more alkaline (basic; **Figure 1**). The $[H^+]$, which is the ion measured when determining pH, has a 10-fold change between each whole pH number. Thus, a solution with pH 5 will have 10 times more $[H^+]$ than a solution with pH 6.

## Measurement of pH

pH is commonly measured either electrometrically using potentiometers or calorimetrically using pH-sensitive dyes.

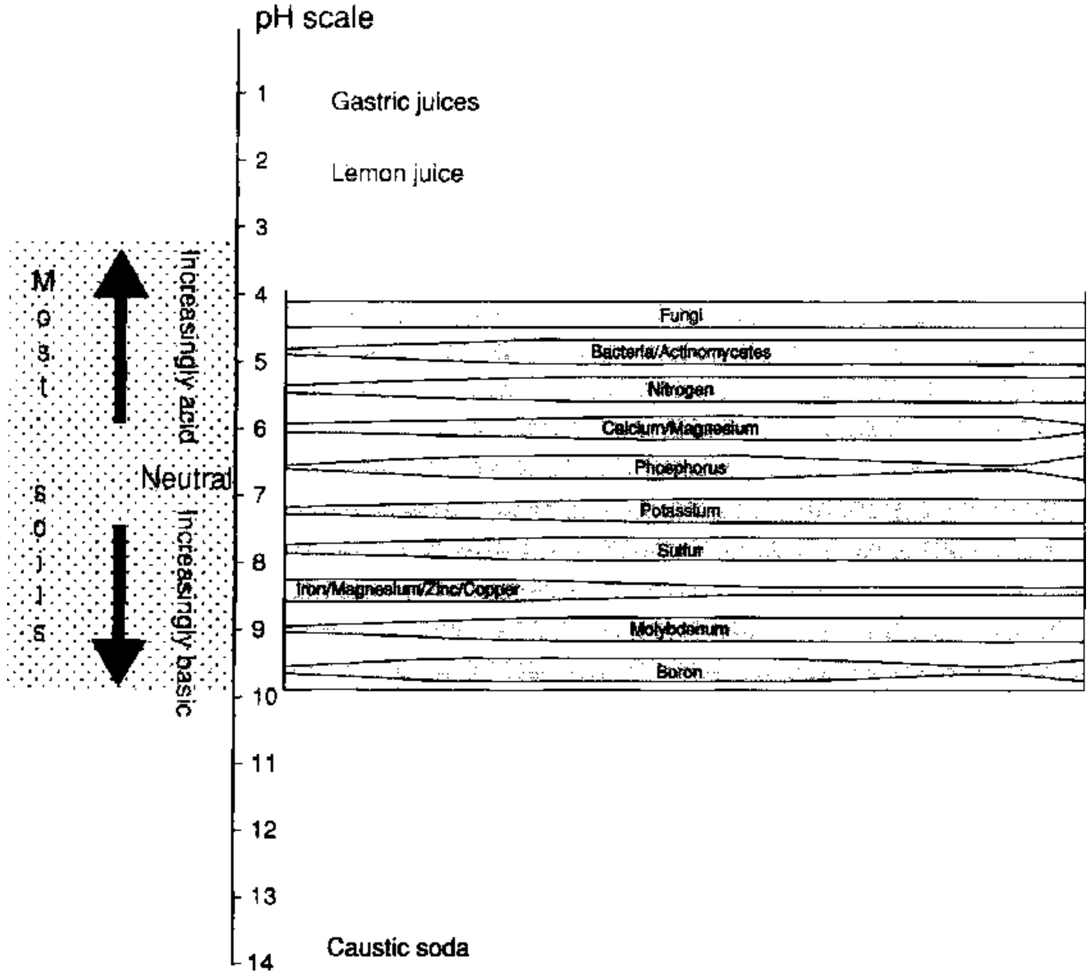


**Figure 1** pH scale and the effect of pH on nutrient availability and microbial activity.

### Potentiometric Method

This method involves the measurement of potential difference between an electrode whose potential is determined by the $[\mathrm{H}^+]$ of the solution and a reference electrode whose potential is independent of $[\mathrm{H}^+]$. The normal hydrogen electrode is the glass electrode, and the reference electrode is a calomel electrode in which calomel ($HgCl_2$) is bathed in a solution of potassium chloride (Figure 2).

The glass electrode consists of a thin glass bulb containing dilute HCl, into which is inserted an Ag–AgCl wire, serving as the electrode with a fixed voltage. The HCl solution is separated from the test solution by a membrane of special glass, usually a lithium silicate of particular composition. Differences in $\mathrm{H}^+$ activity across this membrane cause a difference in electrical potential, which can be measured by a potentiometer.

When the glass bulb is immersed in a solution, a potential difference develops between the solution in the bulb and the solution outside the bulb. The electrical potentials developed by this electrode are the membrane potential, plus the potential of the Ag–AgCl–HCl reaction inside the electrode:

$$\mathrm{AgCl} + e^- \leftrightarrow \mathrm{Ag} + \mathrm{Cl}^- \qquad [7]$$

This reaction is reversible and the Ag–AgCl potential ($E^0 = 0.222$ V at 25°C) is likely to remain constant. This potential is cancelled out when the electrode is standardized against a standard pH buffer solution.

The glass electrode develops a second potential at the membrane separating the standard HCl and the test solutions. The minute current flow required by the pH meter develops a double layer in the solutions, causes ion exchange at the inner and outer surfaces, and causes diffusion of ions across the glass membrane. The potentials across the glass membrane can be closely calibrated to the approximate value of the $\mathrm{H}^+$ activity.

A typical reference electrode consists for a wire dipped into the liquid mercury, which makes electrical contact with the pH meter. Current flows from the electrode to the solution phase through the reversible reaction ($E = 0.268$ V at 25°C):

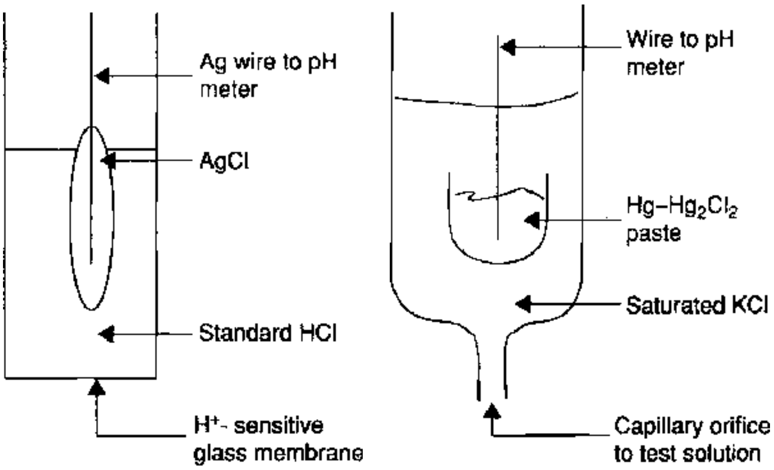


**Figure 2** Hydrogen (glass) and reference (calomel) electrodes. Reproduced from Tan KH (1998) *Principles of Soil Chemistry*. New York: Marcel Dekker.

$$\mathrm{Hg_2Cl_2} + 2\mathrm{e}^- \leftrightarrow 2\mathrm{Hg} + 2\mathrm{Cl}^- \qquad [8]$$

The $\mathrm{Cl}^-$ activity is fixed by the KCl concentration (usually saturated KCl). As a result this potential also cancels out when the pH electrode system is standardized in a standard pH buffer. KCl diffusion through the orifice makes the electrical contact between the reference cell and the test solution. This KCl connection forms a 'salt bridge' between the test solution and the reference electrode. The reference electrode is often built in the same body as the glass electrode to form a combination electrode.

When measuring pH, the potential ($E$) in the glass electrode is formulated by the Nernst equation:

$$E = (RT/nF)\log(K/\mathrm{H}^+) \qquad [9]$$

where $R$ is the gas constant, $T$ the absolute temperature, $n$ the valence, $F$ the Faraday constant, $K$ a constant, and $\mathrm{H}^+$ the activity of $\mathrm{H}^+$ ions. $E$ is called the 'half-cell potential' and cannot be measured alone. If the glass electrode is placed against a reference calomel electrode, the potential difference between the two ($E - E_{\mathrm{ref}}$) is measurable.

Before any pH measurement, the two electrodes have to be laced first in a solution of known pH. This is called 'standardizing' the electrodes and the pH meter. The overall potential of the total cell $E_0$ equals $E - E_{\mathrm{ref}}$:

$$E_0 = [(RT/F)\log(K/\mathrm{H}^+_{\mathrm{std}})] - E_{\mathrm{ref}} \qquad [10]$$

If the two electrodes are now placed in the solution with the unknown $[\mathrm{H}^+]$, the potential $E_c$ is:

$$E_c = [(RT/F)\log(K/\mathrm{H}^+_{\mathrm{test}})] - E_{\mathrm{ref}} \qquad [11]$$

Subtracting eqn [10] from eqn [11] gives:

$$E_c - E_0 = (RT/F)[\log\mathrm{H}^+_{\mathrm{test}} - \log\mathrm{H}^+_{\mathrm{std}}] \qquad [12]$$

$$E_c - E_0 = (RT/F)\log(\mathrm{H}^+_{\mathrm{test}}/\mathrm{H}^+_{\mathrm{std}}) \qquad [13]$$

or

$$E_c - E_0 = (RT/F)(\mathrm{pH_{std}}/\mathrm{pH_{test}}) \qquad [14]$$

where $E$ is the measured potential and is converted to pH units by the scale of the pH meter. $E_{\mathrm{ref}}$ includes all other potentials, which are nullified by standardizing the systems with a standard pH buffer solution.

Since $RT/F = 0.0591$ at 25°C:

$$\mathrm{pH} = (E_c - E_0)/0.0591 \qquad [15]$$

This means that for a change of 1 pH unit, the potential will change by 0.0591 V.

The glass electrode has a uniform response to a wide range of $\mathrm{H}^+$ activities and is less sensitive to interfering ions. The extent of interference is denoted by the selectivity ratio, the concentration ratio of the test ion to interfering ion at which the interfering ion exerts a significant potential at the membrane. The selectivity ratio of the pH glass membrane for $\mathrm{H}^+$ over $\mathrm{Li}^+$, the most serious interfering ion, is approximately $10^9$. That is, $\mathrm{Li}^+$ would cause a significant pH error in a pH 9 solution containing 1 mol $\mathrm{l}^{-1}$ $\mathrm{Li}^+$ or in a pH 10 solution containing 0.1 mol $\mathrm{l}^{-1}$ $\mathrm{Li}^+$. The selectivity ratio for $\mathrm{Na}^+$, the next most serious interference, is approximately $10^{13}$.

## Determination of pH Using pH-Sensitive Dyes

Rapid measurement of pH can also be achieved colorimetrically with pH indicator dyes. Colorimetric methods are based on the change in color that takes place upon dissociation of a weak acid, or weak base organic dye; thus for a weak acid dye:

$$\mathrm{HI} \rightleftharpoons \mathrm{H}^+ + \mathrm{I} \qquad [16]$$

where raising the pH will cause a change in color as the undissociated dye (HI) changes to the dissociated form ($I^-$). Typically a mixture of dyes is used to yield color changes that take place over a wide range of pH. However, the color can only be measured accurately in a solution from which the solid particles are removed, and the color also depends on the concentration of various salts present in the solution. A dye solution is mixed with the solid to make a slurry, the dye is decanted and the color is visually compared with a color chart. The pH measured in this manner is not precise and can vary from the results obtained using the glass electrode.

## Measurement of Soil pH

Because 'pH' is a term that is only defined for solution, in a strict sense, it cannot be applied to a solid-phase material such as soil. However, the chemical properties of the solid-phase components in soil define the pH of the soil solution (i.e., the solution in the soil pores). Further the pH of a soil dispersed in water is not a simple concept like that of a solution. When soil particles, which carry ions attached to them, are dispersed in solution, the ions are not uniformly distributed throughout the solution. The concept of the pH of a soil (or soil suspension) can therefore only be discussed in relation to the properties of the ionic atmosphere around the soil particles (i.e., soil–solution interface).

When a negatively charged soil particle is dispersed in water, the negative charge is neutralized by cations, some of which sit firmly on the clay surface, forming the Stern layer, and some dissociate into the dispersion medium to form the Gouy diffuse double layer. The thickness of this double layer depends in part on the ions dissolved in the dispersion medium, and the thickness increases with a decrease in the concentration of the ions, a decrease in the valency of the cations, and an increase in the size of the cation. The $[H^+]$ in the solution surrounding the soil particles is less than that close to the soil particles, due to the $[H^+]$ gradient in the double layer. As the double layer is made more compact by adding an electrolyte to the soil-water system, the $[H^+]$ gradient across the double layer is reduced, and the pH in the solution becomes almost equal to that close to the surface of the soil particle. Thus the pH of the bulk solution is greater than the pH just outside the Stern layer unless the salt concentration in the solution is high.

The pH of most soils ranges from 3 to 8.5. Soils do not normally have a pH below 3–4 unless they contain free acids (e.g., acid sulfate soils and acid mine drainage). Similarly soils normally only have a pH >8.45 if they contain enough exchangeable sodium (i.e., alkali soils). Further, in calcareous and alkali soils, the effective pH around plant roots can be determined by the concentration of carbon dioxide ($CO_2$) produced during respiration by roots and microorganisms.

Strictly speaking the pH of a soil should refer to the $[H^+]$ in the soil solution. This is very difficult to measure, and hence soil pH is often referred to measurements made in the soil suspension. Three common, standard methods for determining pH in soils involve the suspension in distilled water, 0.01 mol $l^{-1}$ $CaCl_2$, or 1 mol $l^{-1}$ KCl solutions at a soil-to-solution (weight-to-volume) ratio of 1:1 (in USA) or 1:2.5 (in New Zealand and Australia).

The pH of a solution in equilibrium with a soil varies with the composition and concentration of the salts in the solution. This is because cations in solution displace $H^+$ and $Al^{3+}$ ions from soil surfaces. The measure of soil pH in water extract is expected to give a higher value than the pH of the soil solution at natural soil water contents. The excess quantity of water added to prepare the suspension results in a dilution of the salts in the soil solution. Since soils always contain some soluble salts, such as nitrates and chlorides, the apparent pH of the soil will depend on the amount of water added to the soil to make the suspension. Further, the amount of soluble salts in a soil varies continuously throughout the season, depending on factors such as the amount of drainage water percolating through the soil.

Soil pH values measured in 0.01 mol $l^{-1}$ $CaCl_2$ or 1 mol $l^{-1}$ KCl salt solutions are generally less than the values obtained using water, because the cations in these salts tend to displace $H^+$ and $Al^{3+}$ from soil materials to soil solution. These salt solutions are sufficiently concentrated to overwhelm variations in salt concentration in soil solution, thereby minimizing the effect of salt concentration on pH in most soils. These solutions also eliminate the junction potential effect. Calcium is used in the determination of pH because it is the predominant soil solution cation in soils of temperate regions. The 0.01 mol $l^{-1}$ concentration, however, is somewhat higher than that found in most soil solutions, and thus the pH values obtained by this method are usually lower than the values of soil solutions.

For some subsurface horizons of highly weathered soils of the tropics, pH in salt solution can actually be greater than pH in water, especially in subsoils that are low in permanent-charge silicate clays and low in organic matter. Ion exchange in these soils is dominated by oxides and hydroxides of Fe and Al, and at acid pH values they are net positively charged, creating an effective sink for anion adsorption. Thus, on salt addition (i.e., $CaCl_2$ or KCl) more $OH^-$ is displaced by $Cl^-$ than $H^+$ by cations, resulting in an increase in pH.

## Factors Affecting Soil pH

The pH of a soil is affected by the concentration of $CO_2$ in the soil air, salt concentration (salt effect), and the presence of colloidal particles (suspension effect).

### Carbon Dioxide

The higher the $CO_2$ concentration in the soil solution, the lower the pH, and the pH of a neutral or calcareous soil is very sensitive to small changes in $CO_2$ concentration. In calcareous soils, its magnitude can be calculated from the solubility of calcium carbonate and $CO_2$ in water, because in such soils the pH is basically determined by the system, calcium carbonate–carbon dioxide–water, which is given approximately by:

$$2\mathrm{pH} = K + \mathrm{pCa} + \mathrm{pCO_2} \qquad [18]$$

where $\mathrm{pCO_2}$ is the negative logarithm of the partial pressure of $CO_2$ in equilibrium with the solution, pCa the negative logarithm of the activity of the calcium ions, and $K$ is a constant whose value lies between 10 and 10.5, depending on the value used for the solubility constant of calcium carbonate. The decrease in pH is therefore approximately proportional to the logarithm of the partial pressure of $CO_2$.

Thus, the pH of a well-aerated calcareous soil containing exchangeable sodium or magnesium ions cannot exceed approximately 8.5 when in equilibrium with the atmosphere; but if the $CO_2$ concentration rises to 0.1% (i.e., a low value for most soils) the pH drops to 8, and if it rises to 1% (a common value for pasture soils) the pH will be less than 7.5. Many plants suffer from phosphate and minor element deficiencies if the pH rises much above 8, and hence they can only thrive on calcareous soils if they can maintain an adequate concentration of $CO_2$ around their roots. This is more easily achieved in soils of low porosity, such as clays, than in soils of high porosity, such as well-drained sands, and in organic matter-rich pasture soils than in organic matter-depleted arable soils.

### Salt Effect

As discussed above, the pH of a soil depends on the salt concentration in the soil solution. This dependence of measured soil pH on salt concentration can be reduced by making all the measurements in a salt solution that is sufficiently strong to swamp the effects of the changes that occur naturally. But adding a salt solution to the soil will cause cation exchange to take place and, in particular, exchange of $H^+$ ions from clay surface with the cation of the added salt, which will give a pH that is lower than the soil pH under normal conditions. The added salt solution should therefore cause as little exchange as possible, and the use of $0.01\ \mathrm{mol\ l^{-1}}$ $CaCl_2$ or $1\ \mathrm{mol\ l^{-1}}$ KCl solution has been proposed for most arable soils on the grounds that they approximate to the salt concentration in the soil solution and do not give appreciable junction potential with the calomel electrode. The pH of a soil measured in these solutions is more constant and is much less dependent on the soil-to-solution ratio than the pH measured in a water suspension, and it is also presumably closer to the pH of the solution around the plant roots.

### Suspension Effect

The difference in pH values between the suspension and its equilibrium solution is referred to as the 'suspension effect,' and this effect is identical to the Donnan e.m.f.

The electrode portion of a pH measurement circuit is presented in **Figure** 3, where each vertical bar represents a phase boundary and the double bar represents the liquid–liquid junction. Thus, very appreciable junction potentials can be created between a soil suspension and the calomel electrode so that the measured potential difference between the two electrodes cannot be interpreted solely in terms of the $[H^+]$ outside the glass electrode.

A junction or diffusion potential always develops when two dissimilar substances or solutions of different composition come in contact. It is inevitable that a diffusion potential is created at the liquid junction between a salt bridge and a solution of different composition. An example in this case is the saturated KCl of the reference electrode diffusing into the test solution. At liquid junctions, differences in the mobilities of the cation and anion create a diffusion potential which is also influenced by any differences in the ion activities that exist across the junction. The potential is minimal when the rates of $K^+$ and $Cl^-$ diffusion into the test solution are equal. It is therefore appropriate that an electrolyte in which the mobilities of the cation and anion are almost identical is used in the salt bridge. Indeed, KCl is used as the reference-cell electrolyte because of the similarity of $K^+$ and $Cl^-$ diffusion rates in water. Because the current at the liquid junction is mainly carried by $K^-$ and $Cl^-$ ions, which have similar diffusion coefficients, the liquid

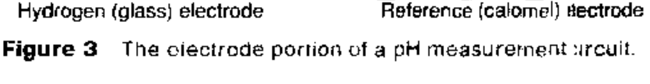


**Figure 3** The electrode portion of a pH measurement circuit.

junction potential should be small. As long as the ionic strengths of the standard and unknown solutions are small compared with the concentrated KCl salt bridge solution, any liquid junction potential should, as well as being small, tend to be rather constant in magnitude and hence can usually be ignored.

Because factors involved in the liquid junction potential during pH calibration and subsequent pH measurement are likely to be different, the liquid junction potential contribution to the total e.m.f. of a cell will also be different. In a colloidal suspension, however, the colloid may cause $K^+$ and $Cl^-$ to diffuse at different rates. Because of attraction and repulsion by the charged colloid, one ion moves ahead of the other. When a salt bridge makes contact through a suspension instead of clear solution, the situation is complicated by a reduced mobility of counter ions which arises from retarding effects created by the oppositely charged colloidal particles. This means that liquid junction potentials ($E_j$) are likely to be much greater in suspensions than in supernatant liquids. Accurate pH measurements require a negligible $E_j$, because such potentials are virtually unpredictable. The potentiometer measures all the potentials of the circuit and cannot distinguish between the $H^+$ ion potential at the glass membrane and the spurious $E_j$. The value of $E_j$ at the interface between saturated KCl and a colloidal suspension of very low salt concentration can be as high as 240 mV, equivalent to more than 4 pH units (1 pH unit = 59 mV). However, values of $E_j$ greater than 30 mV are uncommon in soils, because salt concentrations in even highly leached soils are high enough to suppress $E_j$. Therefore measuring pH in salt solutions of 0.01 mol $l^{-1}$ or more virtually eliminates $E_j$.

Suspension effects can be reduced by introducing the liquid junction of the reference electrode into clear supernatant liquid and not into the suspension. The rates of $K^+$ and $Cl^-$ diffusion are then unaffected by soil colloid particles. The pH glass electrode, on the other hand, can be placed either in the supernatant solution or in the colloidal suspension. The $H^+$ activity is the same in both phases, and the glass electrode is unaffected by the presence of the colloid. However, it has been suggested that when measuring soil pH the glass electrode should be in the suspension on the basis that the suspension is better buffered than the supernatant liquid.

## Importance of Soil pH

The soil pH is considered as the 'master variable' which controls many other soil properties (**Figure 1**). pH greatly affects the solubility of minerals in soils. Most minerals are more soluble in acid soils than in neutral or slightly basic solutions. A low pH may indicate the deficiency of basic cations such as $Ca^{2+}$ and $Mg^{2+}$ in soils. Under acid soil conditions, the concentration of most heavy metals, such as lead, cadmium, and mercury, reach phytotoxic levels. Strongly acidic soils usually have high and toxic concentrations of soluble aluminum and manganese. Azaleas, tea, rhododendrons, cranberries, pineapple, blueberries, and several conifer timber species tolerate acidity and grow well in acid soils. In contrast, alfalfa, beans, barley, and sugarbeets grow well only in slightly acidic to moderately basic soils because of a high calcium demand or inability to tolerate soluble aluminum.

Soil pH can also influence plant growth through its effect of beneficial microorganisms. For example, most nitrogen-fixing legume bacteria are not very active in strongly acidic soils. Bacteria that decompose soil organic matter and thus release nitrogen and other nutrients for plant use are also hindered by strong acidity. Similarly earthworms are not active in acid soils, whereas fungi usually tolerate acidity better than do other microbes.

Soil basicity, although more difficult to manage than soil acidity, may also be undesirable for plants. Alkaline soils may reach pH values of more than 10 with increasing levels of exchangeable sodium. Plants on soils with pH >9 usually have reduced growth or even die. Some plants (halophytes) are tolerant of high salt concentration or pH. Soluble $Al(OH)_4^-$ may exist at these high pH values and is also toxic to plants. The major effect of basic pH is to reduce the solubility of all micronutrients (except chlorine, boron, and molybdenum), especially those of iron, zinc, copper, and manganese. Also, phosphate is often not readily available to some plants because of its precipitation in the soil solution by calcium or precipitation on solid calcium carbonate. Iron deficiency, associated with soils high in carbonates, has long been known, and is referred to as 'lime-induced iron chlorosis.'

## Summary

Soil pH is one of the most important chemical properties which control many other soil physical, chemical, and biological properties. The pH of a soil depends on the $CO_2$ concentration in the soil air and the salt concentration in the soil solution, and these are constantly changing.

## Further Reading

Arnold PW (1978) Surface–electrolyte interactions. In: Greenland DJ and Hayes MHB (eds) *The Chemistry of Soil Constituents*, pp. 335–404. New York: John Wiley.

Bloom PR (2000) Soil pH and pH buffering. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. B333–352. New York: CRC Press.

Bohn H, McNeal BL, and Connor GA (1979) *Soil Chemistry*. New York: John Wiley.

Brady NC and Weil RR (1999) *Nature and Properties of Soils*. Englewood Cliffs, NJ: Prentice-Hall.

Hausenbuiller RL (1978) *Soil Science: Principles and Practices*, 2nd edn. Dubuque, Iowa: WC Brown.

Miller RW, Donahue RL, and Miller JU (1990) *Soils: An Introduction to Soils and Plant Growth*, 6th edn. New York: Prentice-Hall.

Russel EW (1973) *Soil Conditions and Plant Growth*, 10th edn. London, UK: Longman.

Sorenson SPL (1909) Enzyme studies. II. The measurement and importance of the hydrogen ion concentration in enzyme reactions. *Comptes Rendu du Travail Laboratoire de Carlsberg* 8: 1.

Sparks DL (1995) *Environmental Soil Chemistry*. New York: Academic Press.

Tan KH (1998) *Principles of Soil Chemistry*. New York: Marcel Dekker.

Wolt JD (1994) *Soil Solution Chemistry, Application to Environmental Science and Agriculture*. New York: John Wiley.

# PHOSPHORUS IN SOILS

Contents

## Overview

**J T Sims**, University of Delaware, Newark, DE, USA
**P A Vadas**, USDA Agricultural Research Service, University Park, PA, USA

### Introduction

Phosphorus (P) is essential for growth and development of living organisms. In agricultural systems, P inputs to soils are often required to replenish P removed by harvested crops or lost by soil erosion or runoff. Because soils bind P strongly, historically P inputs have been made more to maintain adequate or high levels of soil fertility and less from a concern that excessive P inputs promote P loss in runoff and erosion. However, all soils have a finite capacity to bind P. When this capacity is diminished through repeated P applications in excess of plant needs, soil P mobility and loss to the environment increase. Even when soil P fertility is relatively low, P loss by excessive erosion can be an environmental concern. The concern for P loss is its delivery to surface waters, where P inputs can increase the growth of algae and aquatic plants to undesirable levels. The result is often eutrophication, whose negative ecological and economic effects include anaerobic conditions during decomposition of algal biomass and subsequent deleterious effects on aquatic organisms, increased cost and complexity of water treatment for drinking quality, and impaired recreational water use. Thus, the challenge for soil P management is to maintain adequate soil P fertility for economical crop production and minimize P loss to the environment.

### The Soil Phosphorus Cycle

#### Major Forms of Phosphorus in Soils

Total soil P ranges from 0.01 to 0.30% and occurs in three forms: P dissolved in the soil solution (solution P), organic P, and inorganic P (**Figure 1**). Solution P exists largely as orthophosphate ions ($HPO_4^{-2}$, $H_2PO_4^{-1}$), and can vary from 0.01 to 3.0 mg l$^{-1}$. Some organic P compounds are also present in solution. Organic P comprises between 30 and 65% of total soil P, depending on soil type, and includes phosphate esters, nucleic acids, phospholipids, and other esters. Inorganic P occurs as primary minerals (those derived directly from weathered parent material) and secondary minerals (those formed by precipitation of P with Al, Ca, and Fe), P adsorbed on to the surface of clay minerals, Fe, and Al oxyhydroxides, or Ca carbonates, and P physically occluded within secondary minerals.

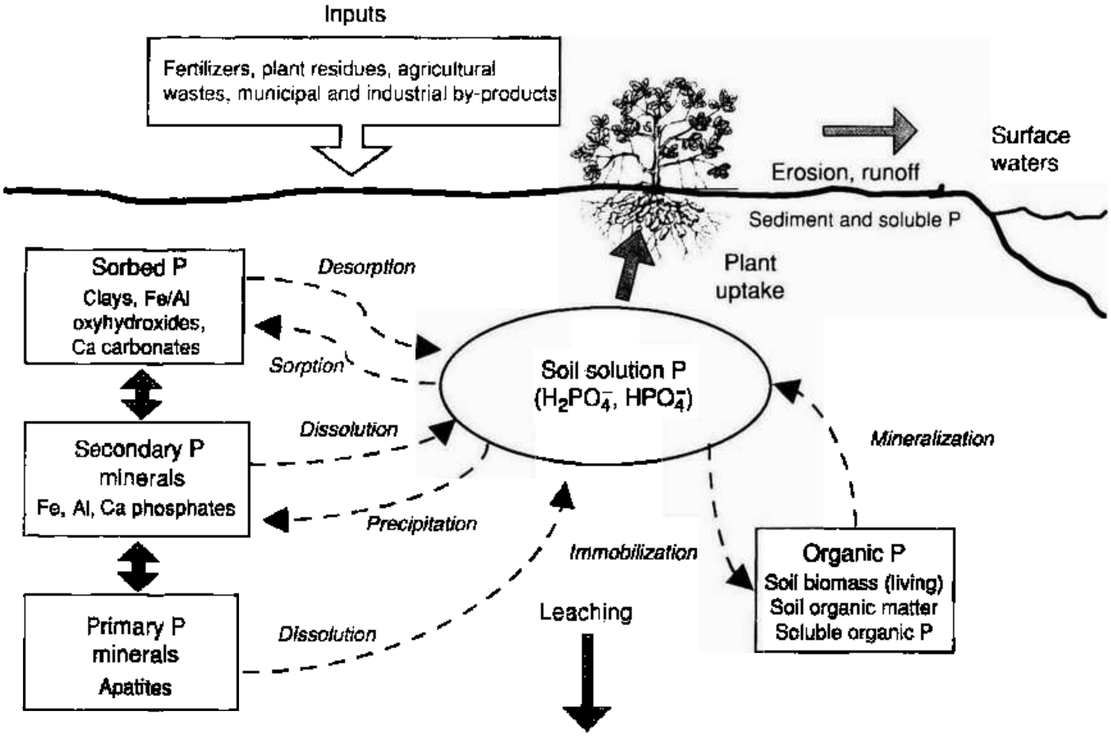


**Figure 1** The soil P cycle. Reproduced with permission from Pierzynski GM (2000) Soils and Environmental Quality. Lewis Publications. CRC imprint.

## Overview of Soil Phosphorus Cycling and Transformations

Solution P concentrations are largely controlled by mineralization and immobilization of organic P and adsorption and/or desorption and precipitation and/or dissolution of inorganic P. Mineralization is the release of organic P to solution, and immobilization is the conversion of solution P to organic P. Soil microbes require P for activity and growth. If organic materials that they oxidize for energy contain less P than required, microbes immobilize P from solution. If organic materials contain more than enough P, microbes mineralize P to solution. Factors that influence P mineralization and immobilization include the quantity of organic P, the C-to-P (C:P) and N-to-P (N:P) ratios of organic materials, temperature, moisture, aeration, pH, cultivation intensity, and P fertilization.

Adsorption is the chemical bonding of P to mineral surfaces, and desorption is release of P from these surfaces. Positively charged surfaces of clays and Fe, Al, and Ca minerals attract orthophosphate ions (e.g., $H_2PO_4^-$, $HPO_4^{-2}$) from solution, which adsorb to the surface by exchanging with $OH^-$ and $OH_2$ groups. When P is bonded through one metal–O–P bond, it is considered labile and can be readily desorbed from the surface to solution. With two metal–O–P bonds, a stable six member ring forms, desorption is more difficult, and P is considered nonlabile.

Precipitation and dissolution reactions are controlled by the solubility of secondary P minerals, the most common of which are Al–P, Fe–P, and Ca–P. A mineral's solubility represents the concentrations of its constituent ions that can be maintained in solution at equilibrium. When the concentration of solution P is greater than that supported by a mineral's chemical solubility, solution P will precipitate with metal cations (e.g., Al, Ca, Fe) and form P minerals. When the solution P concentration is less than a mineral's chemical solubility, P minerals dissolve into solution. Soil physical and chemical properties influencing P adsorption and/or desorption and precipitation and/or dissolution reactions include the nature and amount of soil minerals, soil pH, cation and anion effects, extent of soil P saturation, organic matter, reaction time and temperature, soil oxidation–reduction status, and P fertilization practices.

## Phosphorus Additions to Soils

Because plants absorb P from the soil solution, solution P must be maintained at adequate concentrations, commonly reported as $0.2\,\mathrm{mg\,P\,l^{-1}}$ but which may be as low as $0.03\,\mathrm{mg\,P\,l^{-1}}$. Soils with relatively high concentrations of labile P can provide all P required by plants. Conversely, soils with little labile P require P additions to achieve optimum crop yields. When fertilizer P (inorganic or organic) is added to soil, solution P increases due to dissolution of inorganic P or mineralization of organic P. This increase disturbs the equilibrium between the solid and solution phases, and P moves from solution to the solid phase, mostly through precipitation and adsorption. These initial reactions are rapid and leave the P in labile forms that are easily converted back to solution through dissolution and desorption. Over time, some P in the solid phase stabilizes and becomes nonlabile, so that movement back to solution is more difficult or even impossible. As plants absorb P from the soil solution or soluble P is lost in runoff, labile P dissolves or desorbs from the solid phase to solution.

If P additions consistently exceed plant uptake over time, concentrations of soil P increase. Eventually, labile P concentrations may be great enough to meet plant needs without P fertilization. The greater the concentration of labile P, the longer the soil will be able to maintain plant growth. When such soils are left unfertilized, P fertility levels will progressively decrease, through crop production or P loss, to the point when they no longer adequately supply plant P needs. New additions of P are then necessary.

### Phosphorus Losses from Soils

Phosphorus transfer from soils to aquatic environments is a major nonpoint-source pollution concern. The major pathway for P loss is surface runoff and soil erosion, although P can also be transported by subsurface water flow to nearby streams, especially if artificial drainage is present (e.g., tiles, drainage ditches). As rainfall or irrigation water interacts with the surface soil, P desorbs or dissolves from soil particles, crop residues, fertilizers, or manure. Dissolved P can leave a field in surface runoff or be resorbed by soil or runoff sediment and then move as particulate P. The eroded sediment load in surface runoff determines particulate P loss. In areas where soils are vulnerable to wind erosion, soil P can also be transported with fine, wind-blown sediments and deposited in nearby waters. Years of extensive research show that P loss is minimal when management practices prevent surface accumulation of P and minimize surface runoff and erosion.

Because soil P is relatively immobile, most fertilizer P remains near the soil surface. However, leaching of P from the surface soil into the subsoil may occur in sandy soils and organic soils, and may be greater in soils where organic fertilizer is added than where inorganic fertilizer is added (**Figure 2**). Leaching usually occurs in soils where P additions have increased soil P to excessive concentrations. In such soils, P is more mobile because sorption sites have become increasingly saturated and solution P concentrations have increased. Leaching of P can be exacerbated by preferential water flow where soil water containing relatively high concentrations of P from surface layers flows rapidly through soils in cracks, fissures, biopores, or finger flow. Phosphorus that has leached from surface soils may be retained in subsoils, gradually leach further into groundwaters, or move by subsurface flow to nearby streams.

## Soil Phosphorus Management

### Soil Testing for Phosphorus

Quantifying the ability of a soil to provide plant-available P is usually the first step in P fertility management. Beginning in the 1940s, agronomic soil tests using chemical extractants were developed to estimate rapidly and economically the ability of a soil to supply P to plants over a growing season. Use of the four most common tests, Bray and Kurtz P-1, Mehlich 1, Mehlich 3, and Olsen P, is a well-established practice in the USA. The Bray and Kurtz P-1 test is an acid extractant used in the midwestern and north central USA for acid and neutral soils. The Mehlich 1 and Mehlich 3 tests are acid extractants, with Mehlich 1 used in southeastern and mid-Atlantic states for acid soils and Mehlich 3 used throughout the USA and Canada for both acid and basic soils. The Olsen P test is a basic extractant used in the north central and western USA for calcareous soils. Some other soil P tests used in the USA and Europe include ammonium bicarbonate-DTPA, Morgan and modified Morgan, ammonium lactate–acetic acid, the $P_i$ soil test, and water and dilute salt tests. Soil testing is a reliable method to determine soil P fertility levels and fertilizer P requirements, and can identify soils where P has accumulated to environmentally unacceptable levels.

### Characteristics of Phosphorus Fertilizer Sources

An effective P fertilizer must contain an adequate amount of available P so P will be released quickly and sufficiently enough to replenish soil P reserves (**Table 1**). Mined rock phosphate minerals, or apatites, are the only important raw material for inorganic P fertilizers. Although they contain 12–18% total P, very little of the P is water-soluble, which limits their fertilizer value in certain settings, such as highly acid soils. More common P fertilizers are produced when rock phosphate is acid-treated or heat-treated to break apatite bonds and increase the water-soluble P content. Phosphorus can also be applied to soils in organic fertilizers, including animal wastes, municipal wastes, composts, plant residues, and occasionally industrial by-products (**Table 1**). Organic fertilizers generally contain relatively low concentrations of P, ranging from 0.1 to 3.0%.

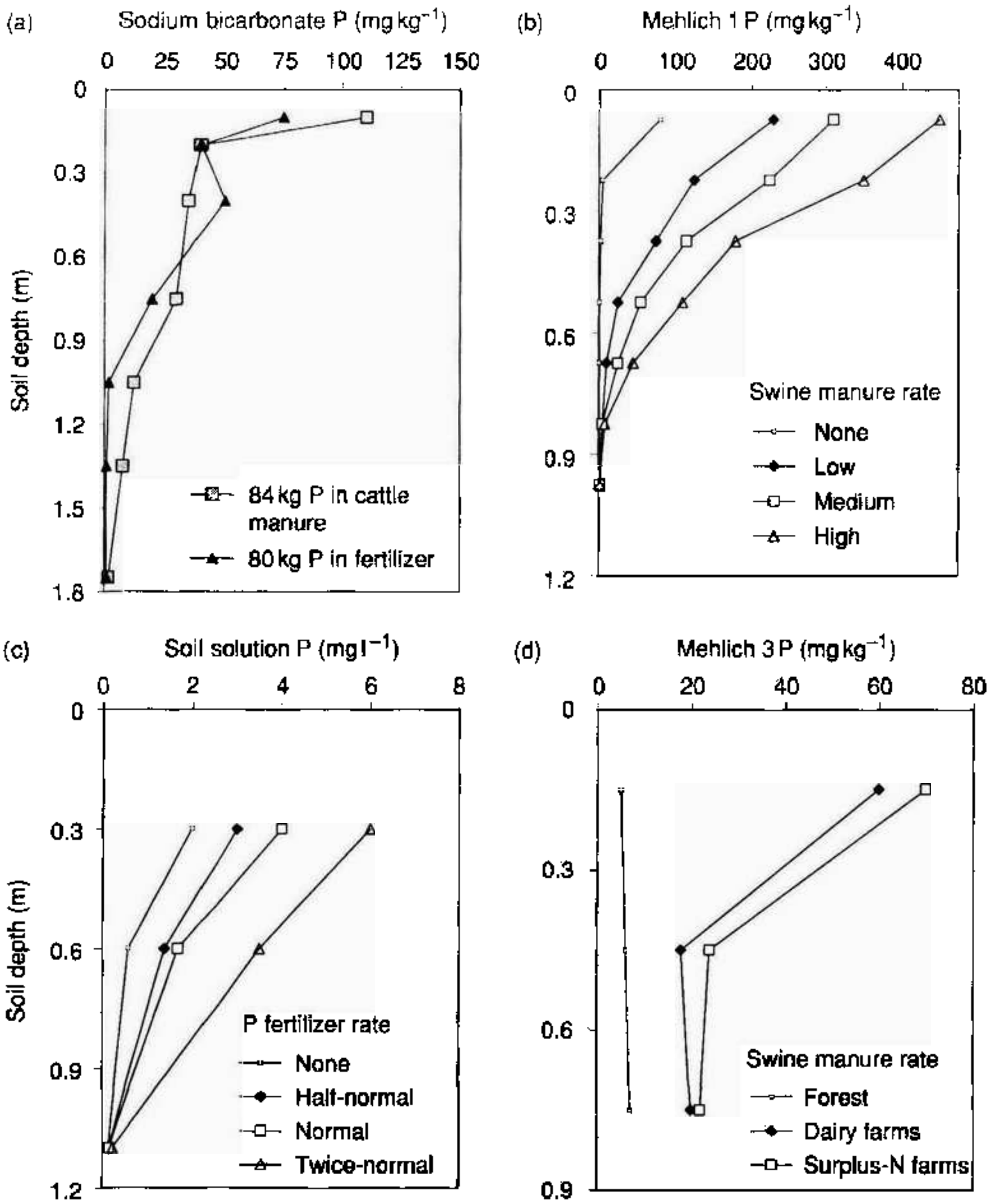


**Figure 2** Examples of P-leaching for (a) soils where organic fertilizers and inorganic fertilizers are applied, (b) sandy soils, (c) organic soils, and (d) soils that have excessive concentrations of P. (Source: Eghball B, Binford GD, and Baltensperger DD (1996) Phosphorus movement and adsorption in a soil receiving long-term manure and fertilizer application. *Journal of Environmental Quality* 25 1339–1343; Hortenstine CC and Forbes RB (1972) Concentrations of nitrogen, phosphorus, potassium, and total soluble salts in soil solution samples from fertilized and unfertilized Histosols. *Journal of Environmental Quality* 1: 446–449; King LD, Burns JC, and Westerman PW (1990) Long-term swine lagoon effluent applications on "coastal" Bermuda grass. II. Effect on nutrient accumulation in soil. *Journal of Environmental Quality* 19: 756–760; Simard RR, Cluis D, Gangbazo G, and Beauchemin S (1995) Phosphorus status of forest and agricultural soils from a watershed of high animal density. *Journal of Environmental Quality* 24: 1010–1017.)

## Nutrient Management Practices for Phosphorus

The fundamental goal of agricultural P management is to identify the most profitable and environmentally sound approach to provide crops with P, to prevent P accumulation in soil, and minimize P loss from soil. Effective P management begins with a nutrient budget that calculates P inputs to and P outputs from a farm. When P inputs exceed outputs over time, the usual result is an accumulation of soil P and an increased risk of P loss to water. If P outputs exceed inputs, soil P fertility declines. A positive imbalance (P surplus) is economically undesirable and represents a waste of a limited natural resource in phosphate ores. A P deficit indicates that soil productivity is being slowly degraded and crop production will probably decline. Therefore, balancing farm-scale P inputs and outputs is critical to long-term P fertility management.

On farms where commercial fertilizers are the only P inputs, soil tests and realistic yield goals will assess soil P fertility, crop P needs, and P fertilizer requirements. Analyzing harvested crops will quantify P leaving the farm. In this situation, a P surplus indicates unnecessary fertilization and a P deficit suggests farmers face economic or logistical limitations to maintaining soil P fertility. In animal-based operations where P inputs also include animal feed, soil testing and yield goals will still identify soil P fertility, crop P needs, and P fertilizer requirements; and a manure analysis will identify available P that can be used as fertilizer. Manure application rates should be based on crop nutrient requirements and manure nutrient concentrations. Crop and animal product nutrient analyses or estimations will quantify farm P outputs. Because animal-based agriculture can produce large quantities of manure and often relies exclusively on land application as the end use for manure, it is especially important that P inputs do not exceed P outputs to prevent the buildup of soil P to concentrations of environmental concern (Table 2). If P inputs significantly exceed P outputs, then management options that reduce P inputs, use on-farm P resources without promoting P losses, and increase P outputs can be investigated. Examples are reducing unnecessary P additions to animal feeds, increasing availability of P in feed with dietary amendments such as phytase, or exporting manures for alternative uses such as compost, fertilizer products, or bioenergy production.

**Table 1** Major inorganic and organic sources of phosphorus for crop production

| *Material* | *$P_2O_5$ (%)* | *Total P available (%)* |
|---|---|---|
| *Inorganic fertilizers* | | |
| Rock phosphate | 25–40 | 14–65 |
| Single superphosphate | 16–22 | 97–100 |
| Wet-process phosphoric acid | 48–53 | 100 |
| Triple superphosphate | 44–53 | 97–100 |
| Ammonium phosphates | | |
| Monoammonium phosphate | 48–62 | 100 |
| Diammonium phosphate | 46–53 | 100 |
| Ammonium polyphosphate | 35–62 | 100 |
| Urea–ammonium phosphate | 16–42 | 100 |
| Nitric phosphates | 14–28 | 80–100 |
| Ammoniated normal superphosphate | 14–21 | 97–100 |
| Ammoniated triple superphosphate | 44–53 | 96–100 |
| Potassium phosphates | | |
| Monopotassium phosphate | 51 | 100 |
| Dipotassium phosphate | 41 | 100 |
| Potassium polyphosphate | 51 | 100 |
| *Organic fertilizers* | | |
| Beef manure | 2.1 | – |
| Dairy manure | 1.4 | – |
| Poultry manure | 4.1 | – |
| Swine manure | 3.5 | – |
| Aerobically digested sludge | 7.6 | – |
| Anaerobically digested sludge | 8.3 | – |
| Composted sludge | 3.0 | – |

Adapted from Pierzynski GM, Sims JT, and Vance GF (1994) *Soils and Environmental Quality*. Boca Raton, FL: Lewis Publishers; Tisdale SL, Nelson WL, Beaton JD, and Havlin JL (1993) *Soil Fertility and Fertilizers*. New York, NY: Macmillan.

Crop response to P additions in either fertilizers or manures is influenced by the rate, timing, and method of P application. At low rates of P fertilization, the water solubility of fertilizer P should be high to get the full benefit of the limited amount applied. In soils that sorb P strongly and quickly, higher rates of P fertilizer may be necessary to achieve the solution P concentration necessary for crop growth. In such soils, it may be more efficient to have smaller, more frequent P additions closer to the times of high crop P demand. If P fertilizers are broadcast and incorporated, much of the added P interacts with soil particles capable of P sorption. If soils are low in P or have high P sorption capacity, high rates of fertilizer P must be broadcast to increase P availability to desired values. For row crops, placing fertilizer P in concentrated bands immediately adjacent to the crop reduces the contact between soil and fertilizer P, allowing for use of lower fertilizer rates and enhanced P uptake by crops.

Loss of P from soils is also influenced by the rate, timing, and method of fertilizer or manure application. The greatest P loss occurs from fields where high rates of P are applied, where P is broadcast and

**Table 2** Phosphorus budget for a typical poultry farm in Delaware, USA. and a typical dairy farm in New York, USA

| | *P input* | | | | *P output* | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Chicks (kg)* | *Feed (kg)* | *Fertilizer (kg)* | *Total $(kg\ ha^{-1})$* | *Birds (kg)* | *Crops (kg)* | *Total $(kg\ ha^{-1})$* | *P balance $(kg\ ha^{-1})$* | *P input in surplus (%)* |
| Poultry farm | 90 | 9300 | 0 | | 2500 | 1300 | | 74 | 58 |
| Dairy farm | | | | 14.5 | | | 8.4 | 6.1 | 42 |

Source: Sims JT and Vadas PA (1997) *Nutrient Management Planning for Poultry-Grain Agriculture*. Fact Sheet ST-11 College of Agricultural Sciences and Cooperative Extension. Newark, DE: University of Delaware.
Source: Rotz CA, Sharpley AN, Satter LD, Gburek WJ, and Sanderson MA (2002) Production and feeding strategies for phosphorus management on dairy farms. *Journal of Dairy Science* 85: 3142–3153.

unincorporated, and when runoff events occur soon after P application. The least P will be lost from fields where little P is applied, where P is incorporated into soil, where good soil conservation practices reduce erosion, and when runoff events occur well after application. Because these optimum conditions reduce the number of manure application opportunities available throughout the year, manure should be stored in suitable facilities until it can be applied.

### Soil and Water Conservation Practices for Phosphorus

An integral component of sound P management is practices that conserve soil and manage water, thus reducing the risk of P loss. To minimize P loss via surface runoff and erosion, it is necessary to reduce rainfall impact on the soil, reduce surface runoff volume and velocity, and increase soil resistance to erosion.

### Plant Cover, Crop Rotation, and Strip Cropping

Rainfall breaks up soil aggregates, allowing soil to move with runoff, and compacts and puddles the soil surface, decreasing water infiltration and increasing surface runoff. Crops can intercept up to 45% of rainfall, dissipating raindrop energy and reducing runoff and erosion. After planting, rapid crop canopy helps minimize soil exposure. Adequate soil fertility, optimum seedbed preparation and planting conditions, pest control, and optimum plant spacing all help promote seed germination, plant growth, and canopy closure. Crop rotations where new crops quickly follow preceding harvests minimize the duration of soil exposure. Rotations also help break insect, disease, or weed cycles, which impede plant growth and reduce plant cover. Strip cropping divides fields into long, narrow segments that help control runoff and erosion while growing crops in rotation. More heavily vegetated strips slow runoff and catch soil eroded from more exposed strips. Strips planted on the contour of slopes (contour strip-cropping) are especially effective in controlling runoff and erosion (Figure 3).

### Companion Crops, Cover Crops, Green Manures, and Crop Residues

With companion crops such as small-grain and forage crops, the small grain starts first and provides quick plant cover. Harvesting the grain leaves residue and the forage to protect the soil. Cover crops grown during cold or dry seasons can provide soil cover when soils would otherwise be left exposed. Cover crops should germinate easily, grow quickly, provide sufficient plant cover, and be hardy against weather and pests to maximize soil coverage. Green manures act as cover crops and, when plowed under, can add nutrients to soil to increase fertility for succeeding crop growth and build soil organic matter content, which promotes the soil structure, permeability, and aeration needed to maximize rain infiltration and minimize surface runoff and erosion. Crop residues also act as mulch to increase soil coverage. Although plowing residues under may decrease soil coverage, it can promote residue decomposition and erosion-control benefits such as those from green manures.

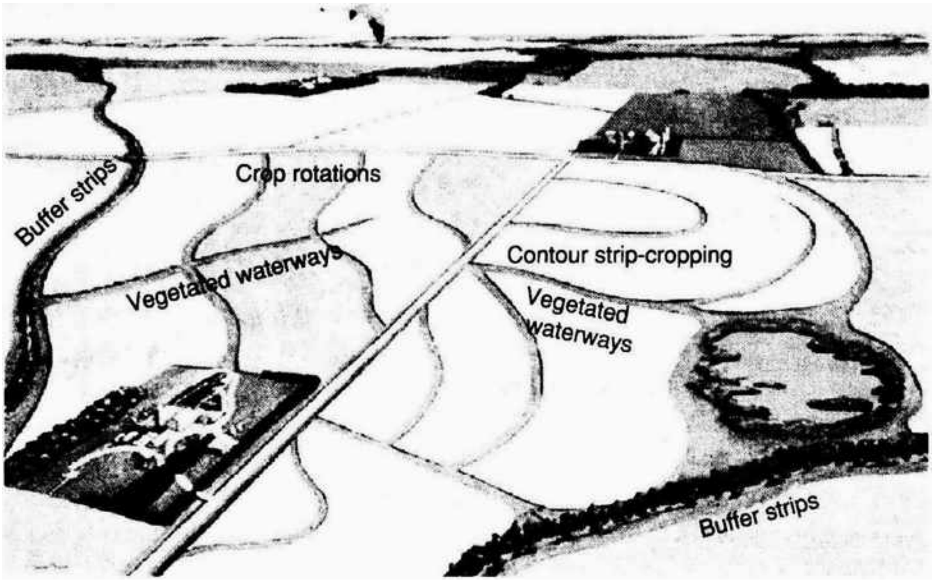


**Figure 3** Soil and water conservation practices that can help reduce P export from agricultural fields. Courtesy of USDA Natural Resources Conservation Service.

### Conservation Tillage, Contour Cultivation, and Deep Plowing

Conservation tillage, such as stubble mulch tillage, minimum tillage, reduced tillage, and no-tillage, leaves at least 30% of the soil surface covered by plant residue after planting to reduce erosion and surface runoff. Contour cultivation follows the contour lines of sloping fields. Contour ridges produced by tillage and planting form barriers that slow downhill water movement and decrease surface runoff and erosion. In applicable situations, deep plowing can break up pans to increase water infiltration and reduce runoff and erosion. It can also help dilute topsoil layers with high P concentrations into subsoils with less P.

Although reducing tillage will help decrease total P loss in surface runoff by reducing erosion, it may increase dissolved P loss. Because soils are not tilled, P from added fertilizers or manures accumulates at the surface, increasing the concentration of dissolved P in surface runoff. Also, in no-till situations where manures and fertilizers are left unincorporated, dissolved P loss in surface runoff soon after application can increase dramatically.

### Terraces, Vegetated Waterways, and Buffer Strips

Terraces divide slopes so that surface runoff is intercepted and carried to a protective outlet. Terraces decrease erosion by shortening slope length, slowing runoff velocity, and trapping sediments. Surface runoff water concentrates in waterways that when bare or unstable are extremely erodible. Vegetated waterways are natural or constructed channels that, when properly established and maintained, transport surface runoff water at a nonerosive velocity from fields, prevent gully formation, and greatly decrease erosion. Vegetated buffer strips act primarily to control water pollution by preventing eroded soil from being carried into streams or ditches.

## Innovations in Soil Phosphorus Management

The past two decades have seen an increased awareness of the importance of P loss from agricultural soils. Extensive research has been conducted to better understand sources and pathways of P loss, to identify areas or fields where risk of P loss is high, to develop management practices to minimize P loss, and determine the best means by which to implement such practices. A number of innovations in soil P management have evolved from this research.

### Animal Feed Formulations

Inorganic P supplements are routinely added to animal feeds to ensure adequate dietary P. Because P is often added in excess of animal needs, significant amounts of P are excreted in manure and applied to soils, where they are a potential source of P loss. Adjusting diets to match P inputs with animal needs, or increasing the availability of native feed P, as with phytase additions to poultry feeds, can potentially reduce the amount of P that is excreted in manure and land applied. This would also help reduce P surpluses that exist in farm-scale nutrient balances (Table 2). However, research must determine whether dietary manipulations will indeed reduce P loss from soils where manure is applied. For example, adding phytase to poultry feeds to increase P availability to birds may increase the solubility of P in manure and thus increase dissolved P loss in surface runoff from fields where manure has been applied.

### Soil and Manure Testing

Soils are important sources of P loss. As soil P concentrations increase due to long-term applications of fertilizer and manures, tests are needed to identify soils that pose a significant environmental risk. Good relationships exist between soil P concentrations as measured by traditional agronomic tests and dissolved P loss in surface runoff or subsurface flow. Research has also shown that some new soil tests, such as extractions with water, iron oxide-coated filter paper, and ion-exchange membranes, and a test for the degree of soil P saturation, can also estimate the risk of P loss from soil. Land-applied manures can too be significant sources of P loss. Simple manure tests, such as extractions with water, can estimate the potential for manure P loss. Although soil and manure tests can identify potential sources of P loss, they give no information about the transport of P. For example, manure applications to a given field increase soil P concentrations and thus the potential for P loss in runoff or erosion. However, if runoff from the field is rare, then transport of P from the field is unlikely. Tools such as the Phosphorus Index can help identify fields where the risk of P pollution from both source and transport factors is high.

### Soil and Manure Amendments

The loss of P in runoff is related to the P concentration of both soils and manures. Amending soils, manures, or municipal biosolids with materials containing high amounts of Fe, Al, or Ca can transform the soil or manure P into less-available forms and thus reduce P loss in runoff. Such amendment materials include an Fe-rich by-product from paint production, Al-rich drinking water treatment residuals, aluminum sulfate (alum), and coal combustion by-products. While these amendments are effective in reducing P availability in

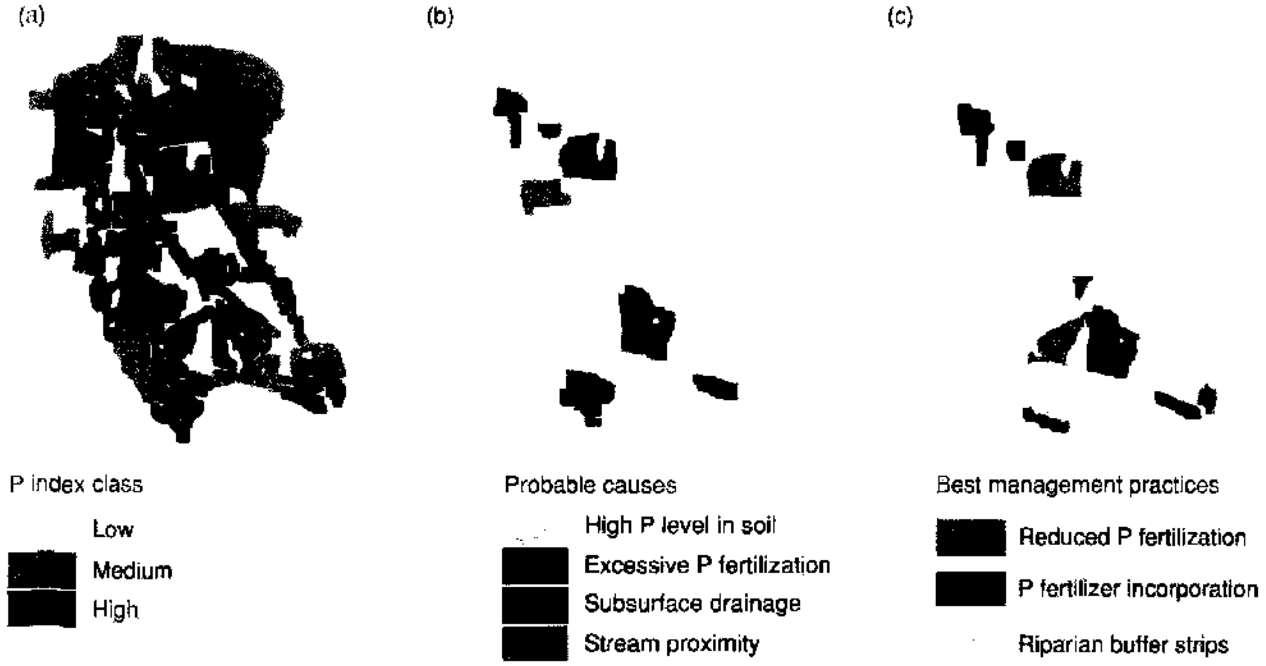


**Figure 4** Examples of the P index for a watershed scale, showing (a) P-index ratings, (b) causes of P export, and (c) management practices that can help reduce the P export. (Source: Djodjic F, Montas H, Shirmohammadi A, Bergström L, and Ulén B (2002) A decision support system for phosphorus management at a watershed scale. *Journal of Environmental Quality* 31: 937–945.)

soils and manures, they may also contain high concentrations of salts or heavy metals that could have adverse agronomic or environmental effects.

### Water-Quality Modeling and the Phosphorus Index

Water-quality simulation models such as EPIC, FHANTM, GLEAMS, AGNPS, SWAT, or ANSWERS are designed to predict P loss from agricultural areas for a variety of soil, hydrologic, or management conditions and are seen as relatively fast, cost-effective tools. Continued research with these models has led to their improvement to simulate P loss better. However, such models are often complex and require large amounts of input data that may be unknown or difficult to measure. They may also be limited in the number of soil, hydrologic, or management conditions they simulate. The Phosphorus Index (P Index) is a much simpler screening tool, developed by the Natural Resources Conservation Service (NRCS) and recently modified; it accounts for source and transport factors controlling P loss and ranks sites according to the risk of P movement (**Figure 4**). The P Index is designed to allow field personnel easily to identify agricultural areas or practices that are most vulnerable to P loss, thus allowing producers more flexibility in developing P-control strategies.

## Further Reading

Abu-Zreig M, Rudra RP, Whitely HR, Lalonde MN, and Kaushik NK (2003) Phosphorus removal in vegetated filter strips. *Journal of Environmental Quality* 32: 613–619.

Akhtar MS, Richards BK, Medrano PA, deGroot M, and Steenhuis TS (2003) Dissolved phosphorus from undisturbed soil cores: related to adsorption strength, flow rate, or soil structure? *Soil Science Society of America Journal* 67: 458–470.

Andraski BJ, Mueller DH, and Daniel TC (1985) Phosphorus losses in runoff as affected by tillage. *Soil Science Society of America Journal* 49: 1523–1527.

Frossard E, Condron LM, Oberson A, Sinaj S, and Fardeau JC (2000) Processes governing phosphorus availability in temperate soils. *Journal of Environmental Quality* 29: 15–23.

Gburek WJ, Sharpley AN, Heathwaite AL, and Folmar GJ (2000) Phosphorus management at the watershed scale: a modification of the phosphorus index. *Journal of Environmental Quality* 29: 130–144.

Harrison AF (1987) *Soil Organic Phosphorus: A Review of World Literature*. Wallingford, UK: CAB International.

Heckrath G, Brookes PC, Poulton PR, and Goulding KWT (1995) Phosphorus leaching from soils containing different phosphorus concentrations in the Broadbalk experiment. *Journal of Environmental Quality* 24: 904–910.

Kleinman PJA, Sharpley AN, Wolf AM, Beegle DB, and Moore PA Jr (2002) Measuring water-extractable phosphorous in manure as an indicator of phosphorus in runoff. *Soil Science Society of America Journal* 66: 2009–2015.

Moore PA Jr, Daniel TC, and Edwards DR (2000) Reducing phosphorus runoff and inhibiting ammonia loss from poultry manure with aluminum sulfate. *Journal of Environmental Quality* 29: 37–49.

Pote DH, Daniel TC, Nichols DJ, Sharpley AN, Moore PA Jr, Miller DM, and Edwards DR (1999) Relationship between phosphorus levels in three ultisols and phosphorus concentrations in runoff. *Journal of Environmental Quality* 28: 170–175.

Sharpley AN, Chapra SC, Wedepohl R, Sims JT, Daniel TC, and Reddy KR (1994) Managing agricultural phosphorus for protection of surface waters: issues and options. *Journal of Environmental Quality* 24: 437–451.

Sharpley AN, Kleinman PJA, McDowell RW, Gitau M, and Bryant RB (2002) Modeling phosphorus transport in agricultural watersheds: processes and possibilities. *Journal of Soil and Water Conservation* 57: 425–439.

Sims JT, Edwards AC, Schoumans OF, and Simard RR (2000) Integrating soil phosphorus testing into environmentally based agricultural management practices. *Journal of Environmental Quality* 29: 60–71.

Vadas PA, Kleinman PJA, and Sharpley AN (2004) A simple method to predict dissolved phosphorus in runoff from surface applied manures. *Journal of Environmental Quality* (in press).

# Biological Interactions

**M D Mullen**, University of Kentucky, Lexington, KY, USA

## Introduction

Phosphorus (P) is essential to life; it is a primary component of many important biomolecules such as adenosine triphosphate (ATP), deoxyribonucleic acid (DNA), phospholipids, the phosphopyridine nucleotides ($NADP^+$), and phytin (P storage molecule – inositol hexaphosphate). Plant requirements for P are, on average, between 0.2 and 0.4% of plant dry weight. While this is relatively modest relative to nitrogen (N) requirements, the challenge with P lies in maintaining enough soluble phosphate to prevent deficiencies of P in agricultural crops. Therefore the management of P in agriculture is of major interest.

Phosphorus enters the soil environment in a number of ways. It is present, in varying concentrations, as the primary mineral apatite from which the plant-available phosphate ion is weathered. Phosphorus is applied as inorganic fertilizers or organic materials such as animal manures and other biosolids, or is simply recycled from plant and animal materials in undisturbed soil ecosystems.

The P cycle in soils has both abiotic and biological components. The abiotic component is primarily concerned with the dissolution of P-containing minerals, the formation of phosphates of Al, Fe, Ca, and Mg, and the sorption of phosphates to clays and Al- and Fe-oxides in soil. The biological component is comprised primarily of mineralization of organic P sources and immobilization of soil solution orthophosphate during microbial metabolism. Soil organisms also produce a variety of acids and chelating agents which facilitate the solubilization of phosphate from the inorganic P in the soil. The organisms involved include bacteria and fungi living in the bulk soil and the rhizosphere, as well as mycorrhizal fungi which supply P to the plant. The biological transformations of P in the soil are very important in the availability of P to plants. However, mismanagement or overapplication of P-containing materials can result in excessive amounts of inorganic P, which can have negative environmental impacts. The introduction of excessive P into surface waters can result in eutrophication, or excessive algal productivity, which can ultimately reduce oxygen content in the water, degrading the quality of the habitat for aquatic organisms.

## Forms of Soil Phosphorus

The original source of most soil P is from the primary mineral apatite, or rock phosphate (RP). Apatite is a highly insoluble Ca-phosphate compound with the general formula of $Ca_{10}(PO_4)_6X_2$, where X is $F^-$ (fluoroapatite), $OH^-$ (hydroxyapatite), or $Cl^-$ (chloroapatite). In highly weathered, acidic soils, little or none of these apatite minerals remain in the soil. The P in these minerals is released, either chemically or by microorganisms and plant roots, then used by plants and microorganisms. Over time, this P is incorporated into organic matter or other relatively insoluble inorganic P minerals with Al and Fe in acidic soils, and with Ca in neutral to alkaline soils. In highly weathered soils, P is tied up by Al- and Fe-oxides. These compounds are only sparingly soluble (**Figure 1**). Generally, solubility of phosphorus from Al and Fe sources increases with pH levels in the range of about pH 6.5. As pH continues to increase, though, Ca-phosphate minerals are formed and these control soluble P concentrations. Therefore as pH increases above neutrality, soluble P concentrations decrease significantly. The low solubility of inorganic soil P makes the use of P fertilizer materials necessary in many high-yield agricultural systems. Additionally, the practice of liming acid soils to increase pH to about 6.5–7.0 has the practical effect of enhancing P availability for plant growth.

Total amounts of P in soils vary widely, depending on parent material and the extent of weathering. In sandy soils with low organic matter, there may be as little as 100 mg kg$^{-1}$, while soils high in primary

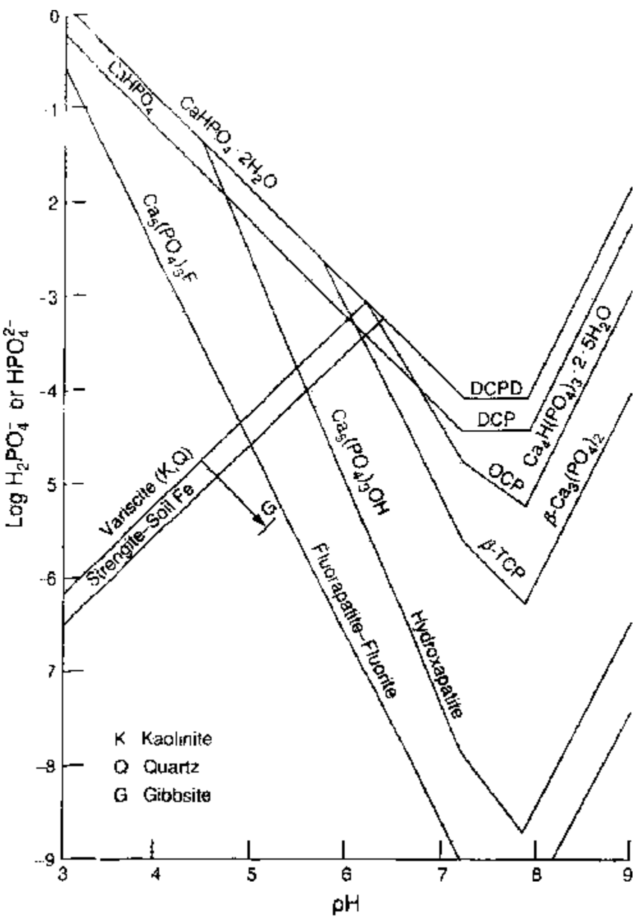
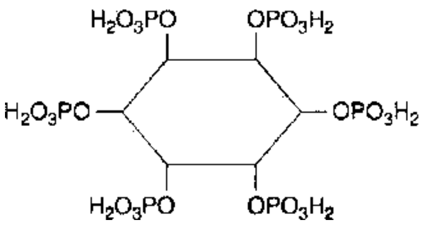


**Figure 1** Solubility of orthophosphate from variscite ($AlPO_4 \cdot 2H_2O$), strengite ($FePO_4 \cdot 2H_2O$), and various Ca-phosphates as a function of solution pH. These equilibria assume a $Ca^{2+}$ concentration of $10^{-2.5}$ mol l$^{-1}$ and $CO_2$ at 0.03 kpa. DCPD, calcium phosphate dehydrate, brushite; DCP, calcium phosphate, monetite; OCP, octacalcium phospate; $\beta$-TCP, $\beta$-tricalcium phosphate. Reproduced with permission from Lindsay WL (1979) *Chemical Equilibria in Soils*. New York: John Wiley.

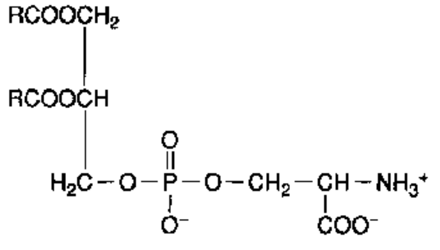
P minerals (e.g., high-phosphate limestone parent materials), clay content, or organic matter may contain well over 2000 mg kg$^{-1}$ soil. In most soils, inorganic P is the dominant form, although organic P increases with increasing soil organic matter.

The plant-available form of P in soils is orthophosphate, $H_2PO_4^-$ (dominant form at pH 2.2–7.2) or $HPO_4^{2-}$ (dominant form at pH 7.2–12.4). Orthophosphate from soluble phosphate fertilizer materials or mineralized from organic sources rapidly forms precipitates as Al-, Fe-, or Ca-phosphates or is sorbed by Al- and Fe-oxides. Due to these reactions, soluble P is present at very low concentrations in soil solution, typically in the range of less than 0.1 to 1 mg l$^{-1}$. In nonfertilized soils the solution orthophosphate must be continuously replenished by chemical or biological dissolution of inorganic P compounds or by microbial mineralization of organic P.

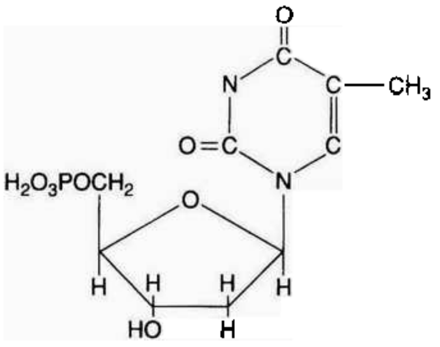
Organic P can account for 3–90% of the total soil P, with 30–50% common in most soils. The P content of soil organic matter ranges from about 1 to over 3%. Total organic P in a soil is a function of organic matter content, so that as organic matter increases, so does organic P. Organic P also decreases with soil depth, as does organic C. The chemical structure of organic P in soil is not well known. The major types of organic P compounds which are added to soil by recycling plant, animal, and microbial remains in the soil environment are shown in **Figure 2**. Of the compounds shown, phytins, or inositol phosphates, are usually found in the greatest quantities, making up 10 to 50% of the total organic P content. The inositol phosphates are typically in a polymeric state, making these compounds relatively resistant to decomposition. The phospholipids and nucleic acids may account for up to 5% of the organic P fraction, while other identifiable forms of organic P are present in trace amounts only.



Inositol hexaphosphate



Phospholipid: phosphatidyl serine



Nucleotide: thymidine 5′-phosphate

**Figure 2** Representative forms of organic phosphorus inputs into the soil environment. Reproduced with permission from Mullen MD (1998) Transformations of other elements. In: Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (eds) *Principles and Applications of Soil Microbiology*, pp. 369–386. Upper Saddle River, NJ: Prentice-Hall.

The microbial biomass represents an important part of the organic P pool in soil. This pool represents actively cycling P in the soil, and is part of the labile or readily available organic P. The microbial biomass is responsible for mineralization and immobilization reactions which convert organic P to inorganic P and inorganic P to organically bound P. The microbial biomass is also active in solubilization of relatively insoluble forms of Al-, Fe-, and Ca-phosphates. The biomass P fraction accounts for up to 5% of the total organic P pool in arable soils, while in undisturbed grassland and forest soils, the fraction may be as high as 20% of the total. Studies have shown that soil disturbance, e.g., tillage, reduces organic matter and total organic P content, and that the microbial biomass P fraction decreases more rapidly than the total organic P pool in general. Therefore, the labile organic P fraction is more susceptible to disturbance than the more stable, complex organic P forms. This indicates the importance of the microbial biomass to making P available for plant use in stable terrestrial ecosystems.

## The Phosphorus Cycle

In some ways, the phosphorus cycle is not as complex as the nitrogen or sulfur cycles because P does not typically undergo oxidation–reduction reactions or exist in gaseous forms. However, as in the N and S cycles, the biological processes of mineralization and immobilization as well as plant uptake are present. Additionally, the microbial fraction is active in the solubilization of relatively insoluble inorganic P materials, making P available for plant or microbial uptake or other transformations.

**Figure 3** shows the P cycle, illustrating the various compartments of P in the soil environment. The P cycle is divided into a geochemical subcycle and a biological subcycle, with the inorganic solution P pool serving as the central point in the overall cycle. This pool of inorganic solution P serves as the source of orthophosphate for plants and soil microorganisms. The remainder of this article will deal with the biological subcycle of P.

In the biological subcycle, soluble orthophosphate is either taken up by plants or immobilized into the microbial biomass. When organic sources of P, such as crop residues, manures, and municipal biosolids, are land-applied, three possible fates of the P can occur. The organic P, especially inositol phosphates, may be incorporated directly into stable humus, becoming relatively unavailable for plant and microbial use. If the residue has a relatively high P content, relative to C and N, a portion of the P will

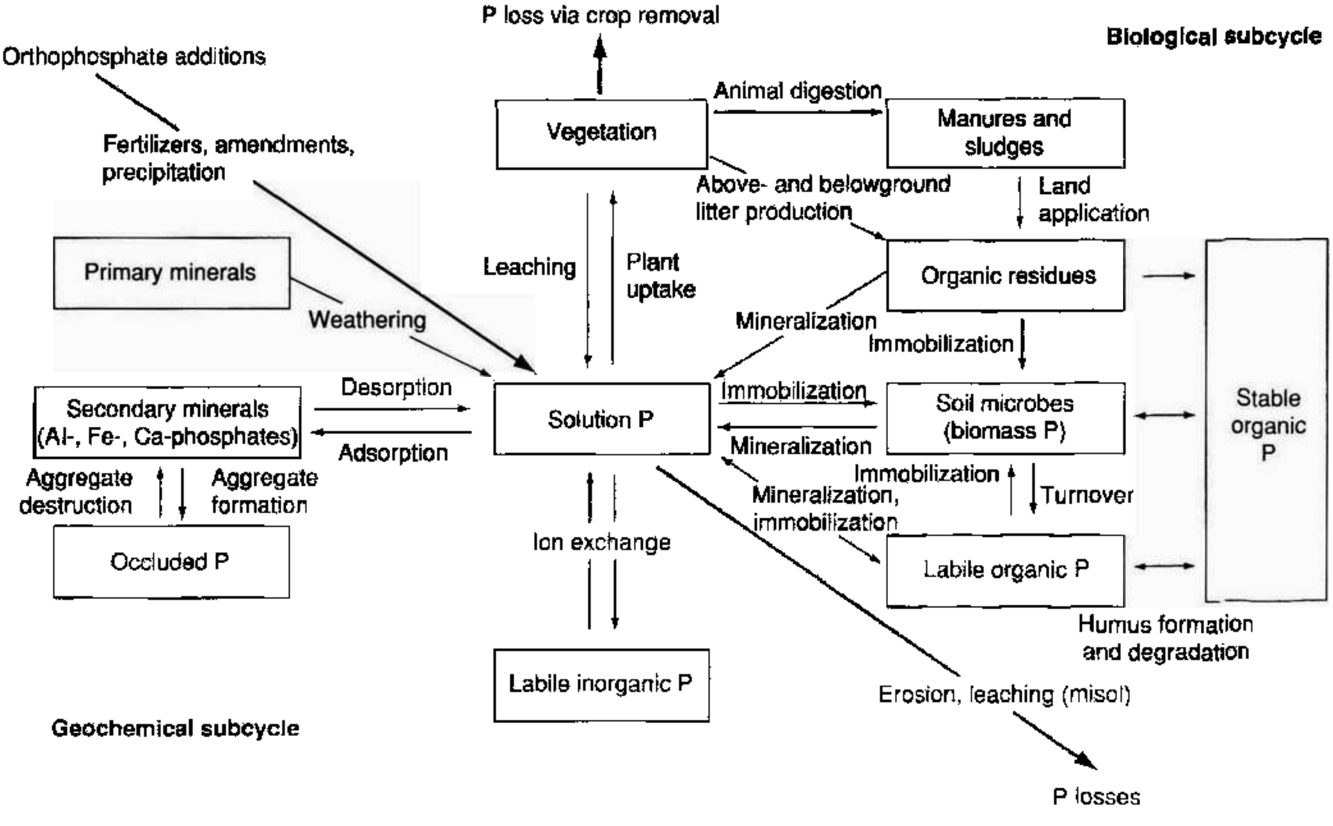


**Figure 3** The phosphorus cycle showing inputs, losses, and transformations. This model illustrates the relationship of the chemical and biological P subcycles through the solution orthophosphate pool. Adapted with permission from Walbridge MR (1991) Phosphorus availability in acid organic soils of the lower North Carolina coastal plain. *Ecology* 72: 2083–2100.

be mineralized as orthophosphate during microbial decomposition of the residue. Finally, part of the P in the residue will be incorporated into the microbial biomass during decomposition. This immobilized P may become available as soluble P when the microbial biomass fraction dies back and is itself mineralized by other microorganisms.

While this model of the P cycle segregates biological activity from geochemical activity, it should be noted that the dissolution of apatite and other secondary P minerals can be enhanced by the activities of microorganisms and plants. Dissolution or solubilization of these minerals releases orthophosphate to solution P, where it can then be utilized by plants or microorganisms. Of course, a portion of this P will again precipitate as secondary P minerals, or be adsorbed by Al- and Fe-oxides and clay surfaces.

## Mineralization and Immobilization of P

As briefly discussed above, key components of the biological cycling of P in soil are mineralization, the transformation of organic P into inorganic orthophosphate P, and immobilization, the transformation of inorganic P back into organic forms. These reactions are catalyzed by the active microbial biomass fraction in the soil, and are largely a consequence of heterotrophic microbial catabolic and anabolic processes. In order to acquire energy, electron-rich organic compounds are oxidized and degraded in order to produce energy. As these compounds are dismantled, inorganic nutrients are released. In order to build new cells, the biomass must acquire inorganic P and other nutrients. The extent to which P is either mineralized or immobilized during these reactions depends on the carbon-to-phosphorus (C:P) ratio of the compound being degraded. If the C:P ratio is 200:1 or less, then the residue is rich in P and the excess will be mineralized. Conversely, if the residue C:P is greater than about 300:1 there will not be enough P available in the residue to facilitate complete degradation. Therefore immobilization of orthophosphate from the soil by the microbial biomass will occur if decomposition of the residue and biomass growth is to continue. While immobilization decreases the amount of orthophosphate immediately available to plants, this is typically a very short-lived phenomenon and may indeed protect orthophosphate from being tied up in relatively insoluble inorganic compounds. After the added organic materials are mineralized, the microbial biomass will begin to die back to previous levels, resulting in the subsequent mineralization of the microbial P pool.

Mineralization is simply the enzymatic release of orthophosphate during the decomposition of organic materials that contain P. The phosphatase enzyme group is responsible for the hydrolysis of esters and anhydrides of orthophosphate-P. A number of different phosphatase enzymes are observed in the soil environment. The most commonly measured soil phosphatases are the phosphomonoesterases, which hydrolyze phosphomonoester bonds. Two phosphomonoesterases have been widely studied: acid phosphatase and alkaline phosphatase. Acid phosphatase is produced by microorganisms and plants and predominates in acid soils, while alkaline phosphatase is only produced by soil microorganisms and is dominant in alkaline soils (although it is present in acid soils as well). The phytase enzyme is also important. Phytins or inositol phosphates (**Figure 1**) can make up as much as 50% of the organic P in a soil. Therefore, hydrolysis of orthophosphate from this source can be important to plant nutrition. These extracellular enzymes are actively excreted into the soil by organisms, and also accumulate in soil, protected by sorption to organo-clay complexes in the soil. These enzymes are important not only for the degradation of freshly added organic residues, but also for the release of inorganic P from soil organic matter.

## Microbial Solubilization of Inorganic P

Microorganisms are also important in the solubilization of P from mineral sources. Orthophosphate is typically present in soil solution at very low concentrations. This solution P is replenished either from mineralization of organic P sources, as discussed above, or from the dissolution of P from phosphate minerals. The primary P-containing mineral in soils is apatite or rock phosphate. In weathered, acidic soils, little apatite typically remains. In these soils, secondary inorganic P minerals in the form of aluminum and iron phosphates, such as variscite ($AlPO_4 \cdot 2H_2O$) and strengite ($FePO_4 \cdot 2H_2O$), dominate. In alkaline soils, the calcium phosphates dominate (e.g., $Ca_3(PO_4)_2$).

Most forms of inorganic P in soil have limited solubility (**Figure 1**). It is generally agreed that microorganisms can facilitate the enhanced dissolution of these compounds by at least two mechanisms: production of organic acids and excretion of protons during $NH_4^+$ assimilation. These mechanisms are performed by soil fungi and bacteria and appear to be potentially important, especially in soils receiving little or no phosphate fertilizer materials.

Phosphate solubilization has been observed for many different species of bacteria, fungi, and actinomycetes in laboratory studies. Bacterial species reported include those in the genera *Bacillus*, *Micrococcus*, and *Pseudomonas*. The most commonly observed P-solubilizing fungi are *Aspergillus* and *Penicillium* species, although a number of other species have been reported to show P-solubilizing activity. In soil, P-solubilization activities are greatest in the rhizosphere. The plant root system provides ample quantities of readily utilized carbon which results in high microbial activity. The activity of P-solubilizing organisms in the rhizosphere can be potentially important in the P nutrition of plants. While not a direct topic of this chapter, it should be noted that mycorrhizal fungi are also active in solubilization and uptake of P which benefits the host plant. (For more detail) on mycorrhizal activities, see **Mycorrhizal Fungi**. In general, it appears that fungi may be more effective in soil than bacteria. Fungal hyphae can grow across distance in soil microenvironments, allowing for exploitation of P resources not available to relatively immobile bacteria.

Organic acids produced by bacteria and fungi have been shown to solubilize various forms of inorganic phosphates (**Figure 4**). Citrate, and other organic acids, can be quite effective in solubilizing only minimally soluble forms of P, such as $AlPO_4$. Organic acids are effective as P-solubilizing agents in liquid medium because they cause a reduction in pH in the medium, facilitating acid hydrolysis of the $AlPO_4$, and the released $Al^{3+}$ is then complexed by the organic anion. This chelation or complexation of the released Al helps to minimize reprecipitation of Al-phosphates and the potential for Al-toxicity. While organic acids appear to be effective in solubilizing Al-phosphates, solubilization is typically more effective in the presence of microbial cells than that observed with organic acids alone. Also, some organisms solubilize Al-phosphates with no appreciable production of organic acids. Proton production during the assimilation of $NH_4^+$ or through respiratory activities appears to be a primary mechanism when organic acids are not produced.

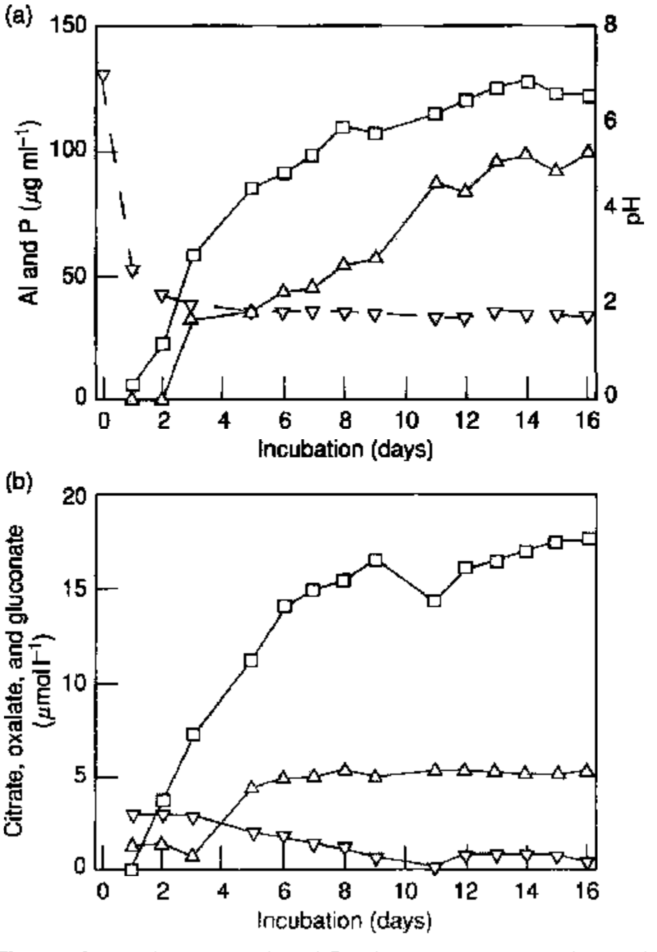


**Figure 4** (a) Al (squares) and P (triangles) concentration and pH (inverted triangles) and (b) production of organic acids (squares, citrate; triangles, oxalate; and inverted triangles, gluconate) during the solubilization of $AlPO_4$ by the fungus *Aspergillus niger*. Reproduced with permission from Illmer P, Barbato A, and Schinner F (1995) Solubilization of hardly-soluble $AlPO_4$ with P-solubilizing microorganisms. *Soil Biology and Biochemistry* 27: 265–270.

Dissolution of Ca-phosphates, such as rock phosphate or $Ca_3(PO_4)_2$, also occurs by organic acid production and/or the production of protons during ammonium assimilation or respiration. Typically, a decrease in medium pH occurs during solubilization of P from these minerals. The pH decrease is most pronounced, and solubilization is usually highest, in the presence of ammonium N sources. While organic acids can be effective in solubilization of Ca-phosphates, they are not always produced by all organisms capable of the process. Again, production of protons by physiological processes is important in this regard.

## P-Solubilizing Microorganisms and Plant Growth

Inoculation of crops with P-solubilizing microorganisms has met with mixed results. In the 1950s in the Soviet Union, researchers reported excellent results when *Bacillus megatherium* var. *phosphaticum* (also referred to as phosphobacterin) spores were applied to soils that were neutral to alkaline in pH and with high organic matter. The mode of action was thought to be the mineralization of organic P. Research in the USA was not able to verify the claims of significant crop yields and the practice was never adopted. However, other P-solubilizing microorganisms have been used with varying degrees of success in greenhouse and field situations.

Many greenhouse studies have shown positive effects of inoculation on yield and P uptake for several crops, including beans, peas, wheat, soybean, and sorghum. Numerous field studies have also been conducted. In some cases, there is little or no effect, while other research has shown positive effects of inoculation on yields and P uptake. Crops reported to respond

**Table 1** Effect of the P-solubilizing fungus *Penicillium bilaji* inoculation on wheat dry-matter production and P uptake under field conditions with rock phosphate (RP) as a P source

| Treatment | *Dry matter*<br>*Grain* *(g plot$^{-1}$)* | *Straw* *(g plot$^{-1}$)* | *P uptake* *(mg plot$^{-1}$)* |
|---|---|---|---|
| *P. bilaji* + RP | 207.7 | 402.7 | 834 |
| *P. bilaji* | 197.9 | 401.8 | 850 |
| RP | 186.8 | 388.3 | 771 |
| Monoammonium phosphate | 208.6 | 419.5 | 975 |
| Control | 180.6 | 378.9 | 742 |
| Tukey's HSD (0.05) | 23.6 | 42.3 | 61 |

HSD, honestly significant difference.

Adapted with permission from Kucey RMN (1988) Effect of *Penicillium bilaji* on the solubility and uptake of P and micronutrients from soil by wheat. *Canadian Journal of Soil Science* 68: 261–270

positively to inoculation with P-solubilizing fungi and bacteria include wheat, canola, soybean, and others. An example of the impact of *Penicillium bilaji* on wheat growth and P uptake when used with rock phosphate is shown in **Table 1**. It should be noted that inoculation of soil or growth media with microorganisms can positively impact plant growth due to production of plant growth-promoting compounds or through suppression of deleterious microorganisms. However, any organism that positively impacts root growth and soil exploration will likely lead to better P availability for the plant.

## Phosphorus and the Environment

Interest in the impact of P on the environment has greatly increased over the past decade. Overapplication of P-containing materials, whether inorganic fertilizers or organic sources such as manures and biosolids, can result in high levels of extractable soil inorganic phosphate. If these soils are close to surface waters, then this P can be exported to those waters either as soluble P in runoff water or as P adsorbed to eroding soil particles. In both cases, this can result in the elevation of the bioavailable P in these waters. This can potentially lead to eutrophication of the waters. Eutrophication occurs when excess nutrients, typically N or P, accumulate in the water, resulting in accelerated growth of algae. These algae can cloud the water, decreasing light penetration, and resulting in decreased growth of aquatic plants. Increased algal biomass production stimulates higher rates of degradation in the water column by bacteria. This accelerated decomposition then decreases available oxygen, often making the habitat unfit for many forms of aquatic life.

**Table 2** Inorganic (Pi) and organic (Po) P content in the 0–15-cm depth of nine Tennessee soils treated with animal manures. Soil textures range from sandy loams to silt loams. Four soils received swine manure, three received dairy manure, and two received poultry litter

| | *Total P* *(mg kg$^{-1}$)* | *Po* *(mg kg$^{-1}$)* | *Pi* *(mg kg$^{-1}$)* | *% Pi* |
|---|---|---|---|---|
| Inorganic fertilizers | 975 | 793 | 182 | 18.7% |
| Manured soils | 1478 | 1049 | 429 | 29.0% |

Many soils that have received regular applications of manure over time have become quite high in total and available P. If manure is applied to a crop to satisfy N requirements, too much P is typically applied, resulting in a P buildup. Although much of the P in these manures is added in the organic form, the result is typically a significant increase in the total amount of inorganic phosphate due to mineralization by soil microorganisms. **Table 2** shows the impact of long-term manuring on total P, organic P, and inorganic P content of nine soils in Tennessee. Note that the inorganic P content is on average 2.4 times higher in the manured soils, and the proportion of inorganic P increased by 10.3%. Microbial mineralization of the manure P releases the phosphate and more of it tends to remain relatively labile. This may be due to increased soluble soil C which may then bind to P-sorption sites or with soil Al and Ca, reducing complexation. Microbial mineralization of these organic P sources over long periods of time can therefore result in potential offsite degradation of water quality.

*See also:* **Eutrophication**; **Fertility**; **Fertilizers and Fertilization**; **Macronutrients**; **Mycorrhizal Fungi**; **Phosphorus in Soils:** Overview

## Further Reading

Germida JJ and Siciliano SD (2000) Phosphorus, sulfur and metal transformation. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. C95–C106. Boca Raton, FL: CRC Press.

Haygarth PM and Jarvis SC (1999) Transfer of phosphorus from agricultural soils. *Advances in Agronomy* 66: 195–249.

Hocking PJ (2001) Organic acids exuded from roots in phosphorus uptake and aluminum tolerance of plants in acid soils. *Advances in Agronomy* 74: 63–97.

Illmer P, Barbato A, and Schinner F (1995) Solubilization of hardly-soluble $AlPO_4$ with P-solubilizing microorganisms. *Soil Biology and Biochemistry* 27: 265–270.

Kucey RMN (1988) Effect of *Penicillium bilaji* on the solubility and uptake of P and micronutrients from soil by wheat. *Canadian Journal of Soil Science* 68: 261–270.

Kucey RMN, Janzen HH, and Leggett ME (1989) Microbially mediated increases in plant-available phosphorus. *Advances in Agronomy* 42: 199–228.

Kuo S (1996) Phosphorus. In: Sparks DL, Page AL, and Helmke PA *et al.* (eds) *Methods of Soil Analysis. Part 3. Chemical Methods.* SSSA Book Series no. 5, pp. 869–919. Madison, WI: Soil Science Society of America.

Lindsay WL (1979) *Chemical Equilibria in Soils.* New York: John Wiley.

Mullen MD (1998) Transformations of other elements. In: Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (eds) *Principles and Applications of Soil Microbiology*, pp. 369–386. Upper Saddle River, NJ: Prentice-Hall.

Oehl F, Oberson A, Probst M, Fliessbach A, Roth H-R, and Frossard E (2001) Kinetics of microbial phosphorus uptake in cultivated soils. *Biology and Fertility of Soils* 34: 31–41.

Pierzynski GM, Sims JT, and Vance GF (2000) *Soils and Environmental Quality.* Boca Raton, FL: CRC Press.

Stevenson FJ and Cole MA (1999) *Cycles of Soil: Carbon, Nitrogen, Phosphorus, Sulfur, Micronutrients.* New York: John Wiley.

Tabatabai MA (1994) Soil enzymes. In: *Methods of Soil Analysis. Part 2. Microbiological and Biochemical Properties.* SSSA Book Series no. 5, pp. 775–833. Madison, WI: Soil Science Society of America.

Walbridge MR (1991) Phosphorus availability in acid organic soils of the lower North Carolina coastal plain. *Ecology* 72: 2083–2100.

Whitelaw MA (2000) Growth promotion of plants inoculated with phosphate-solubilizing fungi. *Advances in Agronomy* 69: 99–151.

# PHYTOTOXIC SUBSTANCES IN SOILS

**M Qadir**, International Center for Agricultural Research in the Dry Areas (ICARDA), Aleppo, Syria; and International Water Management Institute (IWMI), Colombo, Sri Lanka
**S Schubert and D Steffens**, Justus Liebig University, Giessen, Germany

## Introduction

Plants are continuously exposed to their ambient soil environment during the whole span of their life-cycle. Sometimes the presence, application, and/or formation of certain substances in soils reach to the extent that they become phytotoxic – the property of a substance at a specified concentration that restricts or constrains the plant growth. With a wide range of characteristics and impacts, phytotoxic substances have long been accumulating in the terrestrial ecosystem, including soils. Once these substances enter the soil, they become part of a cycle that affects all forms of life. Soils act as a sink and a filter, protecting the groundwater from inputs of phytotoxic substances during initial stages of their addition. With continued uncontrolled applications, the concentrations of potentially harmful substances may increase to levels that become toxic to a wide range of plants, and in certain cases toxicity can occur in soil faunas and floras, higher animals, and humans.

The contribution of several different human activities to accumulation of phytotoxic substances has been several times greater than natural sources. Although the use of chemicals expanded in the nineteenth century, it was not until the middle of the twentieth century that the use of chemicals became widespread. In recent years, the quantities of many products in which phytotoxic substances are used have increased notably, enhancing the opportunity for contamination and phytotoxicity in soils. Although there are several forms of phytotoxic substances in soils, they can be grouped into five broad categories: (1) pesticides, (2) metals and metalloids, (3) soil acidity and aluminum and manganese toxicity, (4) soil flooding and manganese and iron toxicity, and (5) soluble salts and sodium toxicity.

## Pesticides

Pesticides are substances or mixtures thereof intended to prevent, destroy, repell, or mitigate any pest, which may be an insect, rodent, nematode, fungus, weed, or other form of terrestrial or aquatic plant or animal

**Table 1** Class, chemical group, and persistence time of some pesticides

| *Pesticide class* | *Chemical group* | *Pesticide name* | *Persistence time* |
|---|---|---|---|
| Insecticide | Chlorinated hydrocarbons | DDT | 3–15 years |
| Insecticide | Chlorinated hydrocarbons | Chlordane | 2–5 years |
| Insecticide | Chlorinated hydrocarbons | Dieldrin | 2–5 years |
| Herbicide | Triazines | Atrazine | 1–2 years |
| Herbicide | Triazines | Simazine | 1–2 years |
| Herbicide | Benzolics | Dicamba | 2–12 months |
| Herbicide | Phenoxyalkyl acids | 2,4-D | 1–5 months |
| Insecticide | Organophosphates | Diazinon | 1–12 weeks |
| Insecticide | Organophosphates | Malathion | 1–12 weeks |
| Insecticide | Carbamates | Carbaryl | 1–8 weeks |

life that is injurious to growing plants, animal or human health, or the environment. The chemical revolution in agriculture was initiated when the insecticidal properties of dichlorodiphenyltrichloroethane (DDT) were discovered in 1939 along with the herbicidal effects of 2,4-dichlorophenoxyacetic acid (2,4-D) in 1941. Since their discovery, tens of thousands of such chemicals and multichemical formulations have been developed, tested, and put to use. In the past 60 years, synthetic organic pesticides have become a major element in modern agricultural production practices. The advent of pesticide use has coincided with the tremendous increases in agricultural productivity. Although the use of pesticides has provided many benefits to society by protecting crops, animals, and humans, it has also created several problems, including widespread accumulation of residues, with damage to plants, wildlife, fisheries, beneficial insects, and even humans.

The impact of pesticides on plants may be as dramatic and devastating as the impact on insects and animals. Exposure of a sensitive plant may produce remarkable growth distortions and yield loss. During the process of application, pesticides may move from the treatment area to deposit a residue affecting a nontarget species. In the case of some pesticides, the amount remaining after a prior treatment can be a limiting factor for the subsequent crop(s). Even though the quantity of pesticide would be small, it could be sufficient to limit the growth of a sensitive species.

Pesticides are commonly classified according to the target group used to control, deter, or destroy pest organisms: (1) insecticides, (2) fungicides, (3) herbicides, (4) rodenticides, and (5) nematocides. The first three are used in much larger quantities and therefore contaminate soils to a greater extent than rodenticides and nematocides. Whatever the mode of application of pesticides, a high proportion of such chemicals eventually moves into the soil. After finding their way into soils, these chemicals then meet with one or a combination of the following fates as they may: (1) vaporize into the atmosphere without a chemical change, (2) adsorb on soil colloidal complex that consists of humus and clay particles, (3) move with the percolating water to the deeper soil layers or to the groundwater, (4) undergo chemical reactions with other chemicals, (5) be broken down by soil microorganisms, and/or (6) be subjected to absorption by plants. The specific fate of different pesticides in soils is determined by their chemical structures and properties, various processes occurring in soils (retention, transformation, and transport processes and their interactions), climatic conditions, and different forms of anthropogenic activities dealing with the management of pesticides in soils. Some pesticides are not readily biodegradable and persist in soils for many years. Thus there is a large variation in the persistence time of different pesticides in soils (**Table 1**).

## Metals and Metalloids

Metals are a group of chemical elements that exhibit all or most of the following physical qualities: (1) solid at ordinary temperatures, (2) opaque except in extremely thin films, (3) good electrical and thermal conductors, (4) lustrous when polished, and (5) crystalline when in the solid state. Metalloids such as arsenic (As) and selenium (Se) are elements with properties intermediate between those of metals and nonmetals. Some metals and metalloids such as zinc (Zn), copper (Cu), manganese (Mn), nickel (Ni), cobalt (Co), chromium (Cr), molybdenum (Mo), iron (Fe), and Se are essential for living organisms. However, these metals and those that are thought to be nonessential such as lead (Pb), mercury (Hg), cadmium (Cd), and As can pose risks when they build up in soils due to many forms of anthropogenic influences (**Table 2**). When their bioavailability is high, toxicity can result, with negative effects on plants.

Although metals and metalloids are usually present in soil parent materials, they have several different

**Table 2** List of anthropogenic sources of some selected metals and metalloids

| *Chemical* | *Symbol* | *Major uses and sources*[a] |
|---|---|---|
| Arsenic | As | Pesticides, animal feed additives, and manures from intensive animal production, desiccants, fossil fuel combustion, electronics, pigments and paints, medical uses |
| Cadmium | Cd | Electroplating, pigments for plastics and paints, batteries, sewage sludge, fossil fuel combustion, polymer stabilizers, waste disposal, impurities in fertilizers |
| Chromium | Cr | Chrome-plated metals, pigments, electronics, fossil fuel combustion, impurities in fertilizers, wood preservatives, stainless steel |
| Copper | Cu | Fly ash, wind-blown copper containing dust, impurities in fertilizers, warfare and military training, waste disposal, medical uses, electronics, sewage sludge, wood preservatives, pesticides |
| Lead | Pb | Fossil fuel combustion, batteries, pigments and paints, polymer stabilizers, printing, sports shooting, warfare, and military training, waste disposal, impurities in fertilizers, pesticides |
| Mercury | Hg | Pesticides, chlorine manufacture, medical uses, batteries, electronics |
| Nickel | Ni | Alloy manufacture, electroplating, sewage sludge, batteries, catalyst |
| Selenium | Se | Pigments and paints, printing, medical uses |
| Zinc | Zn | Metallurgical industries, electronics, pesticides, impurities in fertilizers, sewage sludge, pigments and paints, polymer stabilizers, printing, medical uses |

[a] In addition to mining of ore bodies in the Earth's crust and smelting of scrap material, and weathering of soil minerals.

anthropogenic sources, including: (1) mining of ore bodies in the earth's crust and smelting of scrap material; (2) sewage sludge produced from the treatment of domestic and industrial waste waters; (3) combustion of fossil fuel consisting of coal and petroleum; (4) manufacture, use, and disposal of metal commodities in metallurgical industries; (5) manufacture, use, and disposal of semiconductors, cables, contacts, and other electrical components of electronic industries; (6) chemical and other industrial sources such as chlorine manufacture, batteries, pigments and paints, catalysts, polymer stabilizers, printing, medical uses, and additives in fuels and lubricants; (7) disposal of household, municipal, and industrial wastes; (8) flood or seepage waters; and (9) sports shooting, warfare, and military training. As an important anthropogenic source of metals and metalloids in soils, the disposal of domestic, municipal, and industrial wastes also contributes to significant quantities of certain substances that can produce harmful environmental effects. In some cases, such wastes may contain less decomposable organic substances. In addition to the nonagricultural sources, agricultural practices constitute important sources of metals which make a significant contribution to their total concentrations in soils in many parts of the world, especially in regions of intensive farming. The main sources are impurities in fertilizers, application of sewage sludge as a source of plant nutrients and organic matter, manures from intensive animal production, use of pesticides, irrigation with metal-rich waters, and corrosion of metal objects.

Atmosphere and water are important transport media for metals from different sources. Soils are often contaminated for up to hundreds of kilometers away from the site of emission. The metals are usually present in air as aerosol particles with a size range of 5–20 $\mu$m, but most are between 0.1 and 10 $\mu$m in diameter and have a mean residence time of 10–30 days. A high proportion of the metals in more-recent dust deposits are of anthropogenic origin. Water, on the other hand, is an integral component of the weathering processes relating to the release of metals from soil parent material. In addition, contamination of either surface water or groundwater impacts the quality of each other and subsequently the soils irrigated with such waters.

The total metal content of a soil is the result of inputs of metals from the abovementioned sources minus losses in metals removed in crop material, leaching, and volatilization. However, it is not the total amounts but the bioavailable concentrations that are crucial and depend on a number of soil and environmental factors, which also control the speciation of metals in soils. Thus the effect of any metal on plant growth depends not only on its chemical properties but also on its bioavailable concentration. The physiological age and species of the plant concerned are also important.

## Soil Acidity and Aluminum and Manganese Toxicity

Soil acidity is usually expressed in terms of pH scale, which quantifies the activity of protons ($H^+$) present in a solution. Soil acidification can develop naturally when a soil profile is leached, but can be accelerated by certain farming practices and acid precipitation, which is commonly known as acid rain. Various amounts of sulfuric and nitric acids ($H_2SO_4$ and $HNO_3$) form in precipitation from certain sulfur (S) and nitrogen (N) gases produced by a number of sources such as lightning, volcanic eruptions, forest fires, and the combustion of fossil fuels. Since $N_2$-fixing legumes release $H^+$ in the root zone, farming practices dealing with continuous cultivation of such legumes may increase the acidity of acid soils. This contribution is crucial when all or a major part of the harvested shoot is not added back to the soil as a measure to control soil acidity. On the other hand, the effects of acid rain are more pronounced on the acidity of water than on soil acidity. Soils generally are sufficiently buffered to accommodate acid rain with little increase in acidity on an annual basis. However, continued inputs of acid rain would have significant effects on the pH of acid soils, especially those that are weakly buffered. Acid rain is also a serious problem for soils that are already highly acidic, since a small increase in acidity could make them even less fertile. Other acidifying processes in soils include: (1) formation of carbonic acid ($H_2CO_3$) from carbon dioxide ($CO_2$) in the root zone; (2) dissociation of acids; (3) oxidation of N, S, and Fe compounds; (4) accumulation of organic matter; and (5) deprotonation of pH-dependent charges. The growth of sensitive plants may be adversely affected by high free-$H^+$ concentration in soils.

Soils become extremely acidic (pH $< 4$) when large amounts of reduced forms of S are oxidized to $H_2SO_4$ if the soils are exposed to oxygen when they are drained or excavated. With a $\geq 0.15$-m-thick sulfuric horizon that is often present, such soils are commonly referred to as acid sulfate soils. These soils could be categorized into potential acid sulfate soils ($H_2SO_4$ may be produced), active acid sulfate soils ($H_2SO_4$ is being produced), or postactive acid sulfate soils ($H_2SO_4$ has been produced). The amounts of S present in these soils provide the basis for estimating the amount of liming material ($CaCO_3$) required to neutralize the $H_2SO_4$ that could be produced. In many cases, the volume of soil that needs to be treated and the amounts of $CaCO_3$ required are huge enough to make such neutralization an impractical solution to the problem.

Aluminum (Al) is the most abundant metal in the Earth's crust, comprising about 7% of its mass, and is found in soils primarily in the form of insoluble aluminum silicates and oxides. Although Al is a nonessential element, its toxicity occurs in acid soils, which is considered as the most common and severe problem associated with such soils. When soils become more acidic, phytotoxic species of Al such as $Al^{3+}$, $Al(OH)^{2+}$, and $Al(OH)_2^+$ are released into soil solution. Since most plant species are sensitive to micromolar concentrations of Al, toxicity of this metal may become a major limiting factor for plant productivity on acid soils. Aluminum toxicity has been recognized as a major constraint for the production of several crops in acid soils for the past 80 years. When plants grow and develop under acid soils in natural conditions, they are continuously exposed to Al and the damage is cumulative. As a general phenomenon, Al toxicity primarily restricts root growth, with an effect on plant metabolism through interference with the uptake, transport, and use of several nutrients by plants. However, it is not known whether these observed effects are the primary cause of Al toxicity or they are only indirect or secondary effects. In addition to plants, many bacteria such as those that are involved in certain transformations in N cycle are also adversely affected by high Al levels associated with low soil pH. In contrast to Al, Mn is an essential plant nutrient, which is readily transported from roots to shoots and, therefore, symptoms of Mn toxicity first appear on shoots. Excessive levels of Mn in soils also interfere with the uptake of other mineral nutrients such as calcium ($Ca^{2+}$) and magnesium ($Mg^{2+}$). Plant species and genotypes within a species vary greatly with regard to their susceptibility to Mn toxicity. Manganese toxicity is not as widespread as Al toxicity.

## Soil Flooding and Manganese and Iron Toxicity

With the exception of rice (*Oryza sativa* L.) and Kallar grass (*Leptochloa fusca* (L.) Kunth.), periodic flooding during the growing season adversely affects growth and productivity of crops. Soil-flooded conditions are also expressed through terms such as 'soil saturation,' 'waterlogging,' 'anoxia,' and 'hypoxia.' Lack of oxygen has been found to be the major problem associated with soil flooding that impacts plant growth. A great deal of information during recent decades has accumulated from research on the molecular, biochemical, and physiological responses of a number of plant species to flooding of the growth medium. The use of molecular plant-breeding and genomic transformations to improve flooding tolerance in crops has been rated as a promising solution to this problem.

In addition to oxygen stress, reduced forms of certain elements can also cause phytotoxicity in flooded soils. For instance, the reduced forms of Mn and Fe are far more soluble than their oxidized forms. Therefore, the toxicity of Mn and Fe is greatly increased by low-oxygen conditions, particularly in acid paddy soils. Other growth-reducing and yield-loss factors in flooded soils could arise from root rot disease, N deficiency, nutrient imbalance, and/or accumulation of high levels of $CO_2$ in the root zone. Such levels of $CO_2$ in flooded soils could cause severe leaf chlorosis and reduction in plant biomass in flood-sensitive crops such as soybeans (*Glycine max* L.).

## Soil Salinity and Sodium Toxicity

Accumulation of soluble salts can occur in soils to levels that can adversely affect crop growth and yield. Salt-affected soils mostly exist under arid and semiarid regions. Major cations in salt-affected soils are $Na^+$, $Ca^{2+}$, $Mg^{2+}$, and, to a lesser extent, potassium ($K^+$). The major anions are chloride ($Cl^-$), sulfate ($SO_4^{2-}$), bicarbonate ($HCO_3^-$), carbonate ($CO_3^{2-}$), and nitrate ($NO_3^-$). Salt-affected soils are generally classified into two main categories, saline and sodic. Soil salinity is generally characterized by measuring electrical conductivity of the saturated soil paste extract ($EC_e$) and expressed in units of deci-siemens per meter. In some parts of the world, saline soils are also known as white alkali soils. Soil sodicity is evaluated in terms of the relative amounts of $Na^+$ on the cation exchange sites (exchangeable sodium percentage, ESP) or in the soil solution (sodium adsorption ratio, SAR), given the accompanying levels of salinity. Other names for sodic soils include 'black alkali soils,' 'solonetz,' and 'slick-spots.'

The presence of voluminous data indicates that there are no single values of $EC_e$ and ESP (or SAR) for saline and sodic soils, respectively. In general, an $EC_e$ of $4\,dS\,m^{-1}$ is used as the boundary between saline and nonsaline soils; the soils of higher $EC_e$ are called 'saline.' Actually, salt-sensitive plants are affected at half of this salinity ($2\,dS\,m^{-1}$) and highly tolerant plants at about twice of this salinity ($8\,dS\,m^{-1}$) or at an even greater value. Similarly, an ESP of 15 (an ESP of 6 in Australia) is considered as the critical limit for soil sodicity; the soils of higher ESP show dispersive behavior in most cases and suffer from serious physical problems when wetted. Different terminology for alkali soils is used in some parts of the world to express sodicity hazard. Alkali soils are characterized by high sodicity (ESP of more than 15) and pH (pH >8.3), and contain soluble $CO_3^{2-}$ and $HCO_3^-$ of $Na^+$, i.e., $C_{Na}:(C_{Cl}+C_{SO_4})$ ratio greater than 1. These soils contain $Na^+$ and $CO_3^{2-}+HCO_3^-$ as the dominant ions and tend to have low salinities and high pH values, which cause an increase in swelling and dispersion. On the other hand, the pH of sodic soils can be either more than or less than 7, and such soils can be either saline or nonsaline. Therefore, use of the term 'alkali' provides a practical distinction in terms of soil management.

Deterioration of salt-affected soils occurs through changes in the proportions of soil solution and exchangeable ions and soil reaction (pH) that lead to osmotic and specific ion effects together with imbalances in the availability of plant-available nutrients. Excess $Na^+$ in soils accelerates the development of structural problems created by certain physical processes (slaking, swelling, and dispersion of clay minerals) and specific conditions (surface crusting and hard-setting) that may affect water and air movement, plant-available water-holding capacity, root penetration, seedling emergence, runoff, erosion, as well as tillage and sowing. These chemical and physical changes have a bearing on the activity of plant roots and soil microbes, and ultimately on crop growth and yield.

Excess soluble salts in salt-affected soils originate either through the weathering of parent minerals (causing fossil or primary salinity and sodicity) or from man-made activities involving improper use and management of land and water resources (contributing to secondary salinity and sodicity). Secondary salinity and sodicity of arable land resources are as old as the history of human settlement and irrigation. A well-documented case occurred in Mesopotamia where, between 4000 and 2000 BC, secondary salinization affected the land in the valleys of the Euphrates and Tigris as a result of improper irrigation practices. History repeated itself in the twentieth century in many countries and regions. Although the vast river systems in arid and semiarid regions with extensive irrigated agriculture supply water for irrigation, they also contain appreciable quantities of salts. The inappropriate use of such irrigation systems accelerates the development of secondary salinity and sodicity, which plague these regions. Dryland (or nonirrigated) salinization is commonly due to rising water tables over essentially saline subsoils caused by clearance of trees or other deep-rooted crops, usually to make land available for rain-fed agriculture. The area under human-induced salinity and sodicity has increased substantially, particularly in the twentieth century.

## Remediation Strategies for Phytotoxic Substances

There is a long history of accumulation and control of phytotoxic substances in soils. The concentrations of such substances in soils are increasing because the inputs are not balanced by the output in terms of removal from the plough layer. Although the accumulation of phytotoxic substances in soils provides a direct stimulus for cleanup, mostly public and political pressure to reduce these substances in soils only occurs when critical levels are reached. The past focus to remove phytotoxic substances from soils was on the use of traditional physical and chemical methods. Physical remediation is the process of correcting the problem by a number of physical means. For example, containment of pollutants could be done by covering the polluted site with a layer impermeable to rain water. Clay or cement has been used in the past to provide an impermeable cap preventing water from reaching the pollutants and leaching them away. Chemical remediation involves the use of a variety of chemicals for different phytotoxic substances. For example, $CaCO_3$ is used to neutralize $H_2SO_4$ in acid sulfate soils, while gypsum ($CaSO_4 \cdot 2H_2O$) has been a commonly used amendment to ameliorate excess $Na^+$ in sodic soils. The enormous cost associated with the physical and chemical means is one of the major reasons for slow progress in the removal of phytotoxic substances from the soils. Therefore, efforts have been made to minimize the cost of amelioration of such soils. In recent decades, strategies assisted by certain microorganisms and higher plants have shown the potential to remediate the soils at relatively low costs. These strategies are gaining an increasing interest among policy makers and scientists. Biological remediation, commonly known as bioremediation, deals with cleaning soils with the assistance of certain microorganisms that have the capability of breaking down some specific phytotoxic substances. Persistent pesticides such as DDT and dioxin have been reported to succumb to the attack of a white rot fungi species, *Phanerochaete chrysosporium*. Phytoremediation is a plant-based approach, which involves the use of suitable crops, grasses, shrubs, and trees that have the potential to grow on contaminated soils with the ability to improve such soils in several different ways.

Phytoremediation of metals and metalloids can be done through phytoextraction, which involves the use of pollutant-scavenging plants to transport and concentrate metals and metalloids from the soil into the aboveground shoot. The efficiency of phytoextraction for a given metal ($Phyto_{Metal}$) is the product of the dry-matter yield of shoot ($S_{DM}$) and concentration of the metal in shoot ($S_{CM}$), as given in the following equation:

$$Phyto_{Metal} = S_{DM} \times S_{CM} \qquad [1]$$

The plants able to accumulate high concentrations of metals are known as hyperaccumulators (**Table 3**). The concentrations of metals accumulated in such plants may be 100 times greater than those occurring in nonaccumulator plants growing on the same substrates. Following the harvest of the metal-enriched plants, the weight and volume can be reduced by ashing. Metal-enriched plants can be disposed of as hazardous material or, if economically feasible, used for metal recovery. Phytoremediation can be done through reduction in the bioavailability of metals in the soil (phytostabilization) and volatilization of pollutants such as Hg and Se from the foliage (phytovolatilization).

**Table 3** Some important metal hyperaccumulators

| *Element* | *Metal accumulator plant species* |
|---|---|
| Cadmium | *Thlaspi caerulescens* |
| Cobalt | *Haumaniastrum robertii* |
| Copper | *Haumaniastrum katangense* |
| Lead | *Thlaspi rotundifolium* subsp. |
| Manganese | *Macadamia neurophylla* |
| Nickel | *Berkheya coddii* |
| Selenium | *Astragalus pattersoni* |
| Thallium | *Iberis intermedia* |
| Uranium | *Atriplex confertifolia* |
| Zinc | *Thlaspi calaminare* |

In contrast to the phytoremediation of metal-contaminated soils, phytoremediation of salt-affected soils works through plant roots to enhance leaching of soluble salts and $Na^+$ from the root zone to the deeper soil layers. The enhanced partial pressure of carbon dioxide ($P_{CO_2}$) in the root zone during cropping is considered as a major factor contributing to the amelioration of calcareous sodic soils. Aqueous $CO_2$ helps in the dissolution of native soil calcite ($CaCO_3$) to provide $Ca^{2+}$ for the $Na^+$–$Ca^{2+}$ exchange at the cation exchange sites. In addition, $H^+$ released by the roots of certain phytoremediation crops such as $N_2$-fixing legumes may react with soil $CaCO_3$ to provide $Ca^{2+}$. Phytoremediation of sodic soils ($Phyto_{Sodic}$) is considered as a function of the following factors:

$$Phyto_{Sodic} = \sum R_{P_{CO_2}} + R_{H^+} + R_{Phy} \qquad [2]$$

where $R_{P_{CO_2}}$ refers to $P_{CO_2}$ within the root zone, $R_{H^+}$ is $H^+$ release in the root zone of $N_2$-fixing crops, and $R_{Phy}$ refers to physical effects of roots in improving soil aggregation and hydraulic properties of the root zone. The collective effects of these factors ultimately lead to soil improvement, provided drainage is adequate and leaching occurs. The salinity–sodicity combination in soil solution during phytoremediation remains suitable to maintain adequate soil structure and aggregate stability, which prevent reduction in soil permeability to air and water.

Phytoremediation of sodic soils has been shown to be advantageous in several economic, environmental, and agronomic aspects: (1) no financial outlay to purchase chemical amendments, (2) financial and other benefits from plant species grown during phytoremediation, (3) promotion of soil aggregate stability and creation of macropores that improve soil hydraulic properties, (4) better plant nutrient availability in soil during and after phytoremedition, and (5) more uniform and greater zone of amelioration in terms of soil depth. In general, phytoremediation

**Table 4** Some plant species used for phytoremediation of salt-affected soils

| *Common name* | *Botanical name* |
|---|---|
| Alfalfa | *Medicago sativa* L. |
| Barley | *Hordeum vulgare* L. |
| Bermuda grass | *Cynodon dactylon* (L.) Pers. |
| Kallar grass | *Leptochloa fusca* (L.) Kunth |
| Millet rice | *Echinochloa colona* (L.) Link |
| Para grass | *Urochloa mutica* (Forssk.) T. Q. Nguyen |
| Rhodes grass | *Chloris gayana* Kunth |
| Rice | *Oryza sativa* L. |
| Sesbania | *Sesbania bispinosa* (Jacq.) W. Wight |
| Sordan | *Sorghum* × *drummondii* (Steud.) Millsp. & Chase |
| Sorghum | *Sorghum bicolor* (L.) Moench |

works well on coarse- to medium-textured and moderately salt-affected soils. On such soils, its performance has been found to be comparable with soil application of chemical amendments such as gypsum. Several plant species of agricultural significance have been successfully grown as a phytoremediation tool on a variety of salt-affected soils (Table 4). Phytoremediation may face certain limitations under conditions of very high levels of salinity and/or sodicity, where the growth of phytoremediation crops is likely to be variable and patchy. Under these conditions, the use of chemical amendments is probably necessary.

*See also:* **Acidity; Pesticides; Sodic Soils**

## Further Reading

Alloway BJ (1995) *Heavy Metals in Soils*. London, UK: Chapman and Hall.

Cheng HH (1990) *Pesticides in the Soil Environment: Processes, Impacts, and Modeling*. Madison, WI: Soil Science Society of America.

De la Fuente-Martinez JM and Herrera-Estrella L (1999) Advances in the understanding of aluminum toxicity and the development of aluminum-tolerant transgenic plants. *Advances in Agronomy* 66: 103–120.

Ghassemi F, Jakeman AJ, and Nix HA (1995) *Salinisation of Land and Water Resources: Human Causes, Extent, Management and Case Studies*. Wallingford, UK: CABI Publishing.

Hillel D (2000) *Salinity Management for Sustainable Irrigation: Integrating Science, Environment, and Economics*. Washington, DC: The World Bank.

Marschner H (1995) *Mineral Nutrition of Higher Plants*. London, UK: Academic Press.

McGrath SP, Zhao FJ, and Lombi E (2002) Phytoremediation of metals, metalloids, and radionuclides. *Advances in Agronomy* 75: 1–56.

Mengel K and Kirkby EA (2001) *Principles of Plant Nutrition*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Ponnamperuma FN (1972) The chemistry of submerged soils. *Advances in Agronomy* 24: 29–96.

Qadir M and Oster JD (2002) Vegetative bioremediation of calcareous sodic soils: history, mechanisms, and evaluation. *Irrigation Science* 21: 91–101.

Qadir M and Schubert S (2002) Degradation processes and nutrient constraints in sodic soils. *Land Degradation and Development* 13: 275–294.

Raskin I and Ensley BD (1999) *Phytoremediation of Toxic Metals: Using Plants to Clean Up the Environment*. New York: John Wiley.

Selim HM and Sparks DL (2001) *Heavy Metal Release in Soils*. Boca Raton, FL: CRC Press.

Terry N, Zayed AM, de Souza MP, and Tarun AS (2000) Selenium in higher plants. *Annual Review of Plant Physiology and Plant Molecular Biology* 51: 401–432.

# PLANT–SOIL–WATER RELATIONS

**R A Feddes and J C van Dam**, Wageningen University and Research Centre, Wageningen, The Netherlands

## Introduction

Water for evapotranspiration from land surfaces is supplied mainly by the soil. The soil water reservoir balances the episodic excesses of water supply from rainfall against the more smoothly varying atmospheric demand for evapotranspiration. The role of soil moisture within the soil–plant–atmosphere system depends on the soil moisture reservoir size and the availability of water in that reservoir, which, in turn depends on the texture and structure of the soil and the characteristics of the root system.

One key function of plant roots is their ability to link the soil, where water and nutrients reside, to the organs and tissues of the plant, where these resources are used by the canopy. Hence roots serve to connect the soil environment to the atmosphere by providing a first link in the pathway for fluxes of water and solutes through the plant to the atmosphere. Fluxes along the soil–plant–atmospheric continuum are regulated by aboveground plant properties such as the regulation of plant transpiration by the leaf stomata, which interact with the atmosphere, and plant root properties such as depth, distribution, and activity of roots, as well as soil properties such as water potential, water content, and hydraulic conductivity.

Plant root systems show a remarkable ability to adapt to soil depth and to changes in availability of water and nutrients and the chemical properties (e.g., salinity) in soils. Root response to soil properties in turn affects the uptake of soil water and nutrients and the storage of carbon belowground. Root distribution may change when ecosystems respond to greenhouse warming and carbon dioxide fertilization.

## Liquid Water Flow Inside the Plant

Water is absorbed in the *liquid state* by the roots from the soil. The general flow pattern is shown in **Figure 1**: soil → root cells → endodermis → xylem → leaf cells → stomata → air (the latter as vapor flow). Most of the transport (±80%) through the root and leaf cells takes place along the microcapillaries situated in the cell wall. However, the roots contain endodermis cells in which the continuous system of microcapillaries in the cell walls is interrupted by the so-called Casparian strips. Consequently most of the water has to pass through the cells itself. This means that the permeability of the endodermis cell and the semipermeable membrane (plasmalemma) can be an important determinant in controlling the water flow. Hence a low soil temperature and/or poor soil aeration will induce a low metabolic activity of the endodermis cell, which may cause a high resistance of the semipermeable membrane against water uptake. The resistance of the xylem vessels is quite low compared with the large resistances encountered in transport from cell to cell.

**Figure 2** shows a daily cycle of transpiration and water uptake which is regularly observed in summer. Before sunrise a plant is usually at full turgor, there is no transpiration, and stomata are fully closed. After sunrise, radiation increases and stomata open. During midday hours the transpiration demand increases, the water head in the leaves drops, because plant roots cannot take up water quickly enough, turgidity declines, and stomata will partially or completely close. This leads to a decrease in transpiration and an increase in leaf temperature. After closing the stomata, water head and turgidity will be restored, which may lead to opening of the stomata again in the afternoon. During nighttime there is complete recovery of water head and turgidity under nonevaporation conditions.

Accurate mathematical description of water uptake by crops with a nonuniform root system is complicated. Therefore it has been customary to describe water flow through the entire soil–plant–atmosphere system with an electrical analog. This model assumes that the water flux density $v_z$ (cubic centimeters per square centimeters per second) through the rooted soil zone and the root–stem–leaf-stomata path is proportional to the total water potential/head difference $\Delta h_{\mathrm{total}}$ (centimeters) and inversely proportional to the total resistance $R$ (seconds) of the system. Thus,

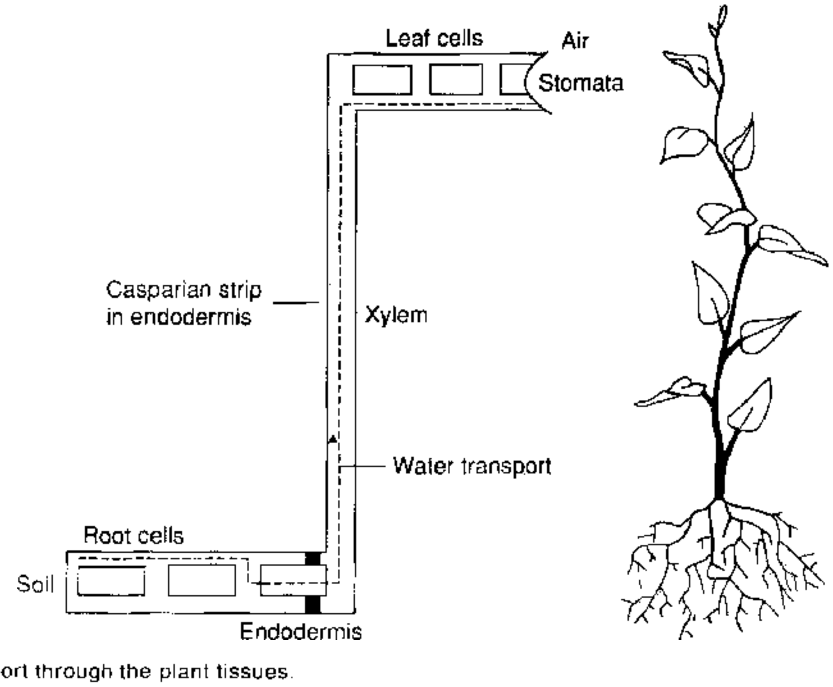


**Figure 1** Liquid transport through the plant tissues.

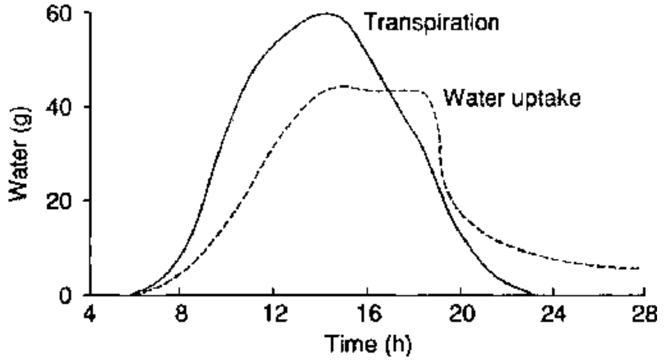


**Figure 2** Daily cycle of plant transpiration and root water uptake.

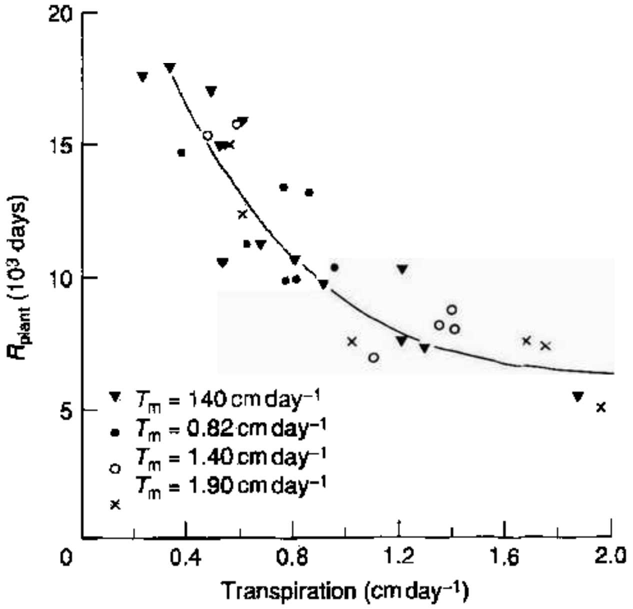


**Figure 3** Plant resistance $R_{plant}$ versus transpiration rate $T$ as computed from an Italian ryegrass column experiment, showing the inverse proportionality of $R_{plant}$ with $T$. (Reproduced from Hansen GK (1974) Resistance to water flow in soil and plants, plant water status, stomatal resistance and transpiration of Italian ryegrass as influenced by transpiration demand and soil water depletion. *Acta Agriculturae Scandinavica* 24: 83–92.)

considering a one-dimensional, steady-state flow in a series network, the liquid flow equation is:

$$T = -\frac{\Delta h_{total}}{R_{total}} = -\frac{(h_{root} - h_{soil})}{R_{soil}} = -\frac{(h_{leaf} - h_{root})}{R_{plant}} = \frac{h_{leaf} - h_{soil}}{R_{soil} + R_{plant}} \qquad [1]$$

where $T$ (centimeters per second) is the transpiration rate, $h_{soil}$, $h_{root}$, and $h_{leaf}$ (centimeters) are pressure heads in the soil, at the root surface and in the leaves, respectively, and $R_{soil}$ and $R_{plant}$ (seconds) are liquid flow resistances of the soil and the plant, which are assumed to be constant; hence $R_{plant}$ does not include stomatal resistance. When the transpiration demand of the atmosphere on the plant system is too high or when the soil is rather dry, $R_{soil}$ and $R_{plant}$ influence $h_{leaf}$ in such a way that transpiration is reduced by closure of the stomata.

Equation [1] can be applied to the root system as a whole by measuring $T$, $h_{soil}$, and $h_{leaf}$ during two periods, thus obtaining two equations with two unknowns, from which $R_{soil}$ and $R_{plant}$ can be computed. The relative importance of $R_{soil}$ and $R_{plant}$ is an important object of study. Under wet conditions $R_{soil}$ is close to zero. Generally one can state, except for very dry soil, that $R_{plant} > R_{soil}$. Most of the plant resistance is concentrated in the roots, to a lesser extent in the leaves, and a minor part in the xylem vessels. Some researchers have concluded that $R_{plant}$ increases with progressive drying of the soil and decreases when the transpiration rate is higher (**Figure 3**). It must be emphasized that this is a controversial issue, as other researchers think that $R_{plant}$ depends neither on soil water content nor on transpiration rate.

The uptake of water and nutrients by plant roots is not always a passive process, it can be actively regulated by the plant. Water often discharges from the cut surface of active roots even if they are growing in unsaturated soil. If a manometer is attached to such a root, the pressure is usually high (0.1–0.2 MPa (1–2 bar)), sometimes even more. This root pressure arises as a result of osmotic effects. The root surface can be considered as a semipermeable membrane: permeable for water in two directions while solutes are held back inside the root. Accumulation of solutes in the root xylem causes a very low osmotic potential. As a result water flows into the roots and a positive root pressure develops in the root xylem, resulting in the discharge from cut surfaces of active roots. Such a root pressure can be high enough to carry water up to a height of 10–20 m. Therefore it has been commonly thought that this active process is the only mechanism that supplies water to the tops of trees. However, if there is a demand for water by the atmosphere, the root pressure is soon unable to satisfy it. The driving force is then no longer the pumping action of the roots, but the sucking action of the leaves, after they have lost water to the atmosphere. This sucking action is more powerful than the force from a mechanical pump could ever produce, as it arises from the chemical potential of water in the presence of solutes. The positive pressure in the xylem vessels of the root then changes to a negative pressure, which may be lower than −1 MPa (−10 bar). The result is that water flows by a passive process from the soil to the xylem vessels in the vascular cylinder. In conclusion: root water uptake due to the osmotic effect plays no essential role in the water economy of plants.

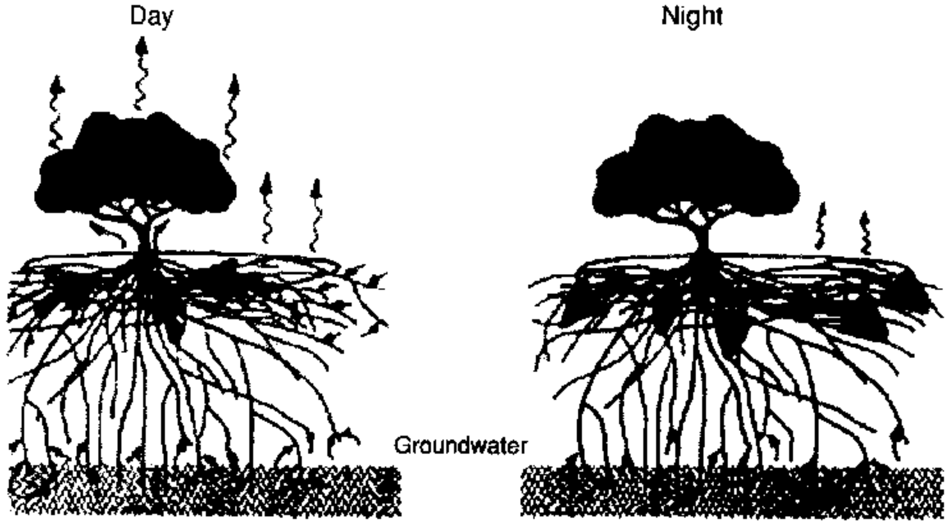


**Figure 4** Mechanism of hydraulic lift by plants during night and subsequent uptake during daytime. (Reproduced from Dawson TE (1977) Determining water use by trees and forests from isotopic, energy balance and transpiration analysis: the roles of tree size and hydraulic lift. *Tree Physiology* 16: 263–272.)

In tall trees the xylem vessels are so strong that they do not collapse when the internal equivalent pressure (a potential) is appreciably lower than the external pressure: possibly more than 1.5 MPa (15 bar!) below the atmospheric pressure of 0.1 MPa (1 bar). Water has a high tensile strength, so when the chains of water molecules are drawn toward the leaves, they hold together against the frictional drag imposed by the walls of the conduit. As a result of this cohesion, cavitation does not occur, not even at very low equivalent pressure. In other words, the water is prevented from boiling. The plant exploits this cohesion of water, to transport water through the xylem vessels at very low equivalent pressures. It is this cohesion that enables trees to grow higher than the critical 'pump lift height' of 10 m.

A potentially important effect that happens in the real world but that models overlook is the marked influence plant roots can have on the distribution and redistribution of soil water via the processes of 'hydraulic lift' (**Figure 4**). During the day, water is absorbed at all depths and passes into the main transpiration stream. During the night, in the case of hydraulic lift, deep-rooted herbs, grasses, shrubs, and trees take in water from deeper, moist soil layers and exude that water during the night into the drier, upper soil layers. For example, sugar maple trees can lift hydraulically 100 l of water through their root systems into the upper soil layer during one night. This water is then absorbed the next day and transpired. But neighboring plants may also utilize this source of water. The result of hydraulic lift is usually a decline in groundwater table depth as well as stream discharge, compared with vegetation systems where hydraulic lift is absent.

## Measurement of Root Water Uptake

Root water uptake can generally be derived by applying the water conservation equation to a given volume of soil. Commonly it is assumed that, in an unsaturated soil, water flows only in the vertical direction $z$. Consider a soil volume of unit cross-sectional area in the horizontal plane, bearing vegetation, with a lower boundary at $z = 0$ (for example, at the groundwater table or a level with constant pressure head) and an upper boundary at the soil surface at $z = z$ (positive upwards). This soil volume is partly explored by roots and partly without roots. The water conservation equation can then be written for every height $z$ and for a short time interval $t_2 - t_1$ as:

$$\int_0^z \bar{S}_z \mathrm{d}z = \bar{v}_0 - \int_0^z \frac{\partial \theta}{\partial t} \mathrm{d}z - \bar{v}_z$$
$$\bar{v}_{\mathrm{root}} = \bar{v}_{\mathrm{total}} - \bar{v}_{\mathrm{soil}} \qquad [2]$$

where $\bar{S}_z$ (cubic centimeter per square centimeter per day) is the time-averaged volume of water taken up by the roots per unit bulk volume of the soil in unit time and considered positive from the soil into the roots, $\bar{v}_0$ (centimeters per day) is the time-averaged volumetric flux density through the lower boundary, $\theta$ (cubic centimeters per cubic centimeter) is the volumetric soil water content and $\bar{v}_z$ (centimeters per day) the time-averaged volumetric flux density through the upper boundary. The latter can be calculated from Darcy's law as:

$$\bar{v}_z = -k(h)\left(\frac{\partial h}{\partial z} + 1\right) \qquad [3]$$

where $k(h)$ is the hydraulic conductivity of the soil (centimeters per day) and $h$ is the soil water pressure head (centimeters). For the calculation of root water uptake patterns, the $k(h)$ curve and the profiles of $\theta$ and $h$ have to be known. When the moisture-retention curve is used, it suffices to measure either $\theta$ or $h$ and infer the unknown one via the retention curve. The flux at the bottom $\bar{v}_0$ should be derived either from measurements (e.g., using a lysimeter) or by means of eqn [3]. Application of eqn [2] gives the integrated root water uptake over a given depth interval in the root zone. Differentiation of eqn [2] gives the water uptake rate by the roots at depth $z$.

An example is shown in **Figure 5**, where a calculation of water extraction by roots at depth $z$ is given for red cabbage on clay for a period of 7 days. The mean pressure head in this period is given in **Figure 5a**; the time-averaged cumulative withdrawal patterns of both crop and soil, $\bar{v}_{\text{total}}$, soil only, $\bar{v}_{\text{soil}}$, and crop only, $\bar{v}_{\text{roots}}$ are presented in **Figure 5b**, while the time-averaged extraction rate at depths $z$, $S(z)$, is given in **Figure 5c**. From **Figure 5b** and c it follows that the effective rooting depth can be found at the depth where the cumulative withdrawal function of total flow and soil flow intercept. From **Figure 5c** it can be seen that for the considered period the maximum extraction rate occurs at a depth of 0.30–0.35 m below surface.

The main conclusions of many reported root water uptake measurements on different vegetation can be elucidated with **Figure 6**, which shows at various times the root water uptake pattern of a red cabbage crop growing on sticky clay in the presence of a 90-cm-deep groundwater table. The magnitude of the root extraction rate is generally small at the top of the profile (unless the soil is wet just after rainfall or irrigation). Downward the extraction rate increases to a certain maximum and next decreases to zero at the root zone bottom. As the soil dries, the zone of maximum root

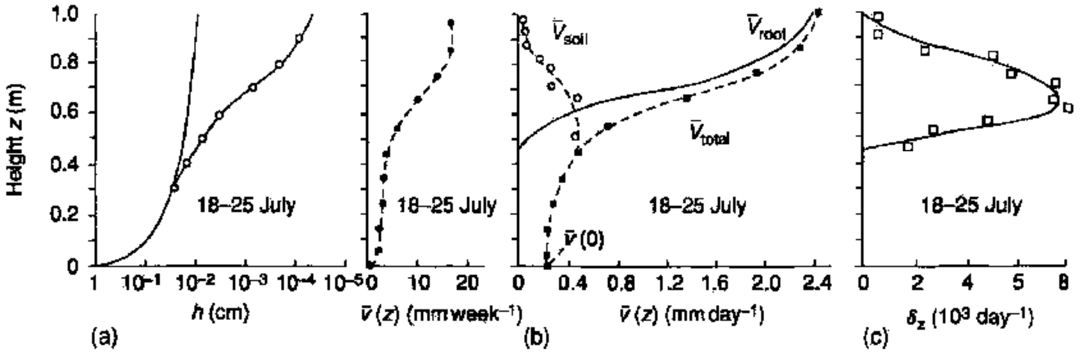


**Figure 5** Profiles of: (a), pressure head $h$; (b), time-averaged cumulative water withdrawal of both a cabbage crop and clay soil, $\bar{v}_{\text{total}}$, of soil only, $\bar{v}_{\text{soil}}$, and of roots only, $\bar{v}_{\text{root}}$; and (c) time-averaged extraction rate $S$ for red cabbage on clay for the period 18–25 July 1967. (Reproduced with permission from Feddes RA (1971) *Water, Heat and Crop Growth*. Dissertation. Wageningen, The Netherlands: Wageningen University Press.)

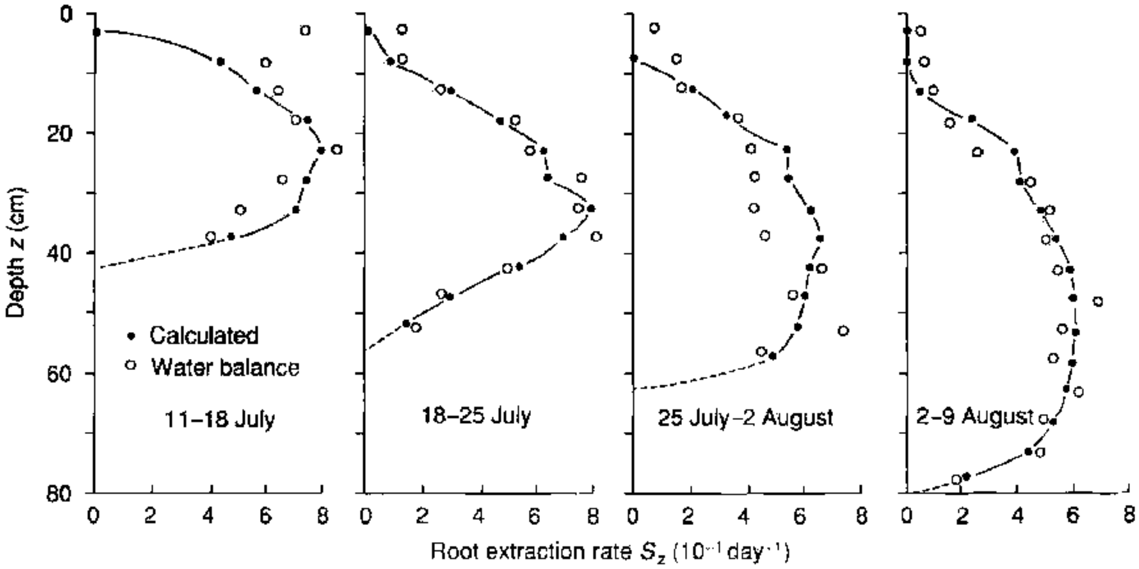


**Figure 6** Example of measured variations of root water uptake with depth and time of red cabbage grown on a clay soil with groundwater table at 90–110 cm depth and obtained from water balance studies over 4 consecutive weeks. (Reproduced with permission from Feddes RA (1971) *Water, Heat and Crop Growth*. Dissertation. Wageningen, The Netherlands: Wageningen University Press.)

water uptake moves from shallow to larger depths in dynamic correspondence with the downward progression of roots into deeper, moist soil. The maximum extraction rate appears to depend on the demand the atmosphere exerts on the plant system, on the depths to which the roots penetrate, and on the soil moisture pressure-head distribution. Later in the season, water uptake from the upper layers becomes relatively less important. Most of the water is absorbed from the zone of low tension near the water table. Thus, a relatively small part of the root system can be responsible for most of the plant water uptake.

We may conclude that root water uptake depends on a number of factors such as soil hydraulic conductivity, rooting depth, root density distribution, soil moisture pressure head, demand set by the atmosphere ('potential' transpiration) on the plant system, and the presence of a water table. This list indicates that it is not simple to model water uptake by roots, nor to generalize on the effect of a single modification at the root zone.

## Modeling Root Water Uptake in Ecologic, Hydrologic, and Atmospheric Communities

### Ecologic Communities: Detailed Radial Root Model

A detailed, plant-based approach is to consider the convergent radial flow of soil water toward and into a representative individual root, taken to be a line or narrow-tube sink. Uniform thickness and absorptive properties along its length are assumed. The root system as a whole can then be described as a set of such individual roots, assumed to be regularly spaced in the soil at definable distances that may vary within the soil profile. The flow equation is cast in cylindrical coordinates and is solved for the distribution of soil-water pressure heads, water contents, and fluxes from the root outward.

In the single-root radial model, the root is viewed as a hollow cylinder of uniform radius and infinite length, having uniform water-absorbing properties. The governing Darcy flow equation can be written in radial coordinates as:

$$\frac{\partial \theta}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left[rk(h)\frac{\partial h}{\partial r}\right] \qquad [4]$$

where $r$ is radial distance from the center of the root (centimeters). For steady-state conditions ($\partial\theta/\partial t = 0$) with water flowing from a distance, $r_2$, to a root with radius $r_1$, the solution under the assumption of constant $k$ is:

$$q_{\text{root}} = \frac{2\pi k(h_{\text{soil}} - h_{\text{root}})}{\ln(r_2/r_1)} \qquad [5]$$

where $q_{\text{root}}$ (cubic centimeters per centimeter per day) is the rate of water uptake per unit length of root, $h_{\text{soil}}$ (centimeters) is the soil-water pressure head and $h_{\text{root}}$ is the pressure head in the soil at the root surface (centimeters). Equation [5] is in hydrology known as the steady-state flux toward a well per unit length of well. For a discrete soil layer of thickness $\Delta z$ and a rooting length per unit volume of soil of $L_{\text{root}}$ (centimeters per cubic centimeter), the water uptake rate $\Delta v_{\text{root}}$ (centimeters per day) can be written as

$$\Delta v_{\text{root}} = L_{\text{root}}\Delta z q_{\text{root}} \qquad [6]$$

Within this soil layer located at depth $z$, eqn [6] can be written as:

$$\Delta v_{\text{root}} = \frac{2\pi}{\ln(r_2/r_1)} L_{\text{root}} k(h_{\text{soil}} - h_{\text{root}} + z)\Delta z \qquad [7]$$

or similarly:

$$\Delta v_{\text{root}} = B L_{\text{root}} k(h_{\text{soil}} - h_{\text{root}} + z)\Delta z \qquad [8]$$

where $B = 2\pi/\ln(r_2/r_1)$ and represents a dimensionless geometry and root distribution factor.

Due to root shrinkage in a drying soil, the surface area of roots in contact with soil decreases and the root membrane resistance increases, and thus a kind of 'contact' resistance may develop. This resistance can be taken into account by assuming that root water uptake is proportional to the relative saturation, $\theta/\theta_{\text{sat}}$, of the soil. Hence:

$$\Delta v_{\text{root}} = B L_{\text{root}} k \frac{\theta}{\theta_{\text{sat}}}(h_{\text{soil}} - h_{\text{root}} + z)\Delta z \qquad [9]$$

Does the concept of contact resistance also hold for heavier-textured soils? **Figure 7** shows computed hydraulic conductivities of a soil root system according to eqn [9] and conductivities of the bulk soil. At low soil-moisture contents, both conductivities are comparable, but at higher moisture contents the hydraulic conductivity of the soil root system becomes not higher than $10^{-6}$ cm per day, far less than the hydraulic conductivity of the bulk soil. Hence, the use of soil hydraulic conductivity is not very realistic to describe root water uptake. It seems better to take an 'effective conductivity,' which represents the conductivity of the soil–root interface and the root tissue. Moreover, in modeling studies, $h_{\text{root}}$ is often taken to be constant over the root zone, while water uptake is assumed to be potential until $h_{\text{root}} = -16\,000$ cm (wilting point).

In addition to an expression such as eqn [9], a second, similar type of equation can describe the flux over the root surface into the root using the difference between $h_{\text{root}}$ and the pressure head inside the root,

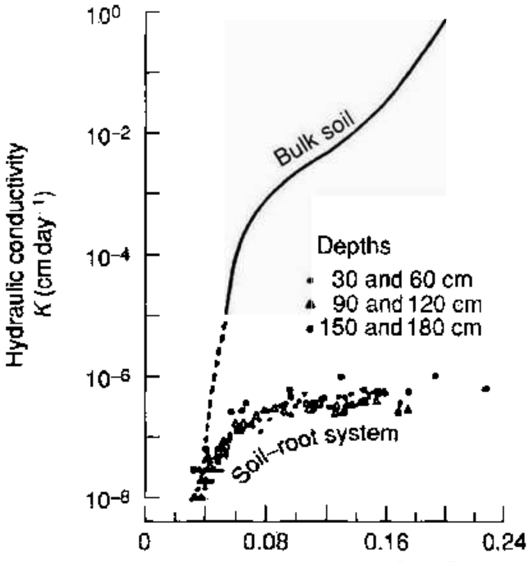


**Figure 7** A comparison of the hydraulic conductivity of a soil–cotton root system obtained at various depths with the soil hydraulic conductivity (*K*) for various soil moisture contents. (Reproduced from Taylor HM and Klepper B (1975) Water uptake by cotton root systems: an examination of assumptions in the single root model. *Soil Science* 120: 57–67.)

$h_{\text{plant}}$ (assumed to be uniform over the whole root system), and the hydraulic conductivity of the root. By equating the flux densities toward and into the root, $h_{\text{root}}$ and $h_{\text{plant}}$ can be derived by a number of iterations.

### Hydrologic Communities: General Root System Model

The second, more hydrologically oriented approach is to regard the root system as a diffuse sink that penetrates each layer of soil uniformly, though not necessarily with a constant strength throughout the root zone. Root water uptake can then be represented as a sink term that is added to the vertical water flow equation through the soil. This macroscopic way of solving the root water uptake problem is to combine the continuity equation of water flow with a sink term representing water extraction by plant roots:

$$\frac{\partial\theta}{\partial t} = -\frac{\partial v}{\partial z} - S \qquad [10]$$

where $\theta$ (cubic centimeters per cubic centimeter) is the soil water content, $t$ (days) is time, $v_z$ (centimeters per day) is the Darcian soil water flux density (eqn [3]) taken positively upward, and $S$ (cubic centimeters per cubic centimeter per day) is the root water uptake rate, which depends on depth, time, soil-water pressure head, and root density.

Combination of eqns [3] and [10] results in the physics-based soil water flow equation, known as the Richard equation:

$$\frac{\partial\theta}{\partial t} = C(h)\frac{\partial h}{\partial t} = \frac{\partial\left[k(h)\left(\frac{\partial h}{\partial z} - 1\right)\right]}{\partial z} - S(z) \qquad [11]$$

where C is the differential water capacity (d$\theta$/d$h$) (per centimeter).

Under optimal moisture conditions, the maximum possible root water extraction rate $S_p(z)$, integrated over the rooting depth, is equal to the potential transpiration rate, $T_p$ (centimeters per day), which is governed by atmospheric conditions. In many applications a homogeneous root distribution over the rooting depth is assumed, which gives:

$$S_p = \frac{T_p}{|z_{\text{root}}|} \qquad [12]$$

where $T_p$ is the potential transpiration rate (centimeters per day) and $|z_{\text{root}}|$ is the root zone depth (centimeters). The potential root water extraction rate at a certain depth, $S_p(z)$ (per day), may for nonhomogeneous root distributions be determined by the root length density $L_{\text{root}}(z)$ (centimeters per cubic centimeter) as fraction of the total root length density over the rooting depth $|z_{\text{root}}|$ (centimeters), according to:

$$S_p(z) = \frac{L_{\text{root}}(z)}{\int_{-z_{\text{root}}}^{0} L_{\text{root}}(z)\partial z} T_p \qquad [13]$$

So far root water uptake under optimal soil-water conditions, $S_p$, has been considered. Under nonoptimal conditions, i.e., too dry, too wet, or too saline, a reduction is applied. A much-used reduction function for water stress is shown in **Figure 8**. Water uptake greater than $h_1$ (oxygen deficiency) and less than $h_4$ (wilting point) is set equal to zero; between $h_2$ and $h_3$ (reduction point), water uptake is maximal. Between $h_1$ and $h_2$, a linear variation is assumed, and between $h_3$ and $h_4$ either a linear (**Figure 8**) or hyperbolic variation is assumed. The value of $h_3$ depends on the demand of the atmosphere and thus varies with $T_p$.

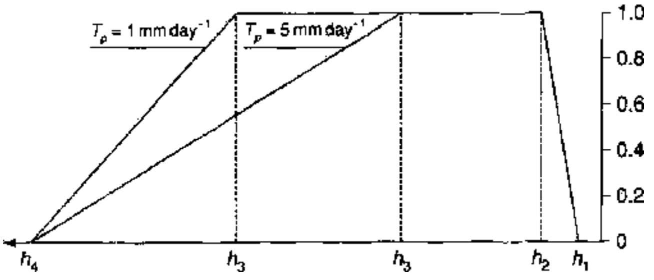


**Figure 8** Dimensionless sink term variable $\alpha$ as a function of the soil-water pressure head *h*. (Source: Feddes RA, Kowalik PJ, and Zaradny H (1978) *Simulation of Field Water Use and Crop Yield*. Simulation Monographs. Wageningen, The Netherlands: Pudoc.)

For an indication of these $h_1$–$h_4$ values for five crops, see Table 1.

For many crops the reduction in crop yield due to salinity can be linearly related to the electrical conductivity of the soil solution. Crops tolerate an increase in the soil salinity up to a threshold value above which relative yields $Y/Y_m$ and thus relative transpiration $T/T_m$ show an approximately linear decrease with increasing salt concentration or electrical conductivity $EC_e$ (**Figure 9**).

In order to simplify parameter calibration and take advantage of existing experimental data, we may assume the water and salinity stress to be multiplicative. This means that the actual root water flux density, $S(z)$ (per day), is calculated from:

$$S(z) = \alpha_{rw}(z)\alpha_{rs}(z)S_p(z) \qquad [14]$$

where $\alpha_{rw}$ (−) and $\alpha_{rs}$ (−) are the reduction factors due to water and salinity stresses respectively. Integration of $S_a(z)$ over the rooting depth then yields the actual transpiration rate $T_a$.

Numerical solution of eqn [11] requires knowledge of the soil hydraulic functions of which various databanks exist in literature for all kind of soil textures. Also data are needed for the initial moisture profile and during the simulation period of the top and bottom boundary conditions of the flow system. At the top the vegetation plays a dominant role in the partitioning of the various fluxes. Hence one needs in principle a coupling of the soil water balance model with a daily vegetation growth model. Only in this way can a proper prediction of vegetation development and growth in dependency of the actual prevailing soil water conditions be obtained, thus assuring proper feedback.

**Table 1** Critical pressure head values of the sink-term function depicted in **Figure 8** for some main crops

| *Crop* | $h_1$ | $h_2$ | $h_{3,high}$ | $h_{3,low}$ | $h_4$ |
|---|---|---|---|---|---|
| Potatoes | −10 | −25 | −320 | −600 | −16 000 |
| Sugar beet | −10 | −25 | −320 | −600 | −16 000 |
| Wheat | 0 | −1 | −500 | −900 | −16 000 |
| Pasture | −10 | −25 | −200 | −800 | −8000 |
| Corn | −15 | −30 | −325 | −600 | −8000 |

Reproduced with permission from Wesseling JG (1991) Meerjarige simulaties van grondwateronttrekking voor verschillende bodemprofielen, grondwatertrappen en gewassen met het model SWATRE (in Dutch). Report 152. Wageningen, The Netherlands: Winand Staring Centre Press.

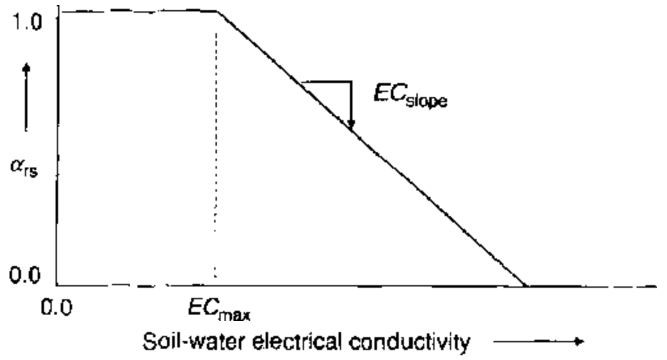


**Figure 9** Reduction coefficient for root water uptake, $\alpha_{rs}$, as a function of soil-water electrical conductivity, EC. (Source: Maas EV and Hoffman GJ (1977) Crop salt tolerance – current assessment. *Journal of the Irrigation and Drainage Division* 103: 115–134.)

## Atmospheric Communities: Large-Scale Root System Model

In global circulation models (GCM), land surface parameterizations are often based on the concept of one big leaf per grid element. However, at the resolvable scale of GCM models, land surfaces are very heterogeneous. On the other hand, various soil–vegetation–atmosphere-transfer (SVAT) schemes have been developed for use in GCM and numerical weather-prediction models. Their weakest component, however, remains their link with the lower boundary. SVAT models, which sometimes may be overparameterized, use a variety of different methods to represent the relationship between roots, soil moisture, and transpiration. Moreover SVAT parameters are generally highly variable in space and difficult to measure. Different SVAT models driven by the same meteorological forcing of air temperature, humidity, wind speed, incoming solar radiation, long-wave radiation, and rainfall can produce remarkably different surface energy and water balances. The question in this context is therefore: What is the role of roots?

With agrohydrologic models, the effect of root distribution on the course of actual transpiration in time can be evaluated. Take, for example, a grass vegetation (covering the soil completely), with a rooting depth of 80 cm growing on a loamy sand containing 10% clay (less than 2 $\mu$m). At the bottom, free drainage prevails; and initially throughout the profile the soil-water pressure head $h = -200$ cm, implying a rather wet soil. If at the soil surface a potential transpiration rate $T_p = 4$ mm day$^{-1}$ is applied for two different relative root density distributions: root 1, where most of the roots are located in the topsoil, and root 2, with a uniform root distribution (**Figure 10a**), the results of this simulation are as shown in **Figure 10b**. The moisture content of the top soil layer decreases more rapidly than the moisture content of deeper soil layers. Hence root 1 produces an earlier onset of moisture stress than root 2, after 30 days showing an actual transpiration rate that is about half that of root 2. This clearly demonstrates that roots influence the behavior of a land surface model, the role of

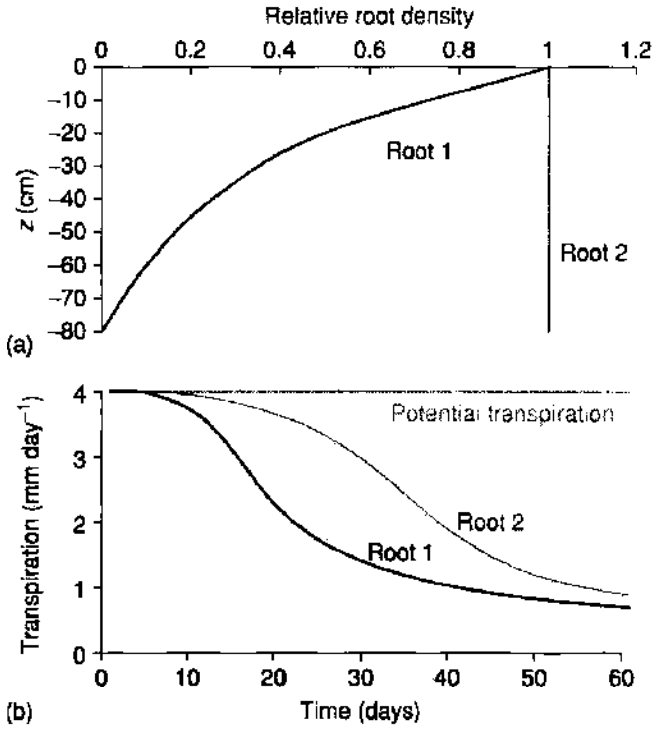


**Figure 10** (a) Two different relative root density distribution functions adopted for grass, with a rooting depth of 80 cm, growing on loamy sand; (b) simulated actual transpiration rate in time for the two root-distribution functions, taking $T_p = 4$ mm per day as the upper boundary condition.

roots being particularly important when soil moisture limits evapotranspiration.

*See also:* **Crop Water Requirements**; **Desertification**; **Irrigation**: Environmental Effects; **Plant–Water Relations**; **Root Architecture and Growth**; **Soil–Plant–Atmosphere Continuum**

## Further Reading

Dane JH and Topp GC (eds) (2002) *Methods of Soil Analysis, part 4, Physical Methods.* Soil Science Society of America Book Series, No. 5. Madison, WI: Soil Science Society of America.

Feddes RA (1981) Water use models for assessing root zone modification. In: Arkin GF and Taylor HM (eds) *Modifying the Root Environment to Reduce Crop Stress*, pp. 347–390. ASAE Monograph 4. St. Joseph, MI: ASAE.

Feddes RA, Kowalik PJ, and Zaradny H (1978) *Simulation of Field Water Use and Crop Yield. Simulation Monographs.* Wageningen, The Netherlands: Pudoc.

Gardner WR (1960) Dynamic aspects of water availability to plants. *Soil Science* 89: 63–73.

Homaee M (1999) *Root Water Uptake Under Non-uniform Transient Salinity and Water Stress.* PhD Thesis. Wageningen, The Netherlands: Wageningen University Press.

Molz FJ (1981) Models of water transport in the soil-plant system: a review. *Water Resources Research* 17: 1245–1260.

Raats PAC (1974) Steady flows of water and salt in uniform soil profiles with plant roots. *Soil Science Society of America Proceedings* 38: 717–722.

Scott HD (2000) *Soil Physics. Agricultural and Environmental Applications.* Ames, IA: Iowa State University Press.

Tanji KK (1990) *Agricultural Salinity Assessment and Management.* Manuals and Reports, No. 71. New York: American Society of Civil Engineers.

van Dam JC, Huygen J, Wesseling JG *et al.* (1997) *Theory of SWAP, Version 2.0. Simulation of Water Flow, Solute Transport and Plant Growth in the Soil–Water–Atmosphere–Plant Environment.* Report 71, Department of Water Resources. Wageningen, The Netherlands: Wageningen University Press.

Warrick AW (ed.) (2002) *Soil Physics Companion.* New York: CRC Press.

Willigen P de and van Noordwijk M (1995) A model for interactions between water and nutrient uptake. In: Kabat P, Marshall B, van den Broek BJ, Vos J, and van Keulen H (eds) *Modelling and Parameterization of the Soil–Plant–Atmosphere System: A Comparison of Potato Growth Models*, pp. 135–153. Wageningen, The Netherlands: Wageningen Press.

# PLANT–WATER RELATIONS

**C Gimenez**, University of Cordoba, Cordoba, Spain
**M Gallardo and R B Thompson**, University of Almeria, Almeria, Spain

## Introduction

Water constitutes the largest single chemical component of plants, yet the volume within the plant is very small in relation to the total volume transpired. Adequate plant water status requires that root uptake of soil water satisfies the atmospheric demand (i.e., transpiration requirements). When water uptake is insufficient, a plant water deficit can develop. Various parameters are used to measure plant water status, the most common being water potential. Numerous plant processes are affected by declining plant water status, the most sensitive being cell expansion, which affects new growth. Stomatal closure is another sensitive process. Plants' responses to declining soil water content are mediated by lowered leaf water potential and/or by chemical agents produced in roots. Some parameters of plant water status can be used for irrigation-scheduling purposes.

## Water in Plants

In herbaceous plants, water normally constitutes more than 90% of fresh weight, although in rare cases it can be less than 70%. In woody plants, over 50% of fresh weight consists of water. Of the total water content of plants, 60–90% is located within cells, the rest (10–40% of total water) is mainly in cell walls. The water in cell walls forms a continuum with specialized transport cells throughout the plant.

The total volume of water within a plant is very small in relation to the total volume of water transpired during its lifetime. Even on a daily basis, the volume of water within plants is insufficient to buffer appreciably daily transpiration requirements on a warm, sunny day. The very large amounts of water transpired by plants, in relation to that retained within plant tissue, can be viewed as the 'cost' that plants incur as a consequence of stomatal opening to allow $CO_2$ absorption for photosynthesis.

A primary function of water contained in cells is the maintenance of cell and tissue turgor. Cell turgor is essential for cell enlargement and therefore for optimal plant growth. Other primary functions are the transport of solutes and participation in metabolic activities. Because of its high dielectric constant, water acts as a solvent for many mineral and organic solutes, enabling their transport within cells and throughout the plant. Also, water is directly involved in chemical reactions in cells such as $CO_2$ reduction in photosynthesis. Cooling is another primary function of water in plants. Because of the high energy requirement for water vaporization (10.5 $\mathrm{kJ\,mol^{-1}}$ at 25°C), water evaporating from leaf surfaces (during transpiration) cools the leaf, thereby avoiding excessive daytime heating from incoming solar radiation.

The water status of plants is a primary determinant of plant growth and development, and therefore of crop productivity in agricultural systems, and of plant survival in natural systems. Almost every plant physiological process is directly or indirectly affected by plant water content. For example, cell enlargement is dependent on the level of cell turgor, photosynthesis is directly inhibited by insufficient water, and stomatal control of transpiration and $CO_2$ absorption is dependent on the water status of stomatal guard cells.

The water status of plants is the sum of the interaction of various atmospheric, plant, and soil factors. The availability of soil water, the atmospheric demand (determined by radiation, humidity, temperature, wind), the capacities of the root system to absorb water and of the plant to transport absorbed water to transpiring leaves, and stomatal responses for regulating transpiration, can all appreciably influence plant water status.

Plant water status is commonly characterized by its water potential ($\Psi$). Water potential is a measure of the free energy status of water, which, because of its applicability to each component of the soil–plant–atmosphere system, enables water movement between these components to be considered. It theoretically represents the work involved in moving one mole of water from a selected point within the plant (or soil) to a reference point of pure water at the same temperature and at atmospheric pressure. $\Psi$ varies from zero at the reference point to negative values within the plant and soil. It is normally measured in units of pressure, with megapascals (MPa) being most commonly used.

Leaves are the plant organs where most of the exchange of $CO_2$ and $H_2O$ between the plant and the atmosphere occurs. The pathway for the inward diffusion of $CO_2$ is much the same as that for the outward diffusion of $H_2O$ vapor. The outward diffusion of $H_2O$ vapor from the saturated surfaces within the plant to the drier atmosphere follows a gradient of the partial pressure of $H_2O$ vapor. To maximize $CO_2$ fixation by photosynthesis, stomata must remain open for as long as possible during daylight periods. This also maximizes the period of water loss by transpiration.

The water status of leaves (considered as $\Psi$) is the balance between the water lost to the atmosphere by transpiration ($T$) and the water absorbed by the plant from soil, which is a function of soil water potential ($\psi_{\text{soil}}$) and the combined resistance to water movement within the roots and shoots ($r$). These relationships are described by the equation:

$$\Psi = \psi_{\text{soil}} - T\,r \qquad [1]$$

Even in saturated soils (where $\psi_{\text{soil}} = 0$), $\Psi$ is negative when transpiration occurs. During daylight periods, absorption of water lags behind transpiration owing mainly to the high resistance to water flow from soil into root xylem tissue. As atmospheric evaporative demand increases during the morning, transpiration increases, which lowers the water potential of cells from which water is evaporating. Within the plant, water then moves from nonevaporating parenchyma cells of leaves, which have a higher $\Psi$, toward the evaporating cells, establishing a $\Psi$ gradient. This gradient is transmitted throughout the plant–soil system, enabling continuous water movement. In the afternoon, transpiration decreases on account of reduced atmospheric evaporative demand. However, water uptake by roots continues until parenchyma cells fully rehydrate, and their $\Psi$ equals soil $\Psi$, which usually occurs during the night. At this stage, plant and soil water are in equilibrium, and absorption by roots ceases. (*See* **Soil–Plant–Atmosphere Continuum.**) In some species under certain climatic conditions (high nighttime vapor pressure deficit, wind), transpiration can occur at night; when it does, it is generally relatively small compared to daytime transpiration. However, it can be sufficient to prevent nighttime equilibration of plant and soil water potentials.

## Plant Water Status Indicators

### Direct Indicators of Plant Water Status

**Relative water content** The water content of plant tissue is the fraction of the total fresh tissue weight that is water. It can be considered either on a fresh- or dry-weight basis. As there can be considerable variation between different plant tissues and organs in the maximum water content, the normalized parameter relative water content (RWC) is used. RWC is the water content of plant tissue or a plant organ relative to its maximum water content when it is fully hydrated (i.e., saturated). RWC provides a measure of the degree of hydration of tissue, which is considered to have an important controlling role on numerous plant functions.

Relative water content is expressed as:

$$\text{RWC} = (\text{FW} - \text{DW})/(\text{SFW} - \text{DW}) \qquad [2]$$

where FW is the fresh weight (grams), DW the dry weight (grams), and SFW is the saturated fresh weight (grams).

**Water potential and components** The most commonly used parameter to characterize plant water status is water potential ($\Psi$). Total water potential ($\Psi$) has four components: the osmotic potential ($\psi_s$), pressure potential ($\psi_p$), matric potential ($\psi_m$), and gravitational potential ($\psi_g$):

$$\Psi = \psi_s + \psi_p + \psi_m + \psi_g \qquad [3]$$

Osmotic potential ($\psi_s$) results from dissolved solutes in cell sap and is proportional to solute concentration and inversely proportional to cell water volume. $\psi_s$ in plants is always negative and decreases as solutes concentrate during plant dehydration. Pressure potential is a measure of tissue turgor produced by the diffusion of water into the protoplast of cells enclosed by largely inelastic cell walls. Matric potential ($\psi_m$) arises from the action, on water, of electrostatic forces of attraction associated with cell wall and colloidal surfaces, and of capillary forces associated with narrow transport vessels. In plants it is considered to be negligible. Gravitational potential ($\psi_g$) results from gravitational forces acting on the water within plants. On account of the gradient of $\psi_g$ with height of 0.01 MPa $m^{-1}$, $\psi_g$ is normally negligible compared with the other components of plant $\Psi$; it is only considered to be significant in very tall trees. In most situations, total plant water potential is considered to be the sum of the pressure potential ($\psi_p$) and osmotic potential ($\psi_s$).

As both pressure and osmotic potential are dependent on tissue water content, there are relationships between RWC and $\psi_s$ and $\psi_p$, and consequently between RWC and $\Psi$. These relationships are schematically represented in **Figure** 1. In a fully hydrated plant, RWC equals 1, $\psi_p$ is positive, $\psi_s$ is negative, and $\Psi$ (the sum of $\psi_p$ and $\psi_s$) equals zero. As RWC progressively declines, both $\psi_p$ and $\psi_s$ decline, and consequently $\Psi$ becomes more negative. When RWC has declined sufficiently for $\psi_p$ to equal zero, the plant loses turgor and wilts.

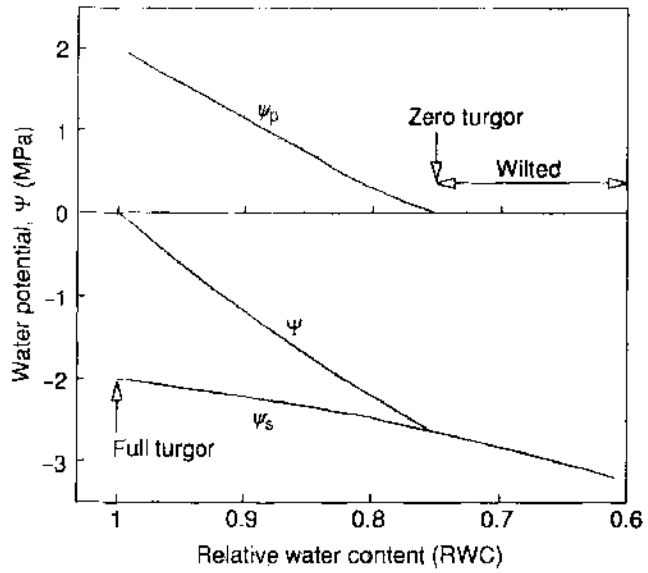


**Figure 1** Höffler–Thoday diagram illustrating the relationships between total water potential (Ψ), pressure potential ($\psi_p$), osmotic potential ($\psi_s$), and relative water content (RWC) as a cell or tissue loses water from a fully turgid state. Full and zero turgor, and wilting are indicated by arrows. (Adapted from Jones HG (1992) *Plants and Microclimate*. Cambridge, UK: Cambridge University Press, with permission.)

Several methods are used to measure the total water potential of plant tissue. Thermocouple psychrometry and hygrometry are the most accurate methods, but both require very stable environmental conditions. Consequently, whilst they can be used under controlled conditions, they are very difficult to use in field studies. The pressure-chamber method is suitable for routine and rapid field measurement.

Pressure potential ($\psi_p$) can only be measured directly by the pressure microprobe, which is inserted into cell protoplasm. This technique is limited to relatively large cells and is generally restricted to laboratory measurement. In practice, the most commonly used procedure is to measure both Ψ and $\psi_s$, and then to calculate $\psi_p$ as the difference between Ψ and $\psi_s$. Osmotic potential ($\psi_s$) is commonly measured by either thermocouple psychrometry or hygrometry after previously freezing and thawing the sample to break cell membranes and reduce $\psi_p$ to zero; normally a correction is made for the dilution by apoplastic water. Another method for measuring $\psi_s$ is by determining a moisture-release curve (also called pressure–volume curve) with a pressure chamber.

### Indirect Indicators of Plant Water Status

**Stomatal conductance** Stomatal conductance ($g_l$) is a measure of the degree of stomatal opening and can be used as an indicator of plant water status. Stomatal conductance is related to leaf Ψ by feedback processes. Reductions in $g_l$ prevent further decreases in Ψ by reducing transpiration; also, reductions in Ψ can induce stomatal closure, resulting in lowered $g_l$. Stomatal responses are discussed more fully in the section Stomatal Responses, below. Stomatal conductance can be measured with both dynamic and steady-state diffusion porometers.

**Stem diameter variations** Transpiring plants and trees undergo diurnal variation in stem or trunk diameter. During the morning, when plants are most vigorously transpiring, some water from stem tissue, mostly from phloem tissue, is incorporated into the transpiration stream, producing a measurable reduction in stem diameter. During the afternoon, as transpiration slows, this tissue begins to rehydrate, with a corresponding increase in stem diameter. Rehydration continues throughout the night as some water uptake continues from soil, when transpiration has usually completely ceased. Commonly, stem diameter has a maximum daily value just before dawn, and a minimum daily value in the early afternoon. The magnitude of the diurnal contraction is dependent on atmospheric evaporative demand and plant water status. Under given atmospheric conditions, the magnitude of daily stem contraction increases with increasing plant water deficit. Consequently, stem contraction can be used as an indicator of plant water status. In some fruit-tree and vegetable species, variations in stem diameter have been shown to be very sensitive measures of plant water status.

In young herbaceous plants and young trees, well-watered healthy plants show a trend of increasing daily maximum stem-diameter values, which is an indication of growth. In such plants, a reduction in the rate of growth is apparent in a change in the slope of daily maximum stem-diameter values, which may be due to a lack of water or also to inadequate crop management or changed environmental conditions. In more mature plants, daily maximum stem-diameter values tend to be more constant. Sensors incorporating sensitive pressure transducers, known as linear variable differential transducers (LVDT), connected to data loggers enable continuous monitoring of stem diameter.

**Sap flow** Sap-flow sensors measure transpiration flow as the ascent of sap within xylem tissue; measurements can be made in stems, trunks, branches, or tillers. Given that transpiration is sensitive to plant water status, with the effect being mediated by stomatal opening (see Stomatal Responses, below), sap flow can be used as an indicator of plant water status.

Sap flow rates can be out of phase with transpiration because of capacitance effects in stems or branches arising from the storage of water. Continuous data recording enables the time course of transpiration to be followed.

Two different techniques are used to measure sap flow; both use heat as a tracer. One is the stem (or trunk) sector heat-balance method, in which a section of the entire stem circumference is electrically heated, and the axial and radial heat-loss measured. The mass flow rate of sap is calculated as a function of the heat dissipated by the ascending sap. The other method is the heat-pulse method in which the heater and temperature sensor probes are placed inside the trunk in a radial direction. The sap velocity is calculated as a function of the time required by the flowing sap to transport heat to a particular location.

**Leaf or canopy temperature** When plant water status is adequate, canopy temperature (i.e., average temperature of total leaf surface) is appreciably less than the air temperature, through the cooling effect of evaporation. As stomata partially close in response to increasing plant water stress, the energy balance of the plant is altered; less heat is dissipated through the evaporation of water, and consequently canopy temperature increases. With increasing plant water stress, the difference between canopy and air temperature lessens and, under conditions of severe water stress, may become positive. Canopy temperature is most commonly measured with infrared thermometers.

### Other Methods

As plant water status affects overall plant growth, measurable parameters such as fruit growth and expansive growth of leaves or stems have been used as indicators of plant water status. Visual characters such as leaf rolling, color change, and visible wilting have also been used. However, visual characteristics are normally an indication that the water stress is severe.

## Water Deficits in Plants

Traditionally, 'plant water deficit' or 'plant water stress' has been defined as being when plant water status is reduced sufficiently to affect normal plant functioning (e.g., plant growth, stomatal conductance, rate of photosynthesis). It is not possible to define water deficit in terms of absolute values on account of the complex interaction of atmospheric, plant, and soil factors involved. Also, preconditioning effects from the prior history of the plant can modify plant response at a given water potential. Consequently, a commonly used approach is to express the measured values of the stressed plants in relation to those of well-irrigated plants which have experienced the same environmental conditions and previous general history.

Plant water deficits occur when any plant process is affected by:

1. Limited water absorption by roots because of dry, cool, or poorly aerated soil;
2. High evaporative demand, on account of low relative humidity, high air temperature, high wind speed, high radiation, or combinations of the four;
3. A combination of limited water absorption and high evaporative demand.

Generally, plant water deficits can be considered as being induced by either insufficient available soil water, or a high atmospheric evaporative demand. Plant water deficits induced by lack of soil water may continue for days, possibly weeks, until they are either alleviated by rain or irrigation or the plant dies. Those that are solely due to high atmospheric evaporative demand ('midday' water deficits) are much more transient, occurring for no more than several hours in the middle part of the day.

When not alleviated, water deficits induced by lack of soil water become progressively more intense. The time course of leaf and soil $\Psi$ during a theoretical drying cycle of several days is shown in **Figure 2**. As soil continues to dry and soil $\Psi$ to decrease, there is increasingly less available soil water for root uptake.

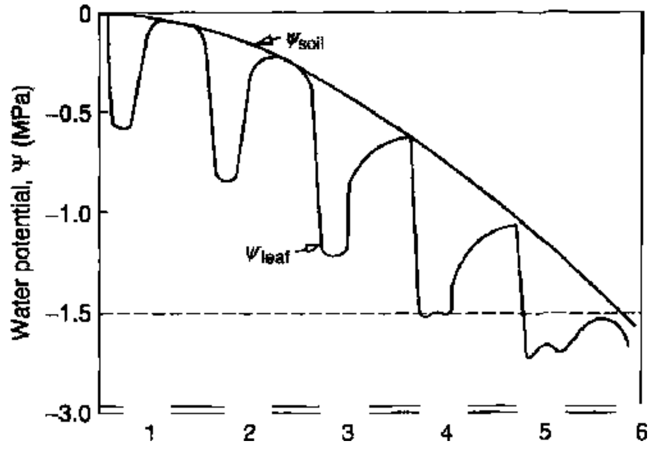
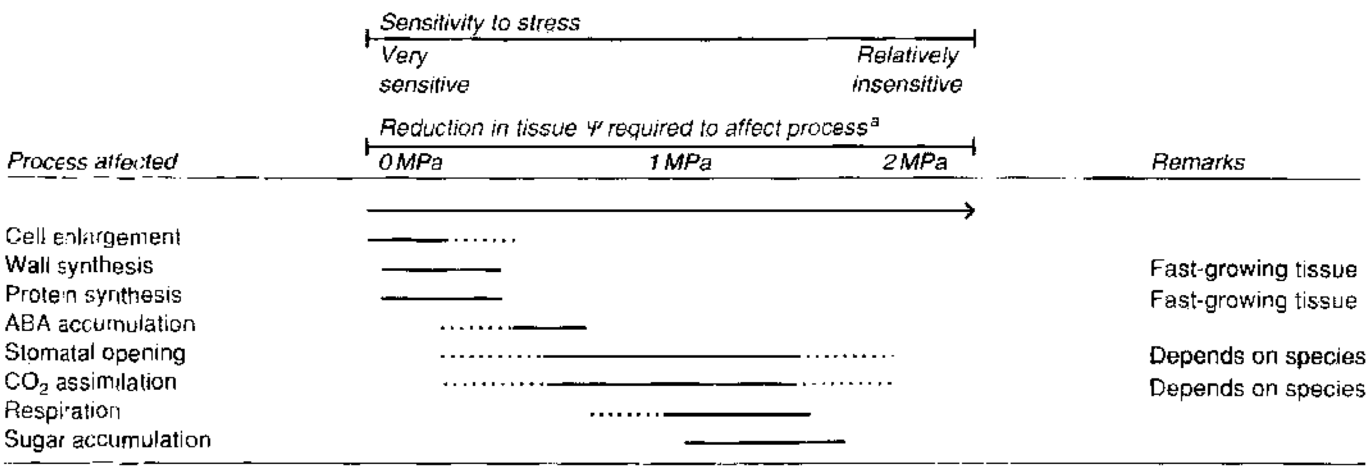


**Figure 2** Time course of soil water potential ($\psi_{soil}$) and leaf water potential ($\psi_{leaf}$) during a drying cycle. Dashed line represents soil water potential at which permanent wilting of the plant occurs. (Adapted from Kramer PJ (1983) *Water Relations of Plants.* New York: Academic Press, with permission.)

**Table 1** Plant process affected by water stress and their sensitivity. Length of the horizontal lines represent the range of stress levels within which a process becomes first affected. Dashed lines signify deductions based on more tenuous data



| Process affected | Sensitivity to stress: Very sensitive → Relatively insensitive<br>Reduction in tissue Ψ required to affect process[a]: 0 MPa, 1 MPa, 2 MPa | Remarks |
|---|---|---|
| Cell enlargement | | |
| Wall synthesis | | Fast-growing tissue |
| Protein synthesis | | Fast-growing tissue |
| ABA accumulation | | |
| Stomatal opening | | Depends on species |
| $CO_2$ assimilation | | Depends on species |
| Respiration | | |
| Sugar accumulation | | |

[a] With Ψ of well-watered plants under mild evaporative demand as the reference point.

Source: Hsiao TC (1973) Plant responses to water stress. *Annual Review of Plant Physiology* 24: 519–570.

This results in a general tendency for plant Ψ to decrease over time. Eventually, the plant is unable to absorb sufficient water for plant Ψ to equilibrate with soil Ψ at night. When plants are unable to recover cell and tissue turgor, which occurs at a threshold soil-water potential value (usually approximately −1.5 MPa), permanent wilting of the plant occurs.

Midday water deficits can even occur in plants growing in moist soil or nutrient solutions. They occur when high atmospheric evaporative demand causes high midday transpiration rates that exceed the rate of water uptake by roots. Where these situations are recurrent, daytime growth is inhibited by these water deficits, and most growth occurs at night, as long as nighttime temperatures are not limiting. Midday water deficits are not restricted to hot and dry environments; they have also been observed in the humid tropics and in greenhouses. Recurrent midday water deficits can reduce yield in cropping situations. In intensive greenhouse crop-production systems, fog systems are commonly used to prevent midday water deficits.

Plant processes affected by water deficits, and an indication of their sensitivity, are shown in **Table 1**. Processes such as cell enlargement, cell wall synthesis, and protein synthesis are very sensitive, being affected by relatively small reductions in Ψ. Other processes, such as stomatal opening and $CO_2$ assimilation, generally require larger reductions in Ψ. As plant water deficits intensify, an increasing number of plant processes are affected (**Table 1**), and there can be complex interactions between them. Two of these processes, expansive growth and stomatal responses, are discussed more fully below.

### Expansive Growth

The plant process most sensitive to water deficit is cell enlargement, which affects expansive growth. The effects of plant water deficits on expansive growth are very important, because they result in reduced expansion of the plant assimilation surface (i.e., leaf area). The main factor controlling cell enlargement and therefore leaf expansion is turgor pressure ($\psi_p$). The relationship between $\psi_p$ and cell expansion can be represented by the Lockhart equation:

$$dV/V\,dt = E_g(\psi_p - \psi_{p,th}) \qquad [4]$$

where the relative change in cell volume ($V$), with time, is dependent on $\psi_p$ above a threshold value ($\psi_{p,th}$) below which no cell expansion occurs. $E_g$ is the gross extensibility of the cell wall (i.e., its capacity to expand). Increases in cell wall extension during growth are irreversible. Values of $\psi_{p,th}$ are normally only slightly less than $\psi_p$ values of nonstressed plants, so even small reductions in $\psi_p$ can result in reduced growth. Neither $E_g$ nor $\psi_{p,th}$ is constant: they can be influenced by previous exposure to water stress and by growing conditions. With the progressive imposition of water stress, plants can adapt by either reducing $\psi_{p,th}$, increasing $E_g$, or doing both in order to maintain growth at lower $\psi_p$.

### Stomatal Responses

When leaf water status is adequate (i.e., well-watered plants) and atmospheric evaporative demand is low, stomatal opening is determined primarily by light conditions and the low $CO_2$ partial pressure of substomatal cavities. Under these conditions, the stomata are fully open during daylight periods, maximizing assimilation of $CO_2$, thereby ensuring optimal rates of photosynthesis.

When plant water deficits develop and leaf $\Psi$ decreases, partial stomatal closure occurs to reduce water loss from the plant, with a consequent reduction in photosynthesis. In many species, $\Psi$ must decline to a threshold value before stomatal closure commences. Threshold $\Psi$ values for stomatal closure vary with species, leaf age, previous exposure to radiation, the stress history of the plant, and environmental conditions. If the plant has previously been subjected to a progressive, mild water stress, it can be preconditioned so that the threshold $\Psi$ value is lowered, thereby enabling assimilation to proceed at lower $\Psi$. Stomata can also close in response to low air humidity, even when leaf water status is adequate.

When recovering from water stress, stomata respond relatively slowly. Unlike $\Psi$, which rapidly recovers to nonstressed values once water stress is alleviated, stomatal conductance can take several hours, even days, to recover to nonstressed values.

Stomatal closure in response to plant water deficits is not solely controlled by turgor pressure of leaf cells. The process is complex, with the additional involvement of solute movement. When water deficits develop, stomatal guard cells simultaneously lose both water and cell solutes, particularly $K^+$ ions. It seems that water stress promotes the efflux of $K^+$ ions from guard cells, resulting in the loss of cell turgor, which induces stomatal closure. In the contrary situation, when water stress is relieved, synchronized changes in $K^+$ concentration and water content of guard cells have been detected prior to stomatal opening.

### Root Signals

Traditionally it has been accepted that drying soil has a hydraulic effect on plant water relations and consequently on cell expansion and leaf gas exchange. The basic theory is that a reduction in soil water potential causes a reduction in plant pressure potential, inducing reduced cell expansion and stomatal closure. In recent decades, it has become clear that stomatal closure, and also leaf expansion, are controlled by a mechanism or mechanisms additional to changes in pressure potential of leaf cells. Reductions in stomatal conductance ($g_l$) of plants, in drying soil, have been related more strongly to changes in soil water status than to leaf water status. These observations indicate that plants can 'sense' that soil in the root zone is drying and can communicate this information to the leaves by a means other than reduced leaf water status. The evidence to date suggests that a chemical signal, abscisic acid (ABA), is produced in roots in drying soil and is transported to leaves in the xylem sap, where it induces stomatal closure before reductions in leaf water potential occur. Although ABA is produced in both roots and leaves, it is accepted that root tips are the major source of ABA produced in response to drying soil. The sensitivity of stomata to ABA is mediated by the nutritional and water status of the plant and can be genetically determined. The triggering of stomatal closure by chemical root signals has been observed in many agricultural species.

Much of the research work to date on chemical root signals has been conducted on the effects of ABA on stomatal responses. However, other plant growth processes sensitive to water stress, such as leaf expansion and leaf initiation, are also influenced by chemical root signals. For example, in some species, even when leaf water potential is artificially maintained, soil drying induces reductions in leaf elongation.

The relative degree of control of hydraulic and chemical root signals over stomatal conductance and leaf expansion, in field-grown plants, is the subject of considerable research and debate. In some studies, changes in ABA concentration in xylem sap do not correlate well with changes in stomatal conductance, suggesting that there may be chemical agents other than ABA that induce stomatal closure. As yet, no other such agents have been identified. Additionally, there is controversy associated with some data from studies conducted under controlled conditions, such as the induction of ABA production by mechanical impedance to root growth in potted plants, where the nature of the experimental conditions may contribute to ABA production.

It is now accepted that plant processes such as stomatal conductance and leaf expansion that are sensitive to water stress are affected by both chemical root signals and changes in leaf water status. Currently, research is being conducted to investigate possible interactions between hydraulic and chemical signals.

## Applications of Plant–Water Relations to Irrigation Scheduling

The most common practical application of plant–water relations is to assist in irrigation management. Different plant water status indicators have been proposed for use in irrigation scheduling (IS). The advantage of plant-based measurements over alternative IS methods based on soil-moisture monitoring or estimation of evapotranspiration (ET) requirements is that the actual unit of production (the plant) is being assessed, rather than using an environmental parameter (e.g., the soil) or a mathematical estimate (e.g., of ET).

Generally, indicators of plant water status are suitable for determining the timing of irrigation (i.e., when to irrigate), but not the amount of water to apply. An additional method, for example that of the Food and Agriculture Organization of the UN (FAO) for estimating crop evapotranspiration requirements (*See* Crop Water Requirements), can be used to estimate the amounts to apply. Interpretation of data of plant water status indicators requires comparison with previously established threshold values that define the need for irrigation (e.g., a maximum allowable daily stem-diameter contraction). Plant water status data are normalized to take into account variations in water status due to evaporative demand, i.e., to distinguish effects mostly due to insufficient soil water from those mostly due to atmospheric demand. Ideally, measurements of plant water status are considered in relation to those obtained from fully irrigated plants.

In some crops such as cotton, the timing of irrigation has been determined by predawn or midday measurements of leaf water potential. To use leaf $\Psi$ for this purpose, the relationship between leaf $\Psi$ and growth is established previously, for the individual species. In some fruit-tree species, stem $\Psi$ has been used in preference to leaf $\Psi$, because variations in leaf $\Psi$, due to rapid changes in evaporative demand, are avoided. Stem $\Psi$ is measured in leaves that have been previously covered to prevent transpiration.

Until recently, plant water status measurements were generally based on discrete manual measurements (e.g., leaf water potential, stomatal conductance), which limited the amount of data that could be obtained on account of the time and labor required for each measurement. Recent developments in sensor and data technology have enabled continuous monitoring of plant water status. Data from infrared thermometers, LVDT (stem-diameter) sensors and sap-flow gauges can be recorded every few minutes, providing detailed information on the dynamics of plant water status. Automatic programmers for irrigation management are being increasingly used; these systems can be integrated with continuously measured parameters of plant water status so that crops can be automatically irrigated when the selected indicator of plant water status reaches a defined limit. Discrete measurements of plant water status cannot be used in this way.

Different methods based on temperature difference between air and canopy have been developed to schedule irrigation. The most widely used is the crop water stress index, which relates canopy ($T_c$) and air ($T_a$) temperature to the vapor pressure deficit (VPD),

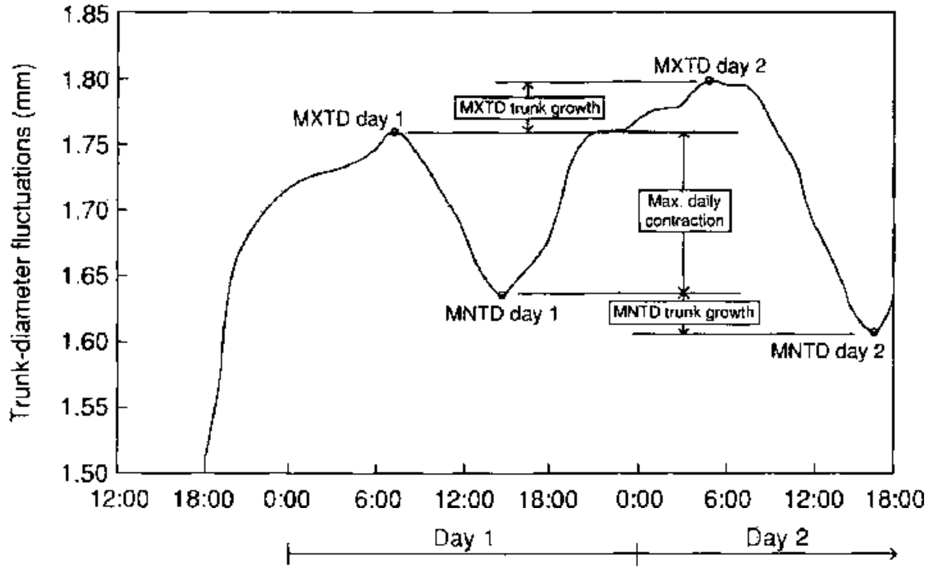


**Figure 3** Parameters that can be derived from trunk-diameter measurements, including maximum daily trunk contraction, and trunk growth expressed as daily differences in maximum and minimum daily trunk diameters (MXTD and MNTD, respectively). (Adapted from Goldhamer DA and Fereres E (2001) Irrigation scheduling protocols using continuously recorded trunk diameter measurements. *Irrigation Science* 20: 115–125, with permission.)

and compares this with $T_c - T_a$ for a well-watered crop at the same VPD.

Trunk-diameter measurements with LVDT sensors have been proposed for IS in a number of fruit-tree species. Good correlations between variations in trunk diameter and changes in tree water status were observed in these species. Applying trunk- or stem-diameter measurements for IS requires firstly the definition of parameters derived from the data, and then the development of criteria for applying the derived parameters, as discussed previously. Maximum daily contraction, the difference between daily maximum values, and difference between daily minimum values (**Figure 3**) have been proposed as the most sensitive parameters to detect water stress in some fruit trees. As of 2004, irrigation protocols are being developed for fruit trees and vegetable species that involve: (1) selecting the derived parameter most suitable for an individual species and particular growth stage, and (2) relating the derived parameters to reference values of well-watered crops and normalizing them for VPD. A consideration when using this method is the number of replicates required to account for between-plant variability.

Sap-flow sensors have also been used for irrigation scheduling in fruit trees. A method for the automatic control of irrigation has been developed, in which sap-flow sensors determine the amount of water consumed by the plant, at short time intervals. An irrigation system can be programmed to replenish this water at specified times or time intervals. Variability between plants is an issue when scaling from single plants to crops.

*See also:* **Plant–Soil–Water Relations**; **Soil–Plant–Atmosphere Continuum**; **Water-Use Efficiency**

## Further Reading

Davies WJ and Zhang J (1991) Root signals and the regulations of growth and development of plants in drying soil. *Annual Review of Plant Physiology* 42: 55–76.

Goldhamer DA, Fereres E, Mata M, Girona J, and Cohen M (1999) Sensitivity of continuous and discrete plant and soil water stress monitoring in peach trees subjected to deficit irrigation. *Journal of the American Society of Horticultural Science* 124: 437–444.

Hsiao TC (1973) Plant responses to water stress. *Annual Review of Plant Physiology* 24: 519–570.

Hsiao TC (1990) Measurements of plant water status. In: Stewart BA and Nielsen DR (eds) *Irrigation of Agricultural Crops*, pp. 243–279. Madison, WI: ASA–CSSA–SSSA.

Hsiao TC and Bradford KJ (1983) Physiological consequences of cellular water deficits. In: Taylor HM, Jordan WR, and Sinclair T (eds) *Limitations to Efficient Water Use in Crop Production*, pp. 227–265. Madison, WI: ASA–CSSA–SSSA.

Jones HG and Tardieu F (1998) Modelling water relations of horticultural crops. *Scientia Horticulturae* 74: 21–46.

Kramer PJ and Boyer JS (1995) *Water Relations of Plants*. San Diego, CA: Academic Press.

Smith DM and Allen SJ (1996) Measurement of sap flow in plant stems. *Journal of Experimental Botany* 47: 1833–1844.

Turner NC (1997) Further progress in crop water relations. *Advances in Agronomy* 58: 293–338.

## Poiseuille's Law *See* Hydrodynamics in Soils

# POLAR SOILS

**J C F Tedrow**, Formerly with Rutgers University, New Brunswick, NJ, USA

## Introduction

Soil science, traditionally allied with food and fiber production, has been viewed primarily through the eyes of the farmer. It is therefore understandable why soils of the polar regions, with negligible agricultural potential, were the last to receive serious scientific attention. Early investigators of polar soils showed all land in northern Siberia as being mantled with Tundra soil. The tundra is described as developed on a treeless plain of the north where conditions are swampy and the land is underlain by permanently frozen ground (permafrost). Investigators tend to use the 50°F July isotherm as the approximate southern limit of the arctic, a term used somewhat interchangeably with the word 'tundra.' Subsequent investigators of the arctic lands have convincingly shown that not only are there a variety of soils present locally, depending upon relief, substrate, and physiographic setting, but also there are major regional differences.

## Climate

Climate of the northern polar regions is characterized by long, cold winters and short, cool summers. Mean monthly temperature for the warmest month in the vicinity of the tree line ranges from approximately 50 to 54°F but, in the northern extremities of ice-free land such as northern Greenland, Svalbard, and Franz Josef Land, the temperatures of warmest month decrease to 32–34°F. Precipitation values approximate 25 cm annually in the main polar belt but decrease somewhat northward.

Vital activity in polar soils is limited to 2–4 months per year, depending upon region, site, relief, vegetation, snow cover, and related factors. Annual snowfall in the polar regions generally is 30–60 cm, with much redistribution by wind action.

In the far north, the ice-free sectors have a desert-like aspect. Naturalists have long had a variety of names for these northern sectors, such as 'high polar,' 'cold steppe,' 'arctic wastes,' 'rock desert,' 'polar tundra,' 'barren-ground tundra,' 'fell field' (desert-like), 'desert glacial,' 'nival zone,' and 'dry periglacial zone.' The far north (polar desert) has a number of desert affinities such as saline soil conditions, and the presence of desert pavement and wind-abraded rocks.

## Biotic Factors

One of the distinctive characteristics of the polar landscapes is the absence of trees. The northern edge of forested land marks one of the most striking vegetative changes anywhere on earth. There are, however, a number of problems associated with plotting a circumpolar tree line in detail, because there are interruptions by high mountains, glaciers, and open bodies of water. Further, over long periods the tree line may shift as evidenced by the presence of tree stumps well north of the present tree line. Also, there are differences between the physiognomic forest line and the tree species line.

The tundra zone is characterized mainly by wet cotton-grass meadows and wet sedge meadows, commonly referred to as 'typical tundra' (**Figure 1**). There are also extensive bogs present. The uplands and mountain sectors, however, have drier, upland and barren plant communities.

The polar desert zone supports a discontinuous, sparse vegetative cover (**Figure 2**). Here the vascular plants do not compete for "open space" as they do within the tundra zone. The plant cover consists mainly of lichens, mosses, liverworts, and a scattering of arctic-alpine herbaceous plants.

The subpolar desert zone, also known as midarctic or arctic tundra, is a blend of high arctic and tundra vegetative elements. **Figure 3** shows the location of the three soil zones of the arctic: tundra, polar desert, and subpolar desert. This is a generalized depiction and does not take into account factors such as littoral climates, high mountains, and special geologic conditions.

## Permafrost

Permafrost, also known as perennially frozen ground, is generally considered to be a thickness of soil, surficial deposits, or bedrock at a variable depth below the surface of the earth in which the temperature of the matrix has existed below the freezing point of water for a period of at least several years. Permafrost therefore may consist of frozen soil, frozen bedrock, and clear ice. Permafrost underlies nearly all of the polar lands, including adjacent offshore sectors. The thickness of permafrost varies greatly, with a few locations in Siberia having a depth of more than 1200 m. In Siberia permafrost extends hundreds of miles south of the tree line (Figure 4) and still much further into high mountain areas of the world.

**Figure 1** Wet meadow in the tundra zone, northern Alaska. © J Tedrow.

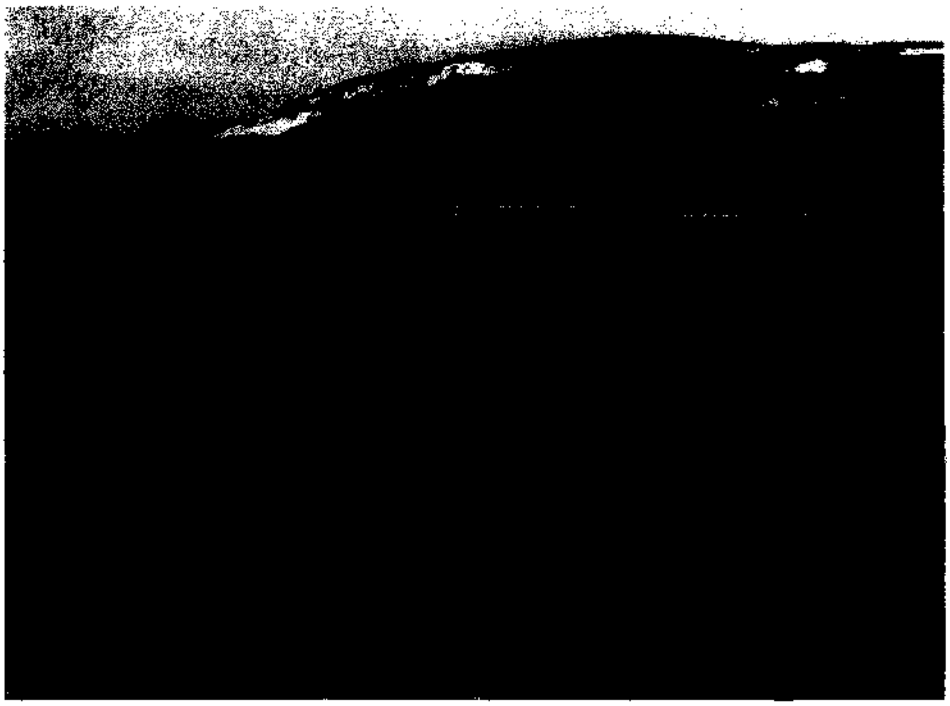


**Figure 2** Polar desert on Cornwallis Island. Most of the landscape is barren except the dark depression, which is mantled with Tundra (Aquiturbel) soil. © J Tedrow.

Of special interest to soil scientists is the rate and depth of thaw in polar soils. Seasonal thaw usually begins in snow-free sectors in early June, but by late July or early August seasonal thaw has largely reached its maximum. Depth of seasonal thaw may be less than 30 cm in the far north, but within the northern fringes of the forested land it may extend to a depth of 150–250 cm or more.

Permafrost generally exists as massive ice or as frozen soil with the pore spaces in the soil filled with ice. There are, however, many locations in mountainous areas and elsewhere where ice is present only at the contact points of the soil particles, a condition referred to as dry permafrost.

## Soil Development and Classification

### Systems of Classification

Up until the mid-1900s, there was very little attention given to detailed classification of polar soils. Tundra soil was recognized, but descriptions were vague and incomplete. The most reliable information was from Russian sources, but even these

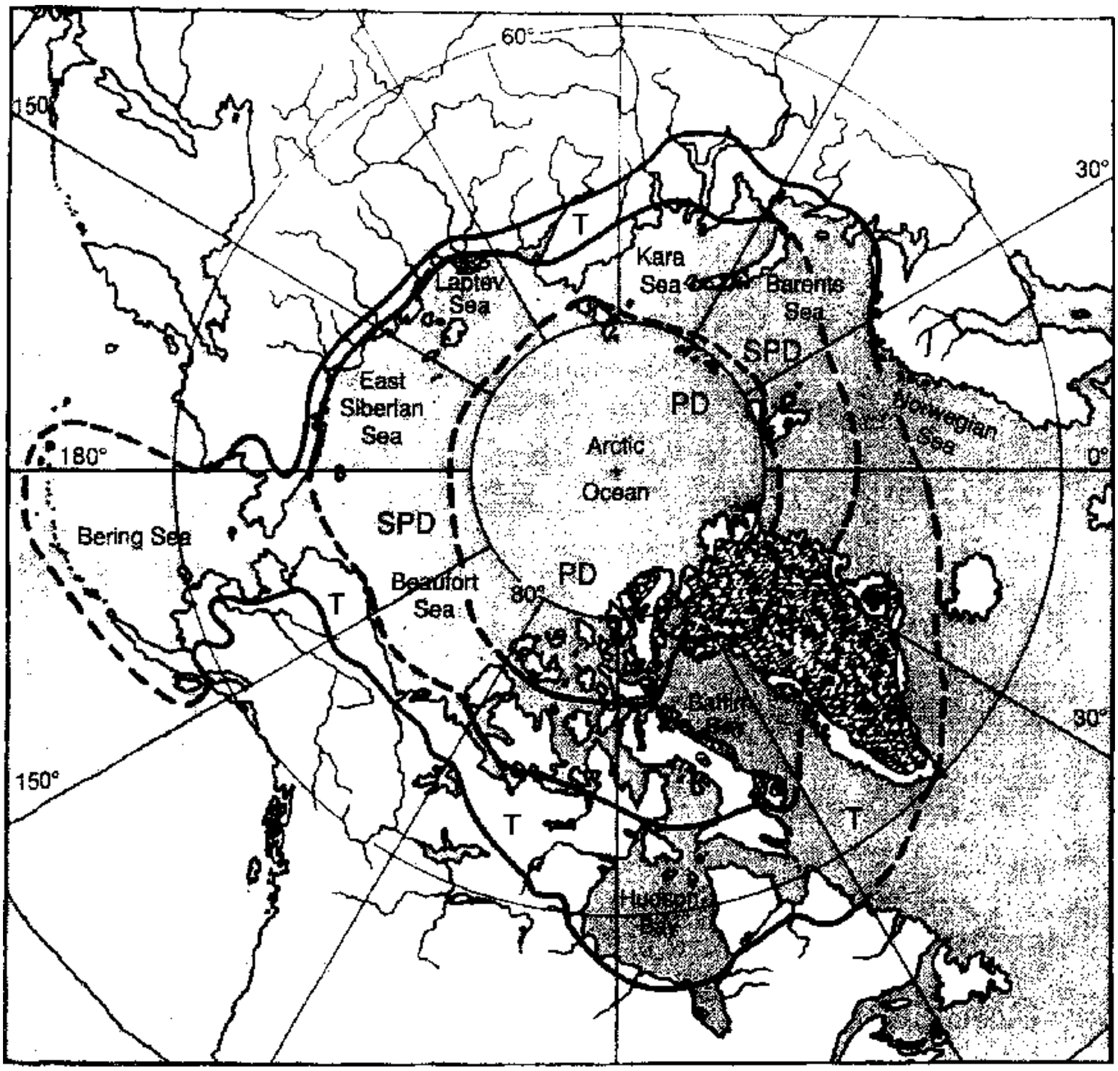


**Figure 3** Delineation of soil zones of the northern polar lands. PD, Polar Desert zone; SPD, Subpolar Desert zone; T, Tundra zone. © J Tedrow.

were preliminary and generally carried out just north of the forested land. During more recent years there have been several proposals advanced for classifying polar soils. **Table 1** shows four systems that have been proposed. Other schemes of classification have also been proposed by Russian, French, and New Zealand (for Antarctica) investigators. In *Soil Taxonomy* (US), soils of the polar regions are collectively referred to as Gelisols (soil order). Gelisols are divided into Histels, Turbels, and Orthels. The Canadian system is similar in principle to that of *Soil Taxonomy* but there are differences in nomenclature. Soils are first described under the Cryosol order followed by Organic Cryosols, Turbic Cryosols, and Static Cryosols. The Food and Agriculture Organization of the UN (FAO) system is similar to that of *Soil Taxonomy* and also of Canada. The Tedrow system uses a somewhat different approach in that the polar regions are first divided into soil zones, then genetic soil divisions. There have been a number of systems proposed by Russian investigators, most of which follow similar lines in that the polar regions are first divided into bioclimatic zones, then genetic soil characterization.

## Tundra Zone

The tundra zone is characterized by a preponderance of hydric mineral soils. These were, for years, the benchmark soils of the northern polar regions. A variety of soils are now recognized within the tundra zone, including the mineral soils that have impeded drainage. The soils are generally silty, but there are wide ranges in texture, from sandy loams and sands to heavy bentonitic clays. Earlier writings have indicated that, with frequent freezing and thawing, individual mineral particles become reduced to silt size, but such a phenomenon has been discounted largely.

**Figure 5** shows an idealized Upland Tundra (Molliturbel) soil. The buried organic-rich layer, as shown at the 24- to 26-inch depth, usually has an age of 8000–11 000 years BP, which puts the time of burial within the Hypsithermal Interval. Pollen from these buried layers is generally rich in Ericaceae, Cyperaceae, *Betula*, and *Salix*.

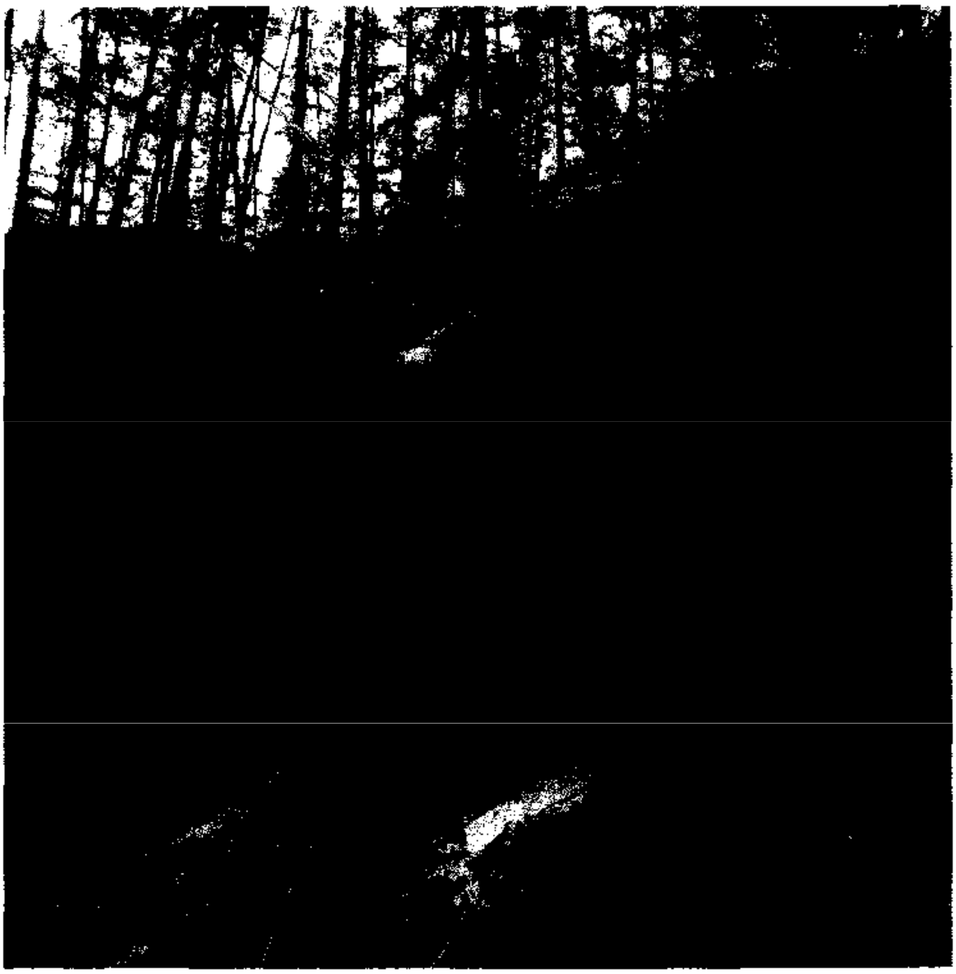


**Figure 4** Ice wedge in permafrost along the Aldan River, Siberia. The top 150 cm active layer above the ice wedge is composed of well-drained silt and fine sand. © J Tedrow.

Clay minerals in Tundra (Molliturbel) soil generally consist of illite and kaolinite. The pH values of wet mineral soils in the tundra zone generally are within the 4.5–6.5 range, but, where there is a dusting of carbonate-bearing minerals from sources such as vegetation-free floodplains, the soils are alkaline.

Within the tundra zone there are extensive sectors of Bog soils (Histels). Organic deposits having a thickness of greater than 10 m have been recorded. The bogs are generally acidic.

There are many small, isolated areas of well-drained soils within the tundra zone. These soils, designated as Arctic Brown (Haplorthels), develop on sites where there is free drainage to a depth of at least 60 cm. The soils develop a brownish color in the solum, much as they do within the northern forested zone. These soils are usually acid, but on wind-swept spurs the soil may be alkaline.

Also, within the tundra zone there are a few Rendzina (Umbrothels) soils formed on limestone bedrock. On the other hand, where the soils from on acid bedrock such as a granite, a Podzol-like (Spodic Psammoturbel) soil may develop. In mountainous sectors of northern lands, the soils may be shallow and rocky, corresponding to a Lithosol (Orthels).

## Polar Desert Zone

Polar Desert soils (Haploturbels) generally form on positive relief elements and on sandy/gravelly plains. Desert pavement is usually present where there is an abundance of gravel in the substrate. Maximum Polar Desert soil development is usually confined to open-textured material underlain by a dry permafrost. A browning effect may be present within a solum with up to 3% of easily reducible iron. Mud coatings and travertine formation may be evident in some locations.

The organic matter content of Polar Desert soil is generally less than 1%, but, where a vascular plant cover is present, the organic matter content increases accordingly. It is believed that most of the organic matter in Polar Desert soil is contributed by algae, including diatoms. The organic matter in Polar Desert soil is quite stable. In northern

**Table 1** Approximate comparison of soil taxa between *Soil Taxonomy*, the World Reference Base for Soil Resources, the Canadian system, and Tedrow's system

| *Soil taxonomy*[a] | *World reference base*[b] | *Canadian*[c] | *Zonal system*[d] |
|---|---|---|---|
| **Gelisol** (soil order) | **Cryosol** (soil unit) | **Cryosol** (order) | **Soils of the Cold Regions** |
| Histels (suborder) | Histic Cryosols (soil subunit) | Organic Cryosols (great group) | Organic (bog) soil (great soil group) |
| Folistels (great group) | (No equivalent) | (No equivalent) | |
| Glacistels | Glacic Histic Cryosols | Glacic Organic Cryosols | |
| Fibristels | Fibric Histic Cryosols | Fibric Organic Cryosols | |
| Hemistels | Mesic Histic Cryosols | Mesic Organic Cryosols | |
| Sapristels | Humic Histic Cryosols | Humic Organic Cryosols | |
| (Terric Fibristels) (subgroup) | (No equivalent) | Terric Fibric Organic Cryosols | |
| (Terric Hemistels) | (No equivalent) | Terric Mesic Organic Cryosols | |
| (Terric Sapristels) | (No equivalent) | Terric Humic Organic Cryosols | |
| Turbels | Turbic Cryosols | Turbic Cryosols | |
| Histoturbels | Stagnic & Gleyic Turbic Cryosols | Gleysolic Turbic Cryosols | Meadow Tundra soils |
| Aquiturbels | Stagnic & Gleyic Turbic Cryosols | Gleysolic Turbic Cryosols | Meadow Tundra soils |
| Anhyturbels | Salic, Gypsic & Calcic Turbic Cryosols | (No equivalent) | Ahumisols, Ahumic (subxerous) soils |
| Molliturbels | Mollic Turbic Cryosols | Brunisolic Turbic Cryosols | Upland Tundra soils, Soils of the Hummocky Ground, Grumusols |
| Umbriturbels | Umbric Turbic Cryosols | Brunisolic Turbic Cryosols | Upland Tundra soils, Soils of the Hummocky Ground, Grumusols |
| Psammoturbels | Regic Turbic Cryosols | Regosolic Turbic Cryosols | Regosols, some Polar Desert soils |
| Haploturbels | Regic Turbic Cryosols | Orthic Turbic Cryosols | Polar Desert soils, Soils of the Solifluction Slopes |
| (Glacic subgroups) | Glacic Turbic Cryosols | (No equivalent) | Meadow Tundra and Upland Tundra soils |
| (Sulfuric Aquaturbels) | Thionic Turbic Cryosols | (No equivalent) | Meadow Tundra and Upland Tundra soils |
| (Andi, Molli-, & Umbriturbels) | Andic Turbic Cryosols | (No equivalent) | Upland Tundra soils, Soils of the Hummocky Ground, Grumusols |
| (Vitrand-, Molli-, & Umbriturbels) | Tephric Turbic Cryosols | (No equivalent) | Upland Tundra soils, Soils of the Hummocky Ground, Grumusols |
| (Spodic Psammoturbels) | Spodic Turbic Cryosols | (No equivalent) | Podzol-like soils |
| Orthels | Haplic Cryosols | Static Cryosols | |
| Historthels | Stagnic & Gleyic Haplic Cryosols | Gleysolic Static Cryosols | Meadow Tundra soils |
| Aquorthels | Stagnic & Gleyic Haplic Cryosols | Gleysolic Static Cryosols | Meadow Tundra soils |
| Anhyorthels | Salic, Gypsic & Calcic Haplic Cryosols | (no equivalent) | Ahumic soils (xerous, ultraxerous) |
| Mollorthels | Mollic Haplic Cryosols | Brunisolic Static Cryosols | Arctic Brown soils, Subarctic Brown soils, Rendzina |
| Umbrorthels | Umbic Haplic Cryosols | Brunisolic Static Cryosols | Arctic Brown soils, Subarctic Brown soils, Rendzina |
| Argiorthels | (No equivalent) | (No equivalent) | Subarctic Grey Wooded soils |
| Psammorthels | Regic Haplic Cryosols | Regosolic Static Cryosols | Regosols, some Polar Desert soils |
| Haplorthels | Regic Haplic Cryosols | Regosolic Static Cryosols | Arctic Brown soils, Subarctic Brown soils |
| (Glacic subgroups) | Glacic Haplic Cryosols | (No equivalent) | Meadow Tundra and Upland Tundra soils |
| (Sulfuric Aquorthels) | Thionic Haplic Cryosols | (No equivalent) | Meadow Tundra and Upland Tundra soils |
| (Andi, Molli-, & Umbrorthels) | Andic Haplic Cryosols | (No equivalent) | Upland Tundra soils, Soils of the Hummocky Ground, Grumusols |
| (Vitrand-, Moll-, & Umbrorthels) | Tephric Haplic Cryosols | (No equivalent) | Upland Tundra soils, Soils of the Hummocky Ground, Grumusols |
| (Spodic Psammorthels) | Spodic Haplic Cryosols | (No equivalent) | Podzol-like soils |

*(Continued)*

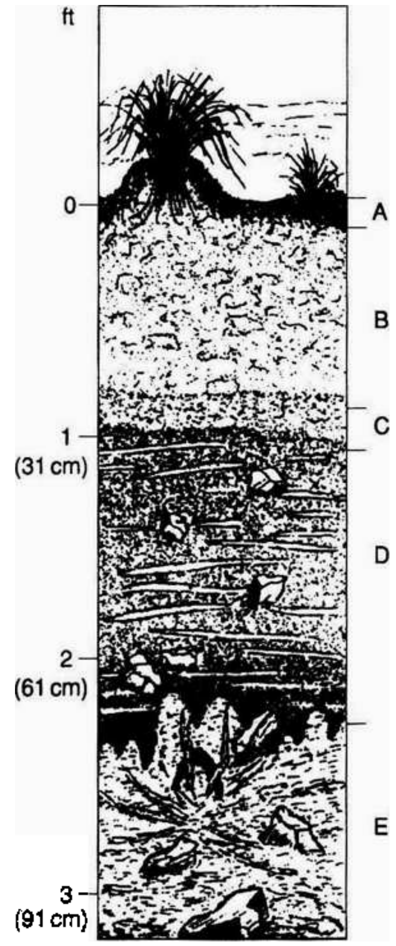


**Figure 5** Idealized profile of Tundra (Molliturbel) soil. A, dark brown to black organic matter, consisting of partially decomposed sedges, sphagnum, and heaths. This horizon may vary from a thin, discontinuous one to an organic mat some 15 cm thick. Usually very loose, fibrous, wet, and strongly acid. B, light olive-brown (2.5Y 5/4) silt loam. Usually very wet, but loose and friable when dried. Nearly always mottled but in varying degrees. Strongly acid. C, dark gray (2.5Y 4/0) silt loam, mottled, very wet. The bottom of this layer approximates the permafrost table. D, very dark gray (2.5Y 3/0) silt loam, permanently frozen. Considerable organic staining in the ice; shreds of organic matter intermixed throughout the horizon. The bottom of the layer is very uneven and is much darker in color than the central portion of the horizon. The upper part of the horizon is also darker than the central portion. The horizon varies from frozen mineral soil to ground ice. Pieces of peaty material are commonly present. E, frozen gray mineral matter interspersed with ground ice. No evidence of organic staining. In many instances the material is nearly clear ice. Weakly acid to calcareous, depending upon parent material. © J Tedrow.

Greenland a sample of topsoil yielded an age of 3300 years BP.

The pH of Polar Desert soils varies widely from approximately 4.4 to as high as 7.9. Likewise electrical conductivity ranges from less than 10 to 66 m$\Omega$ cm$^{-1}$. Salt crusts of thenardite ($Na_2SO_4$) and other minerals form on some of the well-drained to xeric sites. **Figure 6** shows an idealized profile of Polar Desert soil. It is common to have only minor differentiation of the soil horizons other than that induced by the distribution of iron coatings on soil particles.

Within the polar desert zone, Tundra, Bog, Rendzina, and other soils are also present. These soils have similar characteristics to those found in the tundra zone (**Table 1**).

### Subpolar Desert Zone

At one time many polar investigators included the subpolar desert zone as a part of the polar desert zone, but for specialist purposes it is appropriate to recognize a separate zone. The subpolar desert zone has been well recognized by naturalists, particularly plant geographers, who have used a number of climate and phytographic criteria to delineate the zone. Terms such as 'arctic tundra,' 'stony-sedge-moss-lichen tundra,' 'arctic desert,' and 'semipolar desert,' among others have been proposed. The subpolar desert zone is transitional and has few, if any, unique properties of its own. Conditions are somewhat analogous to those in parkland lying between the forest and prairies. There is no single criterion for delineating the subpolar desert zone, instead separation is based on a combination of factors, including soil, plants, and climate. The occasional Polar Desert soils within the subpolar desert zone usually contain little brunification and horizon differentiation, as is the case within the polar desert soil zone. In the subpolar soil zone, Tundra, Bog, Rendzina, and other soils are present (Table 1).

### Patterned Ground

In the polar environment, there are at least two opposing forces operating. One involves pedogenic forces – those resulting in the development of genetic soil horizons. A second set of processes is frost-related and, in effect, consists of destructive forces. 'Patterned

---

Source: JG Bockheim from Bockheim JG and Tarnocai C (2000) Gelisols. In: Sumner ME (ed.) *Handbook of Soil Science*. pp. 256–264. Boca Raton, FL: CRC Press.

[a] Soil Survey Staff (1999) *Soil Taxonomy*. USDA–NRCS. Agricultural Handbook 436. Washington, DC: US Government Printing Office.

[b] Dudal R (1974) Soil map of the world (1). UNESCO, FAO, Rome.

[c] Tarnocai C (1994) Genesis of permafrost-affected soil. Proc. Meeting on classification, correlation, and management of permafrost-affected soils. USDA. Lincoln, Nebraska, USA.

[d] Tedrow JCF (1977) *Soils of the Polar Landscapes*. New Brunswick, NJ: Rutgers University Press.

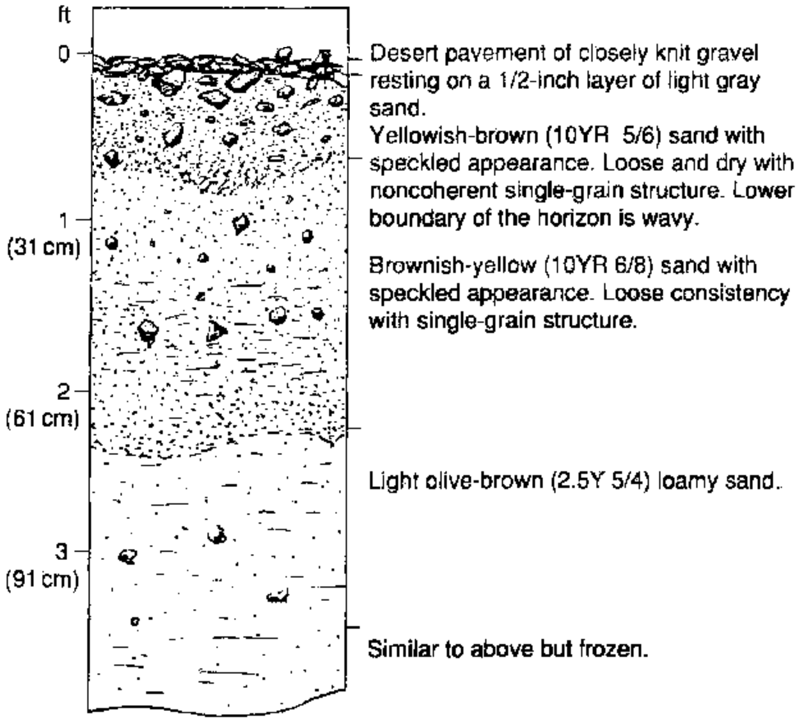


**Figure 6** Idealized profile of Polar Desert (Pasammorthel) soil. © J Tedrow.

ground' is a group term for symmetrical forms such as circles, polygons, nets, steps, and stripes. Various terms have been used to describe the frost structures, such as 'rutmark,' 'råmark,' 'frost polygon,' 'strukturboden,' 'polygonboden,' 'ice-wedge polygon,' 'stone polygon,' 'soil circle,' 'mud circle,' 'stone ring,' 'stone stripe,' 'hummocky ground,' 'frost boil,' 'ice mound,' 'palsa,' 'pingo,' and many others.

One of the more common forms of patterned ground is the ice-wedge polygon. The ground develops a series of ice wedges that result from annual soil temperature amplitudes. Year after year the ice wedges tend to increase in size, exerting exceptionally strong lateral forces that result in a buckling of both the edges and centers of the polygons (Figure 7). Changes of elevation of the soil surfaces may result in the center of the polygon being either raised or depressed. Change in elevation induced by frost action will result in some parts of the polygon being better drained, but other parts will become more poorly drained. This complex phenomenon of changes in drainage induced by frost action has received very little attention from pedologists. In addition to hydric changes in the soil induced by polygon formation, another set of cryogenic processes ruptures and deforms the soil horizons.

Apart from polygonal ground formation, another set of processes involves frost stirring, resulting in the development of stone rings (sorted circles). **Figure 8** shows a form of stone rings in which the pattern of deformation has resulted from an upwelling of the mineral material in the form of a core of raw, earthy material and an outward displacement of the coarser rock fragments. **Figure** 8 shows lichen crusts on the outer rock fragments, whereas the central part of the stone ring is free of lichens, indicating a less stable condition.

Nonsorted circles are common throughout the polar regions, especially where the soils are of medium to clayey texture. **Figure** 9 shows a slight dome-shaped pattern with small, secondary desiccation cracks (zellenboden).

On sloping land, there are many elongated forms of patterned ground that result from at least two processes: frost sorting and solifluction (the slow gravity driven downhill movement of saturated soil). Patterns consist of stone garlands, stone stripes, soil stripes, and others. Where solifluction processes are intense, the genetic soil horizons become highly deformed and, in some locations, completely obliterated.

### Cold Desert Zone

The antarctic continent, also referred to as the cold desert zone, was the last sector of the globe to receive the attention of pedologists. Only since World War II has there been significant pedologic information forthcoming from Antarctica. Soil maps of the globe traditionally provide information only to about 60° S, which excludes all of Antarctica (**Figure 10**). One of the interesting aspects of cold desert zone pedology is that there is virtually no effective organic component in the soils except for a few special conditions such as

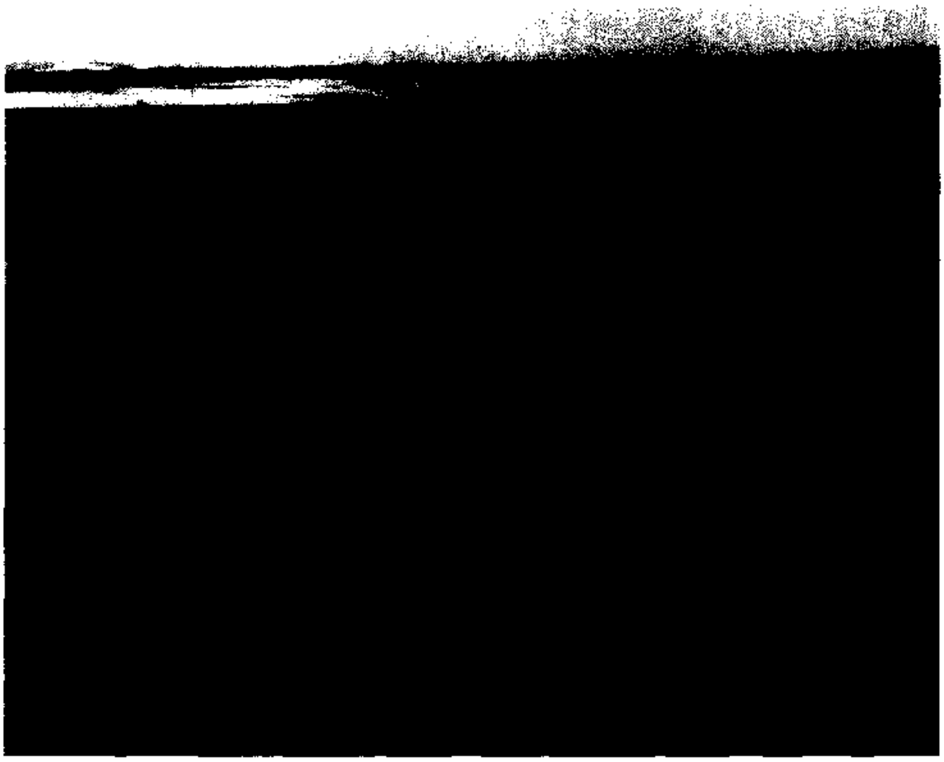


**Figure 7** Ice-wedge (nonsorted) polygons in Prince Patrick Island. © J Tedrow.

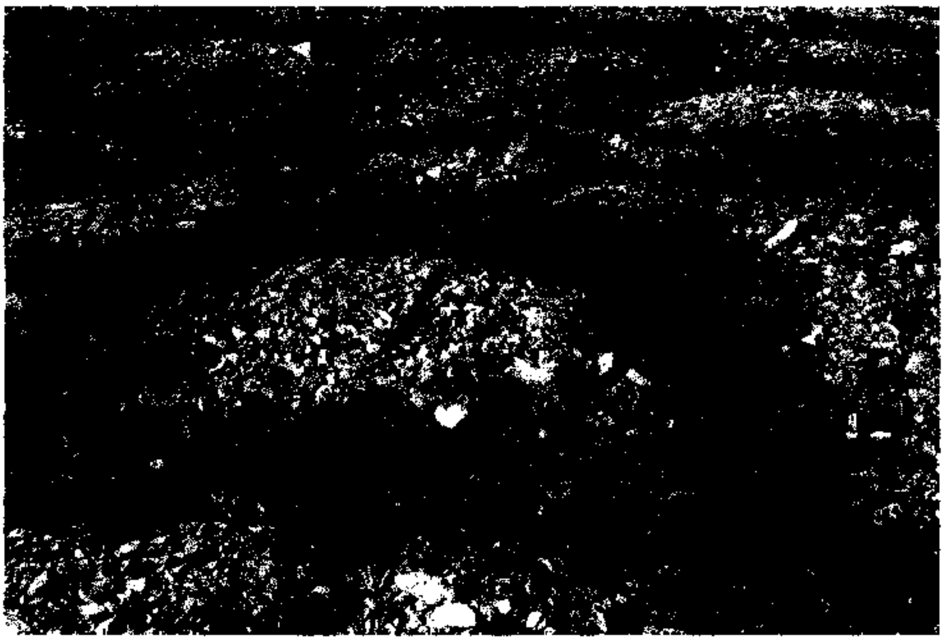


**Figure 8** Stone rings (sorted circles) in Prince Patrick Island. © J Tedrow.

occur at the penguin rookeries, where there are local accumulations of guano. Late in the nineteenth century, pedology was recognized as an independent scientific discipline, and it was agreed that organic matter was an essential component of the soil system. While virtually all pedologists who have worked in Antarctica speak of soils in the traditional sense, plenary bodies have largely ignored the question of antarctic soils. The position of plenary bodies was understandable when, in earlier times, very little information was available concerning Antarctica, but considerable pedologic information is now available.

Of the 2–3% of Antarctica that is free of ice and snow, only a small percentage has surficial deposits adequate for any semblance of soil formation. The remainder of the ice-free land consists of bedrock and boulder fields. Permafrost is believed to be present throughout the cold desert zone, but most

**Figure 9** Nonsorted circles in northern Alaska. © J Tedrow.

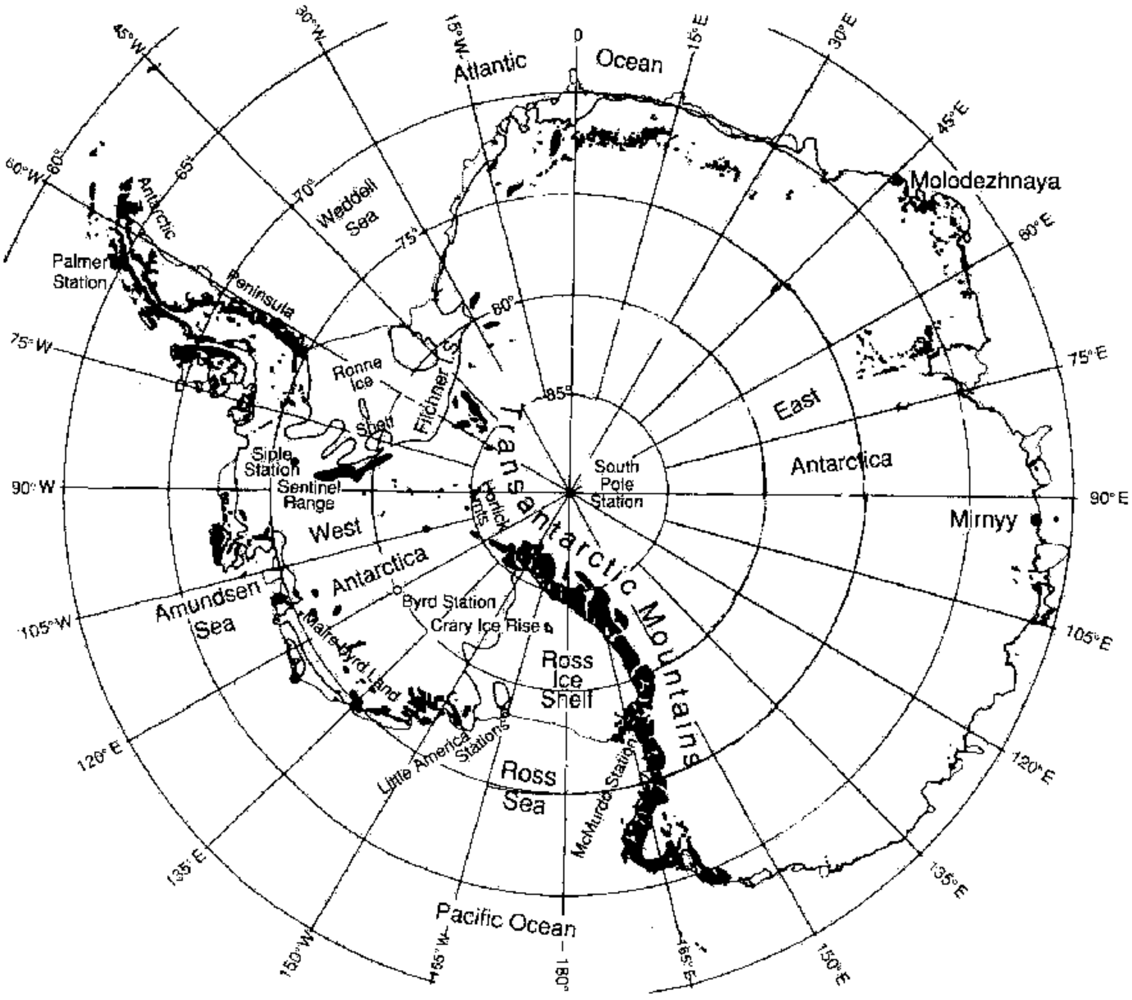


**Figure 10** Map of Antarctica. The dark areas are largely without a snow-ice cover. © J Tedrow.

of the surface deposits are subject to dry frost conditions.

Physical and chemical weathering in the cold desert zone is intense. Boulders exhibit cavernous weathering forms in which shells of exfoliation can be pried off. Wind polishing of rocks is intense with the abrasion caused by the blowing of ice particles. Ice in the cold desert at very low temperatures may have a hardness like that of quartz. Desert varnish is also evident in many places, especially on older land surfaces.

Although air temperatures in the cold desert zone may be well below freezing, solar radiation causes the rock surfaces to attain elevated temperatures on wind-free, sunny days. The temperature gradient in rocks during sunny days results in both physical and chemical weathering. Garnets split, feldspars weather, and clay substances accumulate in the fractured areas. Accumulation of calcite, gypsum, magnesium sulfate, and other minerals is widespread. Also, highly soluble forms of calcium, magnesium, sodium, potassium, chloride, sulfate, and nitrate have been reported. There are a number of forms of patterned ground present in antarctic deserts, with the most common form being of a polygonal nature.

**Table 2** shows two classification schemes proposed for the cold desert zone, both of which follow the same pathways in principle. **Figure 11** shows an idealized soil of the cold desert (Subxerous-Xerous). This depiction is one of maximum soil development on till composed mostly of sandstone, conglomerate, and dolerite. Horizonation is somewhat blurred, with no structural aggregates being present. The 'active' layer is approximately 51 cm thick but, because of the ultra dry, loose condition, depth to the frost table can only be approximated. Cold desert soils, for all practical purposes, are without terrestrial life. Algae including lichens are occasionally present and, in rare

**Table 2** Comparison of several proposed schemes for classifying cold desert soils of Antarctica

| *Campbell and Claridge*[a] | *Tedrow*[b] |
|---|---|
| *Frigic soils* | *Soils of the cold desert zone* |
| *Zonal soils* | *Zonal soils* |
| Ultraxerous | Ultraxerous |
| Xerous | Xerous |
| Subxerous | Subxerous |
| — | Ahumisol |
| — | Protoranker |
| *Intrazonal soils* | *Intrazonal soils* |
| Evaporite soils | Evaporite soils |
| Algal peat soils | Algal peat soils |
| Avian soils | Ornithogenic (Avian) soils |
| Hydrothermal soils | Hydrothermal soils |
| *Azonal soils* | *Azonal soils* |
| Recent soils | Regosols |
| | Lithosols |

[a] Campbell IB and Claridge GGC (1969) A classification of frigic soils – the zonal soils of the antarctic continent. *Soil Science* 107: 75–85.

[b] Tedrow JCF (1977) *Soils of the Polar Landscapes*. New Brunswick, NJ: Rutgers University Press.

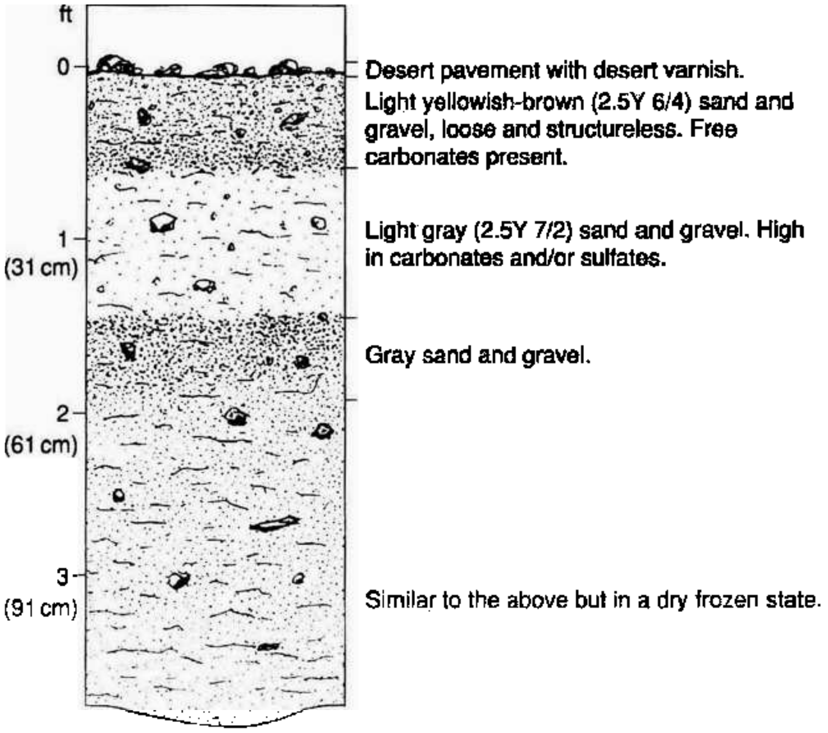


**Figure 11** Profile of Cold Desert (Gelisol) soil. © J Tedrow.

situations, protozoa, rotifers, and nematodes can be found.

There is evidence that there is ionic migration in the soil, even when it is in a dry, frozen state. Buildup of salts is common in the soil as well as in bedrock and boulders. Nearly all soils of the cold desert are alkaline, with pH values being in the 7–10 range. In special situations, however, acid soils have been reported.

Closed basins, depressions, and other sites commonly have saline soils and accumulations of salt encrustations. Saline lakes have been reported in the dry valleys of Victoria Land and in the Mirnyy area. Salt content of water in a closed lake basin in Wright Valley is 34%. The salt consists mainly of calcium and chloride. Clay minerals in cold desert soils consist primarily of micas, kaolinite, vermiculite, montmorillonite, and chlorite. Some of the clay-size fraction also contains quartz, feldspar, and amorphous material.

The cold desert soils can be divided into Ultraxerous, Xerous, and Subxerous varieties (Table 2). In the Molodezhnaya sector of Antarctica, some of the soils show advanced weathering and have been designated as Ahumisol. There are also a very few, extremely small locations, where there are clumps of vascular plants, in which the soils have been designated as Protoranker. Along the margins of some saline lakes, there are shallow accumulations of algae. Also, in the penguin (and skua) nesting areas, the soils, rich in bird droppings, feathers, and bones, have been designated as Ornithogenic or Avian. In the area of volcanic vents such as at the top of Mt. Erebus, the soils have been designated as Hydrothermal.

*See also:* **Factors of Soil Formation:** Climate; **Freezing and Thawing:** Cycles; Processes; **Pedology:** Basic Principles

## Further Reading

Aleksandrova VD (1980) *The Arctic and Antarctic: Their Division into Geobotanical Areas.* Cambridge, UK: Cambridge University Press.

Bockheim JG (2002) Gelisols: the permafrost-affected soils. In: *Encyclopedia of Soil Science.* New York: Marcel Dekker, Inc.

Brown J, Ferrians OJ, Heginbottom JA, and Melnikov ES (1997) *Circum-Arctic Map of Permafrost and Ground-Ice Conditions.* US Geological Survey, Map CP-45. Washington, DC: US Goverment Printing Office.

Campbell IB and Claridge GGC (1969) A classification of frigio soils – the zonal soils of the antarctic continent. *Soil Science* 107: 75–85.

Campbell IB and Claridge GGC (1987) *Antarctica: Soils, Weathering Processes and Environment.* New York: Elsevier.

Glazovskaia MA (1984) *Soils of the World.* Moscow, Russia: Moscow University Press.

Glinka KD (1927) *The Great Soil Groups of the World and Their Development.* Ann Arbor, MI: Edwards Bros.

Gorodkov BN (1939) Peculiarities of the Arctic topsoil. (in Russian). *Izv. Gosud. Geogr. Obsch.* 71: 1516–1532.

Goryachkin SV, Karavaeva NA, and Targulian VO (1998) Geography of Arctic soils: current problems. *Eurasian Soil Science* 31: 467–476.

Jahn A (1975) *Problems of the Periglacial Zone.* National Science Foundation (TT72-54001). Washington, DC: National Science Foundation.

Kimble M (ed.) (2004) *Cryosols–Permafrost-Affected Soils.* Berlin: Springer-Verlag.

King CAM (ed.) (1975) *Periglacial Processes.* New York: Halsted Press.

Linell KA and Tedrow JCF (1981) *Soil and Permafrost Surveys in the Arctic.* Oxford, UK: Oxford University Press.

Rieger S (1983) *The Genesis and Classification of Cold Soils.* New York: Academic Press.

Soil Survey Staff (1999) *Soil Taxonomy,* 2nd edn. US Department of Agriculture, Agricultural Handbook 436. Washington, DC: US Government Printing Office.

Tedrow JCF (1977) *Soils of the Polar Landscapes.* New Brunswick, NJ: Rutgers University Press.

Tedrow JCF (1991) *Pedogenic Linkage Between the Cold Deserts of Antarctica and the Polar Deserts of the High Arctic.* Antarctic Research Series, vol. 53, pp. 1–17. Washington, DC: American Geophysical Union.

Ugolini FC (1986) Pedologic zonation in the well-drained soils of Arctic regions. *Quaternary Research* 26: 100–120.

Ugolini FC (1986) Processes and rates of weathering in Cold and Polar desert environments. In: *Rates of Chemical Weathering of Rocks and Minerals,* pp. 193–235. New York: Academic Press.

Washburn AL (1980) *Geocryology: A Survey of Periglacial Processes and Environments.* New York: John Wiley & Sons, Inc.

Williams PJ and Smith MW (1989) *The Frozen Earth: Fundamentals of Geocryology.* Cambridge, UK: Cambridge University Press.

# POLLUTANTS

Contents

## Biodegradation

**P B Hatzinger**, Shaw Environmental, Inc., Lawrenceville, NJ, USA
**J W Kelsey**, Muhlenberg College, Allentown, PA, USA

### Introduction

Industrialized societies annually release large quantities of synthetic chemicals into Earth's natural environments, both intentionally, in the case of pesticides, and inadvertently, through the production, transport, refining, and combustion of fossil fuels, waste streams from manufacturing, chemical spills, disposal of products by consumers, and discharges from a wide variety of other sources. Currently, there are more than 75 000 chemicals registered in the US Environmental Protection Agency (US EPA) inventory of chemical substances manufactured or processed in, or imported into, the USA. The number of potential chemical contaminants is substantially higher, as the EPA inventory does not include pesticides, radioactive materials, food additives, drugs, or cosmetics. Many of these compounds ultimately end up as contaminants in the atmosphere, water, sediment, or soil. To add to the complexity of the problem, the compounds of concern vary widely in their characteristics. A contaminant can belong to any of a number of disparate classes of compounds that differ in important properties such as charge, polarity, water solubility, $pK_a$, molecular weight, and aromaticity. These chemical characteristics subsequently determine a contaminant's fate in the environment and its hazard to humans and other organisms.

One of the most important processes determining the fate of a given chemical is its susceptibility to biological degradation. There have been more than five decades of research into chemical biodegradation, including microbial mechanisms and the chemical and environmental factors that govern its rate and extent in soils.

### Biodegradation

Biodegradation is the partial or complete breakdown of a chemical pollutant resulting from the physiological reactions of microorganisms. In general, microorganisms carry out such processes to obtain necessary carbon, nutrients, and energy for their survival and growth. Through the millennia, bacteria have experienced extreme competition (primarily from each other) for limiting resources. As a result, they have evolved unrivaled metabolic capabilities, utilizing nearly every thermodynamically favorable reaction for energy and nearly every organic molecule for carbon. For example, whereas humans gain energy only through the oxidation of specific organic substrates using oxygen as an electron acceptor (i.e., through aerobic respiration), bacteria utilize virtually all natural and many man-made organics, numerous inorganic molecules (e.g., Fe(0); Fe(II), $H_2S$, S(0) $H_2$, $NH_4$), and light for energy. Moreover, besides oxygen, specific microorganisms can utilize a variety of chemicals, including an assortment of synthetic pollutants and pesticides, as terminal electron acceptors. Specific examples of microbial biodegradation processes under aerobic and anaerobic conditions are provided in subsequent sections.

The microbial degradation of pollutants can yield many different products. Generally, the biodegradation of a large organic molecule results in smaller, less complex organic molecules that may be more or less toxic than the parent compound. In some cases, an organic contaminant is completely converted to an inorganic product such as $CO_2$. A transformation of this type is referred to as 'mineralization' and is associated with the elimination of toxicity of the parent compound. Although plants, animals, and microorganisms can all metabolize some pollutants, the mineralization of organic contaminants is generally carried out by bacteria and fungi only. These microorganisms are responsible for nearly all of the relevant biotransformations of environmental pollutants.

Whereas biodegradation occurs naturally in soils and other environments without any intervention, bioremediation involves the application of natural processes to remove pollutants from contaminated sites. In short, bioremediation is the use of microorganisms to clean up unwanted compounds, although the specific technology employed will depend on the nature of the site and the pollutant of concern. Contaminants may be treated *in situ* through the addition of nutrients to stimulate indigenous microorganisms (biostimulation) or by the inoculation of specific microorganisms capable of degrading the compound (bioaugmentation). Alternatively, contaminated soils can be excavated and treated *ex situ* using biological reactors, composting, or other approaches.

The biodegradation of soil pollutants is critical because of the impact the resultant transformations can have on human and environmental health. For example, in the absence of appreciable degradation, compounds released into soil may leach to groundwater. Contamination of groundwater poses a potentially serious threat to drinking-water supplies and public health. Moreover, compounds present in groundwater have a tendency to persist much longer than those in soils due to the limited supplies of oxygen and inorganic nutrients and low microbial numbers generally found in subsurface aquifers. Soil pollutants and pesticides can also pose a threat to human and nonhuman health if they are bioconcentrated by soil organisms. Biological uptake directly from soil can adversely affect both individual organisms and food webs. A contaminant may be sufficiently concentrated in soil to harm the organism that contacts it initially. However, certain compounds, even if initially present at low concentration, can be accumulated and magnified through herbivory and predation. A biomagnified compound has the potential to damage organisms at higher trophic levels, despite their very limited contact with the contaminated soil. Alternatively, some compounds are so potent that they are a threat at very low concentrations. Microbial transformations can lower the risk associated with a particular pollutant because they reduce or eliminate the toxicity of the compound.

## Aerobic Transformation Processes

Many different synthetic pollutants and pesticides have been shown to serve as growth substrates for specific bacteria and fungi under aerobic conditions, providing both carbon and energy to these organisms. Often, these compounds are completely mineralized by one organism or by a series of different organisms to $CO_2$ and cell constituents. During this process, the novel pollutant is generally modified by specific enzymes to one or more metabolites that are common in cellular metabolism (e.g., pyruvic acid, acetyl-CoA). Those metabolites are then converted to energy and biomass through the common metabolic pathways. Aerobic mineralization is generally a growth-linked process whereby microbial numbers increase as the chemical concentration decreases. Growth-linked biodegradation is desirable in that it can bring about a rapid loss of a contaminant from soil. Many different organic contaminants are subject to mineralization in soils, including most components of crude oil and gasoline, and numerous pesticides.

A variety of organic chemicals are biotransformed through a nongrowth-linked process termed 'cometabolism.' Cometabolism is the fortuitous transformation of a molecule by an enzyme synthesized within a cell for another purpose. By this process, the cell modifies the chemical structure, but gains no energy, carbon, or inorganic nutrients from the reaction. Several groups of aerobic bacteria present in soils, including methanotrophic bacteria (which oxidize methane for energy), nitrifying bacteria (which oxidize ammonia for energy), and propane-, phenol-, and toluene-oxidizing bacteria, have been observed to transform a variety of different compounds through cometabolic reactions. Each of the above groups of bacteria produces nonselective monooxygenase or dioxygenase enzymes that are responsible for these reactions. For example, methanotrophic bacteria produce methane monooxygenase (MMO), an enzyme that normally functions to oxidize methane for energy (**Figure 1**). However, MMO has a broad substrate specificity and has been observed to catalyze the initial oxidation of numerous alkanes, alkenes, and aromatics as well as the degradation of an array of chlorinated aliphatics, such as *cis*- and *trans*-1,2-dichloroethylene (*cis*- and *trans*-1,2-DCE), trichloroethylene (TCE), and 1,2-dichloropropane. As previously noted, bacteria do not gain carbon or energy from the cometabolic oxidation of most compounds; thus, an additional cosubstrate is often required for bacterial growth (e.g., toluene for toluene

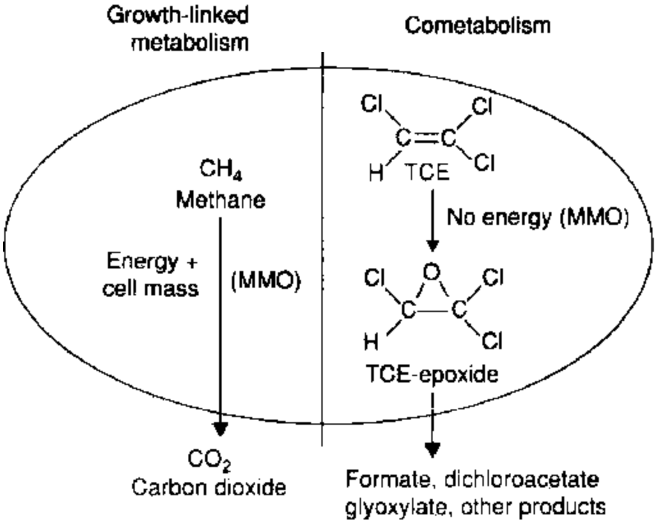


**Figure 1** Representation of an individual methanotrophic bacterium degrading methane to carbon dioxide through growth-linked metabolism and degrading trichloroethylene (TCE) to TCE epoxide via cometabolism. MMO, methane monooxygenase.

oxidizers, methane for methanotrophs, etc.). Although cometabolic reactions sometimes result in the accumulation of persistent intermediates, in many instances, the initial oxidative reaction produces readily degradable compounds that are then mineralized. For example, as shown in **Figure 1**, the initial cometabolic attack of MMO on TCE creates an unstable TCE epoxide, which further degrades to a variety of growth substrates, such as glyoxylate, formate, and dichloroacetate.

### Anaerobic Transformation Processes

A wide variety of pollutants and pesticides are subject to biodegradation under anaerobic conditions. As described for aerobic degradation, both mineralization and cometabolic reactions also occur in anaerobic environments. For some compounds, biodegradation will only occur under such conditions, since the organisms capable of carrying out the process are obligate anaerobes (i.e., they die in the presence of oxygen). Under anaerobic conditions, the degradation of specific pollutants and pesticides can be coupled to the reduction of terminal electron acceptors other than $O_2$, such as $NO_3$, Mn(IV), Fe(III), $SO_4$, or $CO_2$. This process is termed ‘anaerobic respiration.’ The gasoline constituent toluene, for example, has been reported to be biologically degraded with each of these different electron acceptors. For a given substrate, the energy yield during anaerobic respiration with the aforementioned compounds is less than for the same reactions using oxygen ($O_2 > NO_3 > \mathrm{Mn(IV)} > \mathrm{Fe(III)} > SO_4 > CO_2$) as an electron acceptor. Therefore, the anaerobic reactions tend to occur more slowly, and microbial growth yields are lower than for aerobic reactions.

If a large quantity of a readily biodegradable pollutant is added to a soil, it is not uncommon for the dominant, terminal electron-accepting reaction to progress from aerobic respiration (with $O_2$ as the primary electron acceptor) all the way to methanogenesis (with $CO_2$ as the primary acceptor), passing through each dominant electron acceptor in the order shown in **Figure 2**. Some microorganisms (e.g., facultative anaerobes) can utilize more than one electron acceptor, although rarely are these reactions performed simultaneously by a given organism. For example, many denitrifying bacteria grow aerobically in the presence of oxygen, then anaerobically on nitrate when oxygen levels are reduced. Other microorganisms, including many sulfate-reducing bacteria, are obligate anaerobes and are capable of using only one of the common electron acceptors (i.e., sulfate) for growth. As electron acceptors are depleted in sequence, the dominant metabolic reactions (and in many instances, the dominant microorganisms) accounting for the chemical losses shift. The reduction potential declines accordingly as the sequence of terminal electron acceptors progresses from oxygen to carbon dioxide. In the absence of available external electron acceptors, some organic molecules are also subject to fermentation, whereby small amounts of energy are gained by an organism through internal oxidation–reduction reactions. However, when all common electron acceptors are consumed in a contaminated soil, the biodegradation reaction is likely to cease, even though biodegradable compounds are still present. It is for this reason that oxygen is frequently added via sparging or other means to enhance biodegradation in environments that have experienced heavy contamination with organic pollutants. It may also be necessary to add inorganic fertilizers in such instances, as the supply of macro- and micronutrients in the soil may be inadequate to sustain microbial degradative activity.

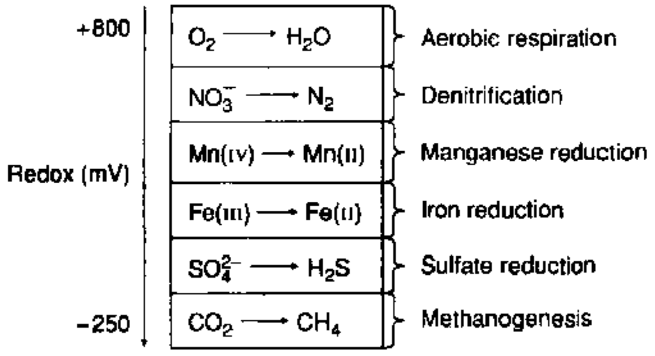


**Figure 2** Sequence of common electron-accepting reactions. Reactions are given in general order of occurrence from top to bottom, with the reactions at the top taking place at the highest redox potential and providing the highest energy yield to microorganisms during degradation of a given organic compound.

Some groups of chlorinated contaminants, including polychlorinated biphenyls (PCBs), chlorophenols, chlorinated solvents, and certain organochlorine pesticides can serve as alternate electron acceptors, by a process known as reductive dechlorination (also called halorespiration). This process, which occurs only in the absence of oxygen, is a critical step in the degradation of these persistent pollutants. During reductive dechlorination, bacteria transfer electrons from a growth substrate (primarily hydrogen) to the chlorinated compounds, causing the release of chloride ion and the production of various organic daughter products. For example, during reductive dechlorination of perchloroethylene (PCE), the daughter products are TCE, *cis*-1,2-DCE, vinyl chloride (VC), and ethylene, respectively (**Figure 3**). The occurrence of these degradation products (particularly *cis*-1,2-DCE and VC) suggests that reductive dechlorination of PCE or TCE is occurring in an environment. A similar reaction has been noted for PCBs and several organochlorine pesticides (e.g., 1,1,1-trichloro-2,2-bis

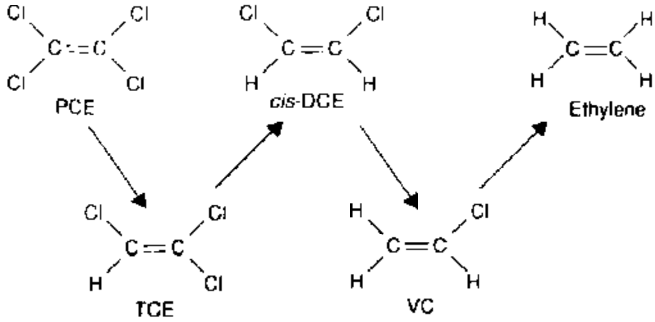


**Figure 3** Biodegradation pathway of perchloroethylene (PCE) during reductive dechlorination. *cis*-DCE, *cis*-1,2-dichloroethylene; VC, vinyl chloride.

(4-chlorophenyl)ethane (DDT); pentachlorophenol (PCP); and 3,5,6-trichloro-4-amino-2-pyridinecarboxylic acid (picloram)) under strongly reducing conditions. In soils, reductive dechlorination of PCBs is observed as a shift in the representative PCB congeners found within an environment (there are 209 PCB congeners based on possible positions and numbers of chlorine atoms on the biphenyl molecule) from more chlorinated to less chlorinated congeners. Other chemical pollutants such as perchlorate ($ClO_4$; an oxidant used in solid rocket fuel) and hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX; an explosive), have also been observed to serve as terminal electron acceptors in soils.

## Chemical Properties Affecting Biodegradation

All natural organic compounds are biodegradable. The occurrence and rates of biodegradation of these compounds vary widely depending on environmental conditions and chemical complexity (e.g., lignin biodegradation is extremely slow, even under ideal conditions), but it is generally believed that no natural organic materials are completely resistant to microbial destruction. Given hundreds of millions of years of evolution and the constant competition among microorganisms for limiting nutrients and energy, it is not surprising that microbial enzymes have coevolved in nature in step with new organic compounds. In contrast to nature, however, rapid advances in organic chemistry have allowed mankind to develop and produce an overwhelming array of new organic compounds in a short time period. A majority of these thousands of new compounds have been introduced to the Earth's environment for the first time since the mid-1940s. Many of these chemicals, particularly those that are similar in structure to naturally occurring organics, have proven to be readily biodegradable. Other, more novel synthetics, however, are highly resistant to microbial attack.

Some of the key factors that influence the degradability of a molecule include the type, number, and location of molecular substituents, the type of bonds, the degree of branching, and the molecular weight of the compound (**Figure 4**). The addition of new organic or inorganic substituents to a common molecular structure is one means of creating novel synthetic compounds. Small differences in the type and location of these substituents can determine whether a molecule is readily biodegradable or highly persistent. Pollutants that are resistant to biodegradation owing to their chemical structure are termed 'recalcitrant' molecules. A number of synthetic organic compounds and pesticides have been developed by adding halogens (particularly chlorine, fluorine, and bromine) to organic compounds. Because the halogen–carbon bond is very stable and resistant to microbial attack, this type of substitution tends to make a molecule recalcitrant, with increasing halogenation leading to increasing persistence. For example, organochlorine pesticides, such as DDT, toxaphene, 2,4,5-trichlorophenoxyacetic acid (2,4,5-T), pentachlorophenol, chlordane, aldrin, hexachlor, and metoxychlor, were once widely used. Many of these pesticides have now been banned from use in the USA and most European countries due to their long-term persistence and negative impacts in various ecosystems. These molecules proved to be highly resistant to aerobic biodegradation and persistent or very slowly degraded under anaerobic conditions. Halogenated pollutants such as PCBs, polybrominated biphenyls (PBBs), freon propellants and refrigerants (e.g., $CCl_3F$, $CCl_2F_2$), and chlorinated solvents (e.g., PCE, carbon tetrachloride) have also proven to be persistent contaminants.

Other molecular traits that may decrease a chemical's biodegradability include high molecular weight, low water solubility, and extensive branching. As molecular weight increases to more than approximately 500–600, chemical biodegradability tends to decrease markedly. This reflects, in part, the inability of large molecules to enter the microbial cell. In some instances, extracellular enzymes produced by bacteria or fungi can initiate an attack on a large molecule, breaking it into smaller pieces that can then be metabolized further. For example, it is large size that prevents the biodegradation of many organic polymers that constitute plastics, such as polyethylene and polystyrene. If reduced to a molecular weight of less than 500 from their typical molecular weights of a few thousand to more than 100 000, many of these polymer structures are biodegradable. Molecular branching also tends to reduce chemical biodegradability. The molecular branches tend to sterically interfere with the common degradative enzymes used by microorganisms. For example, many alkanes

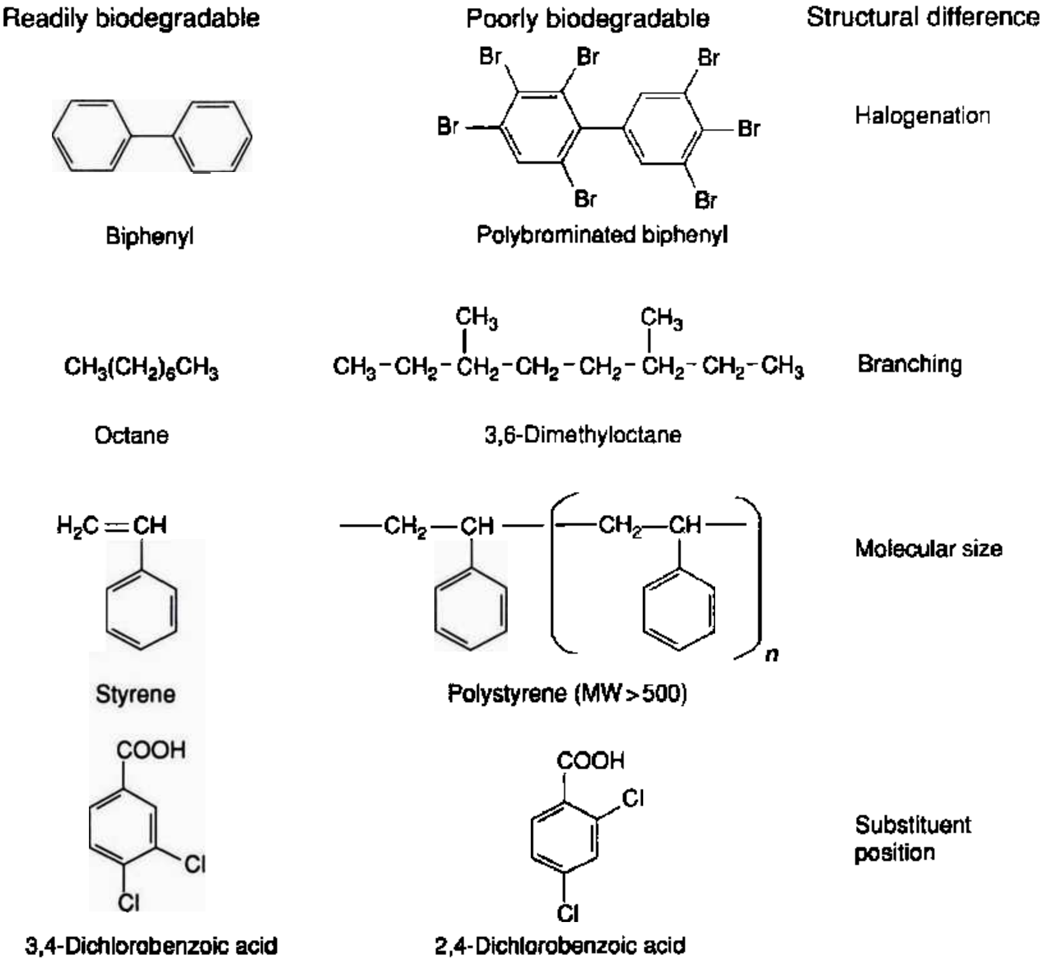


**Figure 4** Influence of molecular structure on contaminant biodegradability.

are transformed by bacteria through the sequential oxidation and cleavage of two carbon molecules at a time, in a process known as beta-oxidation. During this process, the alkane is sequentially oxidized (usually at the terminal carbon) to an alcohol, an aldehyde, and then an aliphatic acid. The acid is then cleaved to form acetyl-CoA, which the cell can utilize for energy. This beta-oxidation process continues until the alkane is completely degraded to acetyl-CoA. Branching of the alkane, particularly branching with one or more methyl groups, can block the two-carbon cleavage of beta-oxidation, causing an otherwise biodegradable molecule to become recalcitrant. For this reason, branched alkanes, such as pristaine and phytane, tend to be among the slowest molecules to biodegrade during a crude oil spill. These molecules are frequently employed as internal standards to assess the biodegradation of other, more labile, components of crude oil after an environmental release.

## Environmental Factors Affecting Biodegradation

A compound that is readily metabolized in a laboratory may escape breakdown and persist unexpectedly in the field, even if an organism capable of causing its degradation is present. Unless pollutants are accessible to microorganisms and environmental conditions are conducive to microbial growth, contaminants will not undergo biotransformations and may remain in soil for extended periods. Thus, the extent of chemical biodegradation will depend in large part upon the physical, chemical, and biological properties of a contaminated site. A number of important factors influencing the growth of microorganisms and therefore biodegradation are described below.

### Bioavailability

Pollutants released into soil will not be biodegraded if they are biologically unavailable. Even the most labile compound may persist if it does not come into contact with microorganisms. Bioavailability of a compound can be reduced if the compound is sorbed to soil particles, entrapped within the soil matrix, or dissolved in a nonaqueous solvent (**Figure 5**).

Sorption Chemical compounds can bind to solid surfaces through the actions of a variety of attractive forces. In some instances, a compound will adhere to the outer surfaces of soil particles in a process known

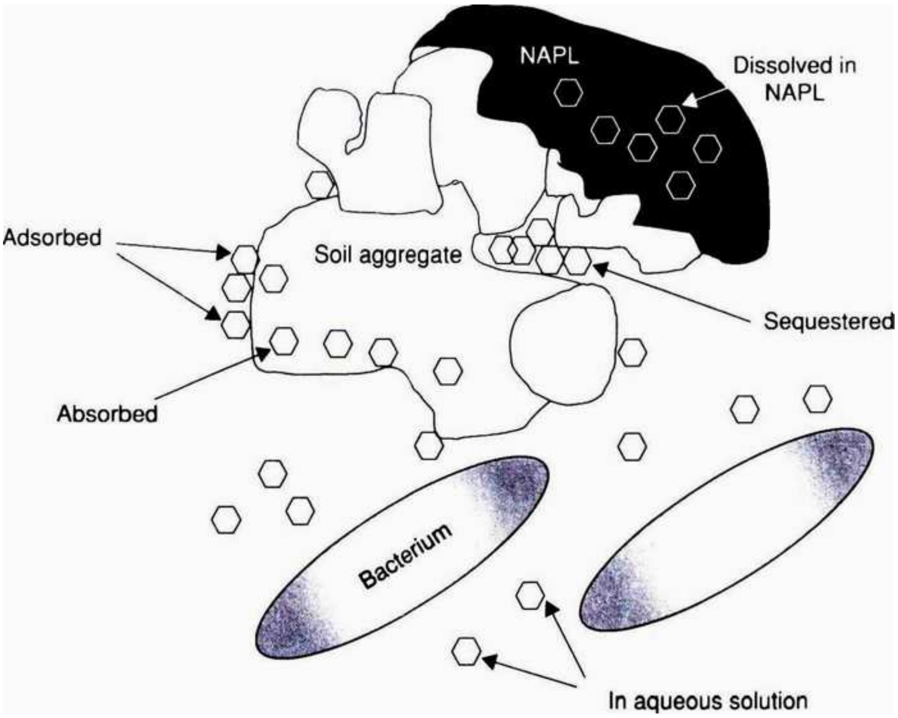


**Figure 5** Representation of different potential states of an organic pollutant (cyclohexane) in a soil environment. (Not drawn to scale.) NAPL, nonaqueous-phase liquid.

as adsorption. Absorption refers to retention within the interior matrix of a solid particle. In either case, a compound partitions from a liquid phase (usually from water) onto or into a solid phase. Typically, the nonspecific term 'sorption' is used to describe this phenomenon. Since compounds must generally be in an aqueous phase and enter the microbial cell to be subject to biodegradation, sorption often increases a pollutant's resistance to biological attack. After a given period of time in a soil, a pollutant establishes equilibrium between the sorbed and the aqueous phase. The extent of sorption at equilibrium, as quantified by the pollutant's sorption coefficient ($K_d$), reflects the characteristics of both the pollutant and the soil matrix. When comparing pollutants of a single structural class (e.g., polycyclic aromatic hydrocarbons, PAHs), those molecules with higher $K_d$ values, which tend to be primarily present in soils in the sorbed phase, generally biodegrade more slowly than those with lower $K_d$ values. In order to enhance pollutant desorption, some bacteria and fungi have been observed to produce extracellular biosurfactants. The release of these surfactant molecules can increase the rate and extent of pollutant desorption and, subsequently, accelerate contaminant biodegradation.

Clay minerals and soil organic matter contain the principle sites to which pollutants can sorb. Charged compounds such as cationic pesticides (e.g., diquat) are often retained by clay minerals. Nonpolar organic molecules, on the other hand, tend to associate with soil organic matter through both adsorption and absorption. For example, PAHs such as phenanthrene, anthracene, and pyrene are subject to sorption and diminished availability as a result of their interactions with humic materials.

**Entrapment within the soil matrix** After their release into soil, some compounds can slowly move from positions within the soil that are accessible to organisms into positions that are inaccessible. These compounds can experience a time-dependent decline in bioavailability termed 'aging.' Nonpolar organic pollutants, for example, can diffuse deeply into hydrophobic regions of soil organic matter and, as a result, evade microbial cells that might otherwise degrade them. Other compounds may diffuse into soil nanopores, thus becoming physically protected from metabolism by microorganisms. Although the details of the mechanisms involved vary, a compound that has moved into remote sites is said to be sequestered in soil.

The degree to which sequestration occurs depends upon a number of variables, including the properties of the soil and the contaminant involved. Although much work needs to be done to isolate the soil components most responsible for the entrapment of pollutants, soil organic matter content appears to be positively correlated with the extent of chemical sequestration. Soil structure also probably plays a key role in the sequestration of some contaminants. Pollutants may be more protected from biodegradation in soils possessing highly developed aggregates than in less structured soils. The chemistry of a contaminant is clearly critical also. Preliminary work has suggested an inverse relationship between a compound's water solubility and its potential for sequestration, but the relative importance of this and other properties remains largely unknown. Pollutants for which sequestration has been observed include the PAHs phenanthrene and anthracene, and various pesticides, such as 1,2-dibromoethane, DDT, and chlordane.

**Nonaqueous-phase liquid** A number of relevant environmental contaminants can be present in soils dissolved in non-aqueous solvents. For example, crude oil is a nonaqueous-phase liquid (NAPL) composed of a number of compounds, many of which have very low water solubility. Since microorganisms are typically active in the aqueous phase only, pollutants in NAPLs can escape biodegradation. The extent to which an individual compound dissolved in a NAPL will be degraded depends upon a number of factors, including the rate of partitioning between the nonaqueous and aqueous phases, the equilibrium concentration of the pollutants in the aqueous phase, the nature of the interface between the two phases, and the toxicity of the NAPL to the microorganisms responsible for degradation. Although individual compounds within a NAPL tend to be degraded at different rates, in general, pollutants in a nonaqueous phase can be extremely persistent.

### Nutrient Availability

Heterotrophic microorganisms in soil require a number of macro- and micronutrients, an energy source, and an electron acceptor to grow. Often, a pollutant provides a carbon or energy source, but all other requirements must be fulfilled if biodegradation is to occur. A biodegradable compound released into a nutrient-poor soil (i.e., low available nitrogen or phosphorus) can persist for an extended period of time, because microorganisms that possess the capacity to metabolize the contaminant cannot appreciably grow. In some instances, as during the biodegradation of the explosives 2,4,6-trinitrotoluene (TNT) and RDX under nitrogen-limiting conditions, specific organisms can degrade an organic compound to obtain a scarce noncarbon nutrient (in this case, nitrogen). Some organic phosphorus-containing compounds, such as triethylphosphate (TEP), can also be utilized by microorganisms to obtain phosphorus under conditions in which the compound is unavailable for microbial growth. During bioremediation of soil contaminants, the addition of nitrogen, phosphorus, or other limiting nutrients is often conducted to stimulate the growth of indigenous or added microorganisms.

### pH

Some bacteria and fungi can survive and grow under very acidic or alkaline conditions, but the optimum pH for most soil microorganisms is near neutrality. Unless organisms have evolved to live under extreme pH conditions, both high and low pH can lead to the denaturing of microbial enzymes and the hydrolysis of other crucial cell components. It is therefore not surprising that biodegradation tends to be fastest near pH 7 and slows considerably as the pH becomes appreciably more alkaline or acidic. In addition to its direct effect on cells, pH will also influence the chemistry of ionizable pollutants in soil. The sorption and subsequent bioavailability of cationic pesticides (e.g., paraquat, diquat), basic pesticides (e.g., atrazine), and acidic pesticides (e.g., 2,4-D, dicamba) will depend in part on the pH of the soil solution in which they are present. To enhance microbial growth and biodegradation, bioremediation protocols sometimes include chemical treatment of the contaminated soil (e.g., by lime addition) to adjust pH to the optimum for the biodegradative organisms.

### Temperature

The growth of microorganisms is highly dependent on temperature, and the prevailing temperature in soil is an important factor controlling the rate and extent to which a pollutant is biodegraded. Most microorganisms in soil are mesophilic, with temperature optima ranging from approximately 25 to 35°C, and growth possible from 15 to 45°C. Therefore, optimal temperatures for biodegradation in soils tend to be within the best growth range for these organisms, with rates increasing as temperatures increase within this range. However, it should not be implied that biodegradation processes cease below 15°C or above 45°C. There are many bacterial species that are psychrophilic, with growth optima less than 15°C, and those that are thermophilic, with growth optima exceeding 40°C and sometimes with extremes as high as 65°C. These organisms can be responsible for contaminant biodegradation in extreme environments such as in polar climates (for psychrophiles) or in compost (for thermophiles). In fact, thermophiles have been applied in numerous *ex situ* bioremediation applications due to their ability to rapidly degrade some contaminants.

### Water Content

The moisture content of soil influences the growth and metabolism of microorganisms. As with the other abiotic factors discussed above, tolerance to desiccation varies among organisms. Environments characterized by low water content tend to support only limited microbial activity. The optimal activity of most microorganisms occurs at soil moistures ranging from 60 to 80% of the soil's water-holding capacity (WHC). Biodegradation can be very limited in dry soils, and pollutants may persist in arid or drought-affected regions longer than in wet areas. Conversely, excessive moisture can limit oxygen diffusion in soil and cause a shift from aerobic to anaerobic conditions. As previously noted, the absence of available oxygen can severely inhibit the disappearance of some pollutants and pesticides, although such conditions may be necessary to promote the degradation of other compounds (e.g., by reductive dechlorination).

### Presence of Appropriate Microorganisms

Microorganisms capable of performing a given biodegradation reaction are frequently indigenous in soil. For example, bacteria that aerobically degrade the common gasoline components benzene, toluene, ethylbenzene, and xylenes (BTEX) are ubiquitous in most surface soils. If these contaminants persist for long times in a given soil, it is likely that environmental factors such as nutrient availability, oxygen levels, contaminant availability, or the presence of toxic cocontaminants are impeding the biodegradation process. For some contaminants, however, biodegradative populations are less universally distributed and/or are present only at very low densities. For example, the gasoline oxygenate methyl *tert*-butyl ether (MTBE) has been observed to biodegrade readily in some aerobic soils and aquifers but to be highly persistent in others. In environments where the contaminant persists, the absence of a biodegrading population has been cited as a key factor limiting its destruction. Bioaugmentation has been used at some sites to promote MTBE biodegradation. Similarly, the absence of a specific dehalorespiring microorganism (*Dehalococcoides ethenogenes*) has recently been cited as the key factor determining the extent of biodegradation of PCE and TCE in some anaerobic environments (see **Figure** 3). If this organism is absent, dechlorination of PCE (or TCE) appears to stall at *cis*-DCE or VC, rather than progressing all the way to ethylene. As with MTBE, bioaugmentation has been used to add this organism to specific sites in order to enhance contaminant biodegradation.

### Predation

Bacteria are subject to predation by protozoa. Such grazing can influence the biodegradation of soil pollutants, with both inhibition and stimulation possible. Predation by soil protozoa can reduce the number of cells metabolizing a contaminant and cause a decline in the rate at which the compound disappears. Alternatively, due to the positive effect predation has on nutrient availability, grazing may stimulate the growth of microorganisms. Through their metabolism of microbial cells, protozoa release available nitrogen, phosphorus, and other nutrients back into soil and, consequently, stimulate the surviving microbial cells to grow faster. The dynamics of predator–prey relationships in soil are complicated and are not completely understood. Whether predation decreases or increases biodegradation rates depends upon a number of factors, including the starting density of the bacterial prey and predator organisms, the size of the cells, and the properties of the soil in which the organisms live.

### Presence of Other Compounds

Microorganisms metabolizing a pollutant can be affected by the presence of other soil contaminants. In particular, it is not uncommon for a cocontaminant (such as a heavy metal or low molecular weight hydrocarbon) to be present in toxic concentrations in a heavily polluted soil, thereby inhibiting the biodegradation of a target compound by killing the biodegradative population. An organism may also preferentially use a compound other than the pollutant as a carbon or energy source and, as a result, biodegradation of the selected contaminant can be slowed. Alternatively, the presence of a second compound can stimulate microbial activity and accelerate biodegradation. For example, bioremediation of a compound that is cometabolized can be enhanced by the addition of a substrate on which the cometabolizing microorganism can grow. The addition of biphenyl to soils has been observed to enhance the rate of degradation of some of the less-chlorinated PCBs. In some instances, however, the growth substrate will actually inhibit degradation of the cometabolized compound owing to competitive enzyme inhibition. The presence of toluene above a given concentration, for instance, can reduce the rates of TCE cometabolism by specific toluene-oxidizing strains. For bioremediation applications, intermittent addition of the growth substrate may be necessary to promote degradation of a cometabolized substrate.

## Summary

Biodegradation is an important process by which chemical pollutants may be removed from soil. Microorganisms have a remarkable ability to metabolize an extremely diverse range of synthetic compounds and can be invaluable in the remediation of contaminated environments. The rate and extent of biodegradation of a given contaminant in soil depends on a number of variables, including chemical structure, contaminant bioavailability, and an array of environmental factors, such as nutrient availability, pH, temperature, soil moisture, predation, and the presence of cocontaminants.

*See also:* **Pesticides**; **Pollutants:** Effects on Microorganisms; Persistent Organic (POPs)

## Further Reading

Alexander M (1999) *Biodegradation and Bioremediation*, 2nd edn. New York: Academic Press.

Alexander M (1981) Biodegradation of chemicals of environmental concern. *Science* 211: 132–138.

Atlas R and Bartha M (1998) *Microbial Ecology: Fundamentals and Applications*, 4th edn. Wokingham, UK: Benjamin/Cummings Science Publishing.

Bollag J-M and Liu S-Y (1990) Biological transformation processes of pesticides. In: Cheng HH (ed.) *Pesticides in the Soil Environment: Processes, Impact, and Modeling*, pp. 169–211. Madison, WI: Soil Science Society of America.

Chaudhry GR (ed.) (1994) *Biological Degradation and Bioremediation of Toxic Chemicals*. Portland, OR: Timber Press, Inc.

Dragun J (1998) Organic chemical biodegradation in soil. In: *The Soil Chemistry of Hazardous Material*, 2nd edn., pp. 664–742. Amherst, MA: Amherst Scientific Publishers.

Focht DD (1997) *Principles of Biodegradation*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Howard PH, Boethling RS, Jarais WF, Meylan WM, and Michalenko EM (1991) *Handbook of Environmental Degradation Rates*. Chelsea, MI: Lewis Publishers, Inc.

Maier RM (2000) Microorganisms and organic pollutants. In: Maier RM, Pepper IL, and Gerba CP (eds) *Environmental Microbiology*, pp. 363–400. New York: Academic Press.

Matsumura F (1982) *Biodegradation of Pesticides*. New York: Plenum Publishing Corp.

Wackett LP and Hershberger CD (2001) *Biocatalysis and Biodegradation: Microbial Transformation of Organic Compounds*. Washington, DC: ASM Press.

# Effects on Microorganisms

**M E Fuller**, Shaw Environmental, Inc.,
Lawrenceville, NJ, USA

## Introduction

Many human activities have resulted in the intentional or unintentional release of pollutants into the soil environment. Some of these compounds are natural or derived from natural sources (i.e., petroleum hydrocarbons, metals), and would hence not be considered pollutants *per se* in their native environment. Their presence in soil is considered to be polluting because they are not native to soil or their quantities are much higher than normally occurs in soils. In contrast, many pollutant compounds are partially or totally man-made (i.e., pesticides, explosives), possessing no naturally occurring analogs.

While many of the pollutants that have been introduced into the soil environment have been shown to be biologically degraded by soil microorganisms, some pollutants have also proven to be toxic to various members of the soil microbial community. The adverse affects of toxic pollutants can be manifested as changes in the structure (composition) of the microbial community, alterations in the functions (activity) of the microbial community, or quite often both.

The effect of pollutants in terms of overall soil 'health' or ecosystem functioning has undergone considerable theoretical and empirical expansion since 1979. A conceptual framework for assessing pollutant effects is presented in **Figure 1**. This framework was developed based on side-effects testing of agricultural pesticides but is widely applicable to all classes of pollutants. The main aspects shown in **Figure 1** are that pollutant effects can range from negligible to critical (compared with a no- or preexposure control) depending on both the magnitude and the duration of the observed deficit in a given soil parameter. The soil parameter can be an aspect of either the microbial community structure (i.e., specific microbial populations) or function (i.e., nitrogen-cycling activities). Under this conceptual framework, large parameter deficits can be acceptable if they are short in duration, and small, long-term parameter deficits can also be acceptable.

## Types of Effects

### Microbial Community Composition

The introduction of a pollutant into the soil ecosystem can result in several types of general and specific changes in the microbial community composition. These effects are summarized in **Figure 2** and are a function of the concentration of the pollutant, the duration of the exposure, and the ability of the compound to be biodegraded.

General If the pollutant is very toxic or is present at very high concentrations, it can reduce the abundance of all microorganisms present in the soil due to cell lysis or death. In contrast, if the pollutant is introduced at a moderate concentration and can also serve as a carbon and/or energy source for soil microbes, an overall increase in the abundance of microorganisms (bacteria, fungi, and protists) may be observed.

The general effects of pollutants on microbial community composition can be assessed using a variety of methods. Systematic laboratory experiments are designed to determine the abundance of total microorganisms before and after exposure to the test pollutant, which is usually introduced into the soil at several concentrations. Field samples collected from inside a contaminated area and from an adjacent uncontaminated area (matched as closely as possible in terms of soil type, vegetation, etc.) also allow the general effects of pollutants to be evaluated.

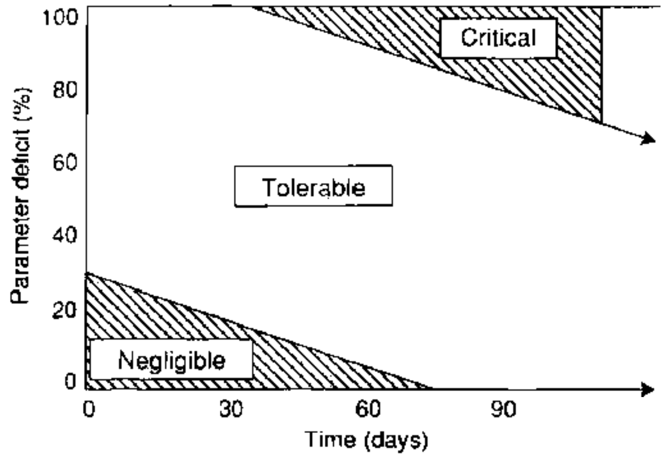


**Figure 1** A conceptual framework for assessing the severity of pollutant effects based on the magnitude and duration of the observed effects. Adapted from Domsch K-H, Jagnow G, and Anderson T (1983) An ecological concept for the assessment of side effects of agrochemicals on soil microorganisms. *Residue Reviews* 86: 65–105, Figure 10, with permission. © 1983, Springer-Verlag, New York, Inc.

Microbial numbers or microbial biomass can be determined by a number of methods, including, but not limited to: spread or pour plate and/or most probable number (MPN) methods, using general microbiological media to enumerate total culturable heterotrophs; epifluorescent direct microscopic counts of total acridine orange- or 4′,6′-diamidino-2-phenylindole (DAPI)-stained cells; chloroform fumigation/extraction procedures to quantify microbial biomass carbon (converted to number of cells based on a factor); extraction and analysis of the total phospholipid fatty acid (PLFA) concentration (converted to number of cells based on a factor); and extraction of nucleic acids combined with quantitative polymerase chain reaction (PCR) techniques using Bacteria-, Fungi-, and Protista-specific primers.

**Specific** Some pollutants exert toxic effects on specific classes of soil microbes. The pollutants may kill or reduce the number of active (vegetative) cells, while spores and other resistant microbes may be relatively unaffected. In many cases, reductions in the abundance of a specific microbial population are due not to direct killing of specific cells, but to inhibition of specific cellular activities and functions to the extent that the cell cannot remain viable or reproduce. Similar to general effects, if the pollutant is present at moderate levels, the abundance of specific

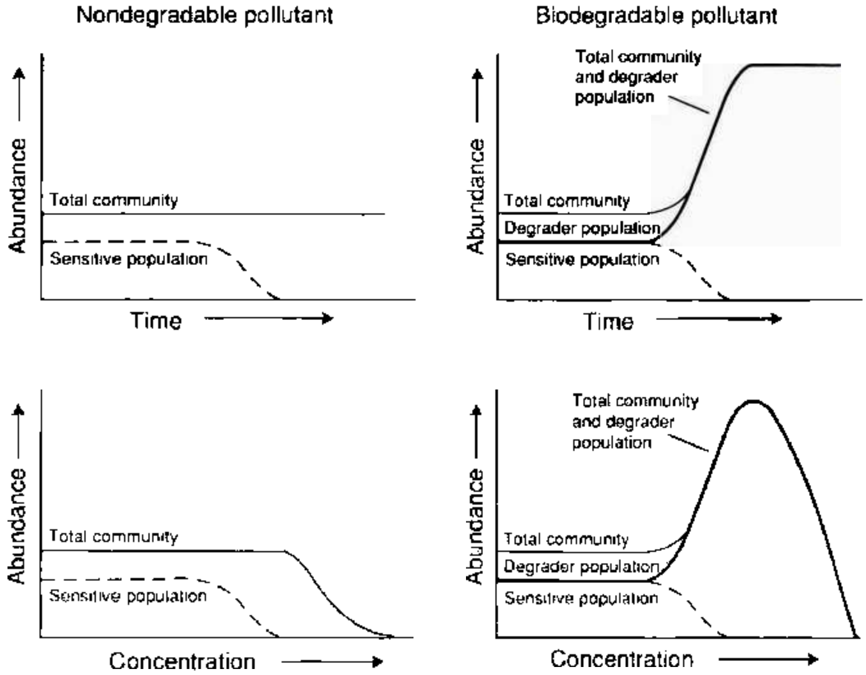


**Figure 2** Expected changes in soil microbial community structure in response to biodegradable and nonbiodegradable pollutants with respect to duration and magnitude of exposure.

microorganisms able to use the pollutant as a carbon and/or energy source may increase substantially compared with their preexposure level.

Effects of pollutants on specific microbial populations are assessed in a manner similar to that described above. Most of the enumeration methods have variants to allow the concentration of specific types of microorganisms to be determined. Plate counts and MPN techniques can be performed using media tailored to certain types of organisms (i.e., purified agar incubated in the presence of hydrocarbon vapors to enumerate hydrocarbon degraders; anaerobic liquid media with nitrate or sulfate as the sole terminal electron acceptor to enumerate nitrate- or sulfate-reducing organisms, etc.). Epifluoresecence microscopy can be done using fluorescently labeled deoxyribonucleic acid (DNA) probes directly to count specific cells of a given microbial strain in contrast to total cells. The profile of different fatty acids that is generated using most PLFA methods can be interpreted to determine the proportion of different microbial populations (Gram-positive versus Gram-negative; sulfate reducers; methanogens; anaerobes, etc.) in a given sample. Lastly, the PCR methods can be adapted for use with genus- and even strain-specific primers/probes, allowing very small subsets of the overall microbial community to be enumerated.

### Microbial Community Function

Pollutants have the ability to decrease (or increase) different activities or functions of the soil microbial community. Some of these changes are in direct response to the pollutant itself, while others may be the indirect result of changes brought about by the pollutant (i.e., changes in pH). In addition, the changes may be in general activities common to almost all microbial cells, or they may be activities that only a small subpopulation of the microbial community is capable of performing. Redundancy in the microbial community (i.e., more than one microbial strain able to perform the same activity) can be a major determining factor for whether the pollutant causes short- or long-term changes in the overall function of the soil microbial community. However, the heterogeneous nature of microbial distribution does limit the potential for redundancy to mitigate the adverse effects of pollutants at the microscale, and is most likely to be observed at the meso- and macro-scale (i.e., soil ecosystem *in toto*).

**General** Carbon mineralization (conversion of carbonaceous material to carbon dioxide) is the most general function of soil microbial communities that can be affected by exposure to pollutants. The majority of microorganisms in most soils are chemoheterotrophic, metabolizing available organic material as their carbon and energy source for growth and cell maintenance. Pollutants can interfere with carbon mineralization by direct interaction with specific degradative enzymes, but often act by disrupting or decoupling the cellular systems needed for energy generation and flow (i.e., production of adenosine triphosphate (ATP), regeneration of nicotinamide adenine dinucleotide (phosphate) (NAD(P)), and flavine adenine dinucleotide (FAD), etc.). Biodegradable pollutants will generally increase the net carbon mineralization (respiration) of a soil because of overall increases in microbial activity and microbial cell densities. However, nonbiodegradable pollutants can also cause increased respiration by triggering stress responses in the microbial community whereby more energy (and hence, carbon throughput) is needed to cope with the pollutant stressor.

Changes in carbon mineralization are usually determined by measuring the basal respiration of soil in the presence of the pollutant and comparing it with the same soil without the pollutant present. In many cases, substrate-induced respiration is also measured by adding a labile carbon source such as glucose and measuring the resulting respiratory production of carbon dioxide by soil in the presence and absence of the pollutant. Any pollutants (or pollutant concentrations) that cause significant decreases in carbon metabolism are also likely to affect adversely the input of fresh carbon (i.e., plant root exudates, plant biomass), reducing overall carbon cycling and raising concerns about long-term soil health. Finally, if the pollutant is biodegradable, its effects will probably decrease over time as its concentration is reduced.

**Specific** The cycling of nitrogen in the soil is by far the most studied microbial activity with respect to pollutant effects. The nitrogen cycle is shown in **Figure 3**. The consensus that has emerged from decades of research is that ammonification and denitrification activity is widely distributed throughout the soil microbial community, while ammonium oxidation, nitrate oxidation, and nitrogen fixation activity is concentrated in only a few microbial genera. This underlies the observation that nitrification, and especially ammonium oxidation, is very sensitive to many types of pollutants. Two bacterial genera, *Nitrosomonas* and *Nitrobacter*, are believed to be the predominant autotrophic nitrifying populations in soil, although some heterotrophic nitrifiers also exist. These two strains are highly sensitive not only to pollutants, but also to a range of environmental conditions, especially pH. Nitrification, therefore,

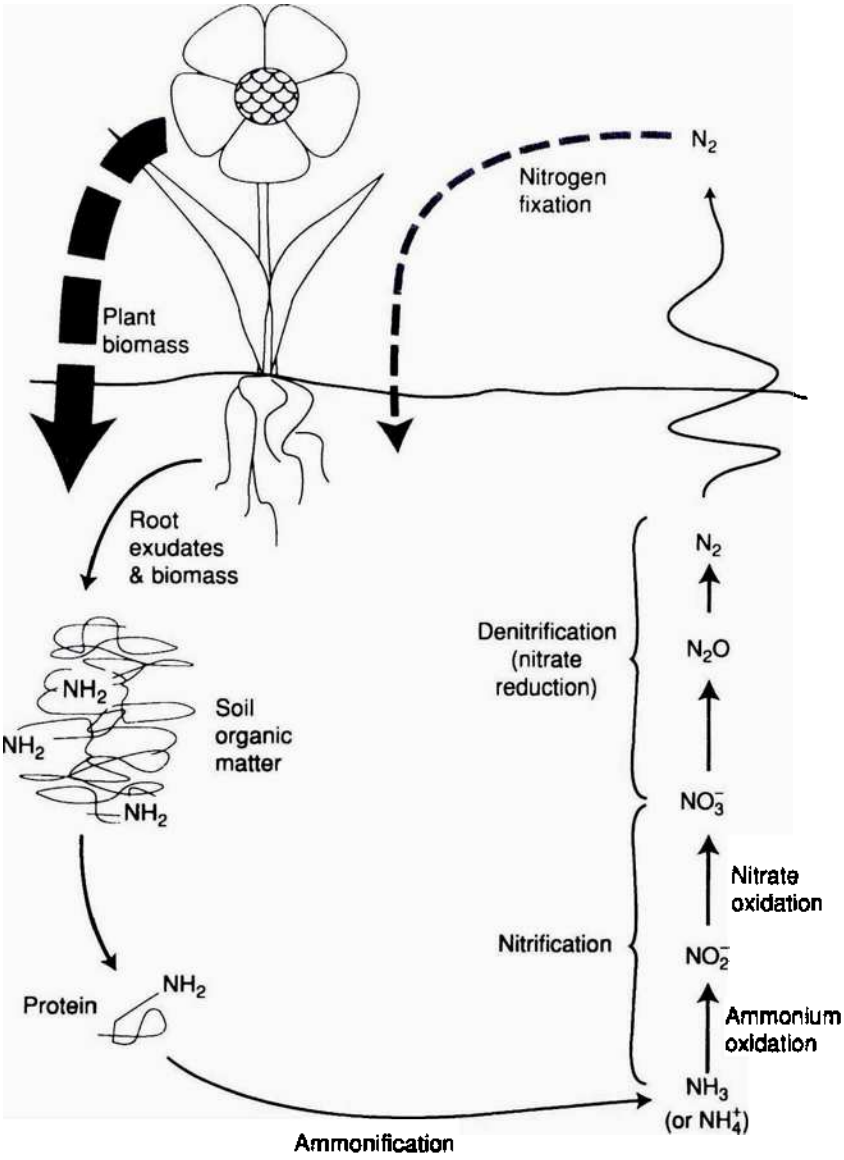


**Figure 3** Soil nitrogen cycle.

may be thought of as a 'sentinel' soil activity, serving as the first indicator of soil ecosystem disturbance by pollutants. Although nitrogen-fixation ability is limited to a few symbiotic and free-living bacterial genera, nitrogen-fixing activity *per se* seems to be less sensitive to most pollutants than nitrification.

The effects of pollutants on the microbial activities associated with the cycling of other nutrients such as sulfur, phosphorus, etc., have not been extensively studied. This is probably due to the perceived lesser importance of these elements compared with carbon and nitrogen to the overall health of the soil microbial community. It has been assumed that these elements are present in excess what is needed by the soil biota, while carbon, and more so nitrogen, are the limiting nutrients. Because the activities involved in the cycling of phosphorus, sulfur, etc., appear to be widely distributed within the microbial community, only in extreme cases would dramatic pollutant effects be expected to be observed (i.e., massive decimation of total microbial biomass).

A final category of specific effects of pollutants relates to changes in the activities of soil enzymes. Some of the more generic enzymes include: dehydrogenase, acid and alkaline phosphatase, phosphodiesterase, arylsulfatase, amidases, urease, and protease. Many soil enzymes are released from microbial cells into the open soil environment, with the goal of liberating the soil nutrients associated with soil organic matter (SOM). Whereas disruption of carbon mineralization (respiration) is a general effect of pollutants on soil microbial community function, the effects of pollutants on soil enzymes disrupt other necessary aspects of carbon cycling in soils (i.e., conversion of large, extracellular organic molecules to smaller molecules available for cellular uptake and metabolism).

Several reports have focused on the effects of pollutants on soil enzymes, but the exact nature of the effects have been less-rigorously investigated (i.e., are the pollutants directly interacting with the enzymes, or with the organisms that are producing and releasing the enzymes?). Again, due to the relatively widespread and generic activities of these enzymes, pollutant effects seem to be relatively small except in cases of excessive and/or prolonged contamination. Generally, dehydrogenase and urease have been shown to be more inhibited by pollutants, while phosphatase has often been observed to be stimulated.

## Effects of Different Classes of Pollutants

The effects that have been observed to be caused by different types of soil pollutants are presented in Table 1. This table, and the discussion in this section, reflects the general consensus of the current literature and hence should be considered as a guide regarding potential pollutant effects, in contrast to a comprehensive summary of all the available data. Indeed, in many cases the published literature presents contrasting results, pointing to the need for continued research with more standardized (and environmentally relevant) experimental procedures. Many of the effects are highly dependent on other physiochemical soil parameters (i.e., pH, redox, cation exchange capacity, organic matter, or clay content) which modulate the bioavailability of the pollutants. Information on the effects of a specific compound should be sought out in the peer-reviewed literature.

### Petroleum Products

The effects of petroleum-based products, ranging from short, straight-chain hydrocarbons to highly condensed polynuclear aromatic hydrocarbons (PAH), have been studied largely in response to the reality of widespread pollution by these compounds due to accidental releases at gas stations and oil refineries. However, there are considerable 'natural' releases of these compounds into the environment from seeps (e.g., the La Brea tar pits in California, USA; undersea seeps in the Gulf of Mexico; etc.).

The most general effect of petroleum products is an increase in microbial biomass and changes in the community structure due to utilization of the compounds as carbon and energy sources (Table 1). Respiration is also usually increased, but effects on other activities are minimal.

### Halogenated Compounds

The compounds in this class include compounds used in chlorinated degreasing and dry-cleaning such as perchloroethylene (PCE) and trichloroethylene (TCE), polychlorinated biphenyls (PCBs), and other halogenated industrial chemicals. Compounds containing chlorine and bromine have been more closely studied than fluorinated and/or iodinated chemicals.

Halogenated compounds enter the environment through accidental releases from industry, leaking waste drums, and landfill leachate. Very high levels of contamination usually coalesce into the groundwater below the soil as dense, nonaqueous-phase liquids (DNAPL), but vapors from these areas may still rise up through the subsurface and affect the soil community. In general, very few effects are observed from contamination with moderate levels of halogenated aliphatic compounds (Table 1). However, chlorinated aromatic compounds have been observed to affect some aspects of the microbial community structure (biomass and diversity) and function (especially nitrification).

### Pesticides

Pesticides (including fungicides, nematocides, insecticides, and herbicides) are the only class of compounds that are intentionally introduced into the environment on a regular basis. The use (and misuse) of pesticides, the development of new pesticides, and the registration of new pesticides (particularly in Europe), has been one of the primary driving forces behind the study of the effects of pollutant on soil microorganisms. The major focus for pesticides has been on 'side-effects testing,' during which the effects of the compound(s) on soil organisms other than the target 'pest' organism are examined. Initially, broad measurements of general effects on soil functions that were deemed important indicators of overall soil health (especially from an agricultural perspective) were studied. More recent work, however, has employed molecular methods to examine effects down to the level of single microbial species and how pesticides may influence long-term soil health.

In a broad study of more than 50 pesticides, the recommended agricultural doses caused no long-term effects on the structure or function of the soil microbial community. Interestingly, when effects are observed, it is often attributed to the pesticide breakdown products rather than the parent compounds.

### Explosives

These compounds have only recently been recognized as potentially significant pollutants. The major compounds in this class include: the nitroaromatics 2,4,6-trinitrotoluene (TNT), 1,3,5-trinitrobenzene (TNB), 2,4,6-trinitrophenol (picric acid), and N-methyl-N,2,4,6-tetranitroaniline (tetryl); the nitramines hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX,

**Table 1** Summary of effects of different classes of pollutants on soil microorganisms

| | *Parameter affected* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Community structure* | | *Community function* | | | | | | | |
| *Pollutant class* | *Total biomass* | *Specific populations* | *Respiration* | *Nitrogen fixation* | *Ammonification* | *Nitrification* | *Denitrification* | *Soil enzymes* | *Sulfur cycling* | *Phosphorus cycling* |
| *Hydrocarbon* | | | | | | | | | | |
| Aliphatic | + | + | + | None/ | ? | – | ? | ? | ? | ? |
| Aromatic | – | + | + | None | None/– | | None/ | None | ? | ? |
| PAH | None/+ | + | – | None | ? | | ? | ? | ? | ? |
| *Halogenated* | | | | | | | | | | |
| Aliphatic | None | None | None | ? | None | | None | ? | ? | ? |
| Aromatic | /None/+ | + | – | ? | ? | ? | | ? | ? | ? |
| *Pesticides*[a] | None | None/+/– | None/– | None/+/– | None/– | None/ | – | ? | /None | None |
| *Explosives* | None/ | + | None/– | – | ? | +/ /None | ? | ? | ? | ? |
| *Metals* | –/None | +/–/None | +/ | –/+ | ? | – | + | – | | |

[a] Applied at recommended rates for agricultural use.

+, increase in parameter; –, decrease in parameter; ?, no or not enough data available; PAH, polyaromatic hydrocarbon.

and octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocine (HMX); and the nitroesters nitroglycerin (NG), nitrocellulose (NC), and pentaerythritol tetranitrate (PETN). Gross pollution with explosives is most often associated with manufacturing and training activities, and historical soil contamination for 50 years or longer has been reported at many military installations where these activities have occurred.

The toxic effects of these compounds in the environment have only recently been initiated, and reports of effects on soil (micro)organisms are sparse. Most of the explosives are sparingly soluble and, when mixed with the binders or waxes used in actual explosives formulations, their bioavailability is quite limited. The summary in Table 1 indicates that some effects have been observed, but more research is clearly needed.

### Metals

Metals enter the soil environment primarily through atmospheric deposition downwind of smelting operations or via leaching from mine spoils. Some metals are essential for microbial cells, some are needed in trace quantities, and others are nonessential and/or toxic. The most toxic metals are the heavy metals cadmium, lead, and mercury, but high concentrations of copper, zinc, chromium, iron, manganese, nickel, cobalt, arsenic, and silver can also exert adverse affects on microorganisms.

The toxicity can be manifest by replacement of essential metal reaction centers of enzymes, reaction of metal ions with the sulfhydryl and phosphate groups of proteins, or displacement of key structural elements. Metal toxicity is highly dependent on the pH and redox state of the soil, as these strongly influence the oxidation state and solubility of metals and hence their bioavailability to soil microbes. Microorganisms themselves have also developed a broad range of resistance mechanisms to counteract the effects of metals. These include facilitated transport of the metal ions out of the cell, or complexation of the metal with reactive molecules (glutathione, etc.) that render it less bioavailable.

*See also:* **Bacteria:** Soil; **Biodiversity; Carbon Cycle in Soils:** Dynamics and Management; **Enzymes in Soils; Fungi; Heavy Metals; Hydrocarbons; Microbial Processes:** Community Analysis; **Nitrogen in Soils:** Cycle; **Organic Residues, Decomposition; Oxidation–Reduction of Contaminants; Pesticides; Phosphorus in Soils:** Biological Interactions; **Pollutants:** Biodegradation; Persistent Organic (POPs); **Pollution:** Industrial; **Quality of Soil; Remediation of Polluted Soils; Sorption:** Metals; Organic Chemicals; **Sulfur in Soils:** Biological Transformations; **Waste Disposal on Land:** Municipal

## Further Reading

Domsch K-H, Jagnow G, and Anderson T-H (1983) An ecological concept for the assessment of side effects of agrochemicals on soil microorganisms. *Residue Reviews* 86: 65–105.

Hicks R, Stotzky G, and van Voris P (1989) Review and evaluation of the effects of xenobiotic chemicals on microorganisms in soil. *Advances in Applied Microbiology* 35: 195–253.

Lal R (1982) Accumulation, metabolism, and effects of organophosphorus insecticides on microorganisms. *Advances in Applied Microbiology* 28: 149–200.

Lal R and Saxena D (1982) Accumulation, metabolism, and effects of organochlorine insecticides on microorganisms. *Microbiological Reviews* 46: 95–127.

McGrath SP, Chaudri AM, and Giller KE (1995) Long-term effects of metals in sewage sludge on soils, microorganisms and plants. *Journal of Industrial Microbiology* 14: 94–104.

Nies DH (1999) Microbial heavy-metal resistance. *Applied Microbiology and Biotechnology* 51: 730–750.

US Environmental Protection Agency (1996) *Terrestrial (Soil-core) Microcosm Test.* Ecological effects test guidelines. OPPTS 850.2450. Washington, DC: Government Printing Office.

US Environmental Protection Agency (1996) *Soil Microbial Community Toxicity Test.* Ecological effects test guidelines. OPPTS 850.5100. Washington, DC: Government Printing Office.

van Straalen NM (2002) Assessment of soil contamination – a functional perspective. *Biodegradation* 13: 41–52.

# Persistent Organic (POPs)

**D Johnson**, Rothamsted Research, Harpenden, UK

## Introduction

Organic pollutants are ubiquitous environmental contaminants and, for many of the compounds, the environmental burden has risen, largely as a result of anthropogenic activity in the nineteenth and twentieth centuries. Soil can be a major environmental sink for many of the organic pollutants released into the atmosphere. Once in a soil, nonionic organic pollutants can often be highly persistent. The primary rate-limiting factors determining the long-term persistence of persistent organic pollutants (POPs) in soil are thought to be fundamental sorption/desorption mechanisms that control the distribution of chemical between the solid and aqueous phases of the soil. Soil POPs are potentially available to many organisms, including soil microflora, macrofauna (earthworms are particularly important), and grazing animals. However, the availability of soil POPs tends to decline with time. The bioavailability of POPs to soil organisms is important in determining the potential hazard presented by a contaminated soil and may also influence the efficiency of bioremediation of contaminated soils.

## Persistent Organic Pollutants

During the nineteenth and twentieth centuries humans have produced and released vast quantities of organic compounds into the biosphere. These are widely distributed throughout the environment and many are thought to be carcinogenic and mutagenic. This has led to the prioritization of organic pollutants by the United States Environmental Protection Agency (USEPA) and the European Community (EC). These lists include polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), polychlorinated dibenzo-*p*-dioxins and furans (PCDD/Fs), chlorobenzenes (CBs), phthalate acid esters, and other miscellaneous compounds. These compounds enter the environment through a wide variety of processes including inefficient combustion, dry and wet deposition from the atmosphere, widespread pesticide use, and wastewater drainage from industrial and domestic sources into sewers. Hence, the soil is a major sink for POPs. Once in a soil, high-molecular-weight non-ionic organic compounds may persist for long periods.

POPs are characterized by their relatively high octanol/water partition coefficient (log Kow) and low water solubility. Thus, once in a soil they tend be resistant to microbial degradation and prone to bioaccumulation. Atmospheric transport processes such as 'global distillation' result in an almost ubiquitous contamination of soils, often thousands of kilometers from major POP sources. The highest soil concentrations of POPs are usually found near industrial point sources, urban areas, and in organic-rich soils such as those that have received sludge amendmen. The recent decline in the manufacturing industry in many western economies has left large areas of soil contaminated by POPs and requiring remediation before they can be redeveloped.

Legislation relating to organic compounds in soil is largely based on toxicological criteria with the ultimate aim of protecting wildlife and human health. Thus, it is important that the 'availability' of compounds to transfers from the soil to the organisms is accurately understood in a quantitative manner. Recent research on the fate and behavior of organic chemicals in soils suggests that the availability of an organic compound from the soil into the food chain may be highly dependent upon the form in which a compound is present in a soil. This is influenced by a number of factors, including its residence time within the soil, physicochemical properties of the compound and soil, and how it entered the soil. Desorption of sorbed compounds into the soil solution enhances biodegradation, leaching, volatilization, plant uptake and other chemical/photochemical transformations. Whilst these processes are generally enhanced by the chemical being present in the bulk soil solution, chemical degradation, photochemical transformations and biodegradation of soil-sorbed compounds are also possible. Furthermore, soil-sorbed compounds may also be lost from areas of soils by the erosive action of wind or running water, uptake by invertebrates or direct ingestion of soil by grazing livestock.

The persistence and relative significance of each loss pathway in a given soil depends on the physicochemical properties of the organic chemical and environmental conditions. However, regardless of the main loss pathway, the overall loss of organic chemicals from soils is often biphasic. This means that a short period of rapid initial loss is most often followed by a longer period of slow dissipation. Loss processes in soils exhibit similar patterns. The primary rate-limiting factors governing this biphasic behavior are thought to be fundamental sorption/desorption mechanisms that control the distribution of chemical between the solid and aqueous/gaseous phases of soil and hence, the supply of chemicals available to the various loss pathways. The availability of organic compounds in a soil will affect both the toxicological threat and the potential to remediate a contaminated site. Further details relating to specific technologies available for contaminated soils can be found in the article Remediation of polluted soils (*See* **Remediation of Polluted Soils**).

## Physicochemical Properties of Persistent Organic Pollutants in Soils

The principal persistent organic pollutants found in soils are PAHs, PCBs, CBs, and PCDD/Fs. **Table 1** displays comparative physicochemical data for a number of selected organic chemicals that occur in contaminated soils.

## Persistence and Aging Behavior of Organic Chemicals in Soils

Figure 1 illustrates the decay curves of three model nonionic organic compounds after they enter a soil. The persistence and overall loss of organic chemicals from soils following their entry into a soil are dependent on the physicochemical properties of the soil and the chemicals themselves, environmental

**Table 1** Physicochemical properties of selected chlorobenzenes, polycyclic aromatic hydrocarbons, polychlorinated biphenyls, and polychlorinated dibenzo-*p*-dioxins and furans

| *Compound* | *Log Kow* | *Log Koc* | *K′ H* | *Log Koa* |
|---|---|---|---|---|
| 1,4-DCM | 3.5 | 3.43 | $6.46 \times 10^{-2}$ | 4.69 |
| HCB | 5.5 | 5.25 | $2.72 \times 10^{-2}$ | 7.07 |
| Naphthalene | 3.37 | 3.14 | $1.74 \times 10^{-2}$ | 5.13 |
| Phenanthrene | 4.57 | 4.4 | $1.31 \times 10^{-3}$ | 7.45 |
| Benzo(a)pyrene | 6.04 | 6.24 | $1.86 \times 10^{-5}$ | 10.77 |
| Coronene | 6.75 | 6.4 | $1.72 \times 10^{-7}$ | 13.51 |
| 2,3,7,8-TCDD | 6.8 | 6.2 | $1.35 \times 10^{-3}$ | 9.67 |
| OCDD | 8.2 | 6.97 | $2.76 \times 10^{-4}$ | 11.76 |
| PCB-18 | 5.72 | 5.15 | $1.95 \times 10^{-2}$ | 7.43 |
| PCB-52 | 5.83 | 4.99 | $1.75 \times 10^{-2}$ | 7.59 |
| PCB-101 | 6.3 | 6.3 | $8.74 \times 10^{-3}$ | 8.36 |
| PCB-138 | 6.69 | 6.24 | $1.00 \times 10^{-3}$ | 8.69 |
| PCB-206 | 7.94 | 6.66 | $5.59 \times 10^{-3}$ | 10.19 |

log Kow, octanol/water partition coefficient; log Koc, octanol/carbon coefficient; K′ H, Henry's constant; log Koa, octanol/air coefficient.

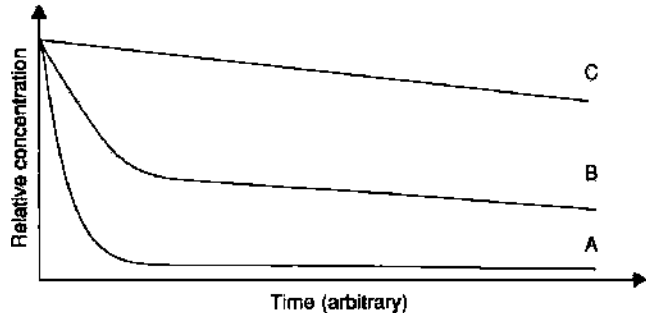


**Figure 1** Hypothetical loss profiles for organic chemicals in a soil. A–C are discussed in the text.

factors, including temperature and precipitation, and anthropogenic factors such as cultivation and drainage.

Curve A (**Figure 1**) is typical of organic contaminants with short half-lives that exhibit relatively rapid dissipation with most of the residues ultimately being removed from the soil. Such decay curves are typical of compounds that are volatile, water-soluble, or easily degraded (or a combination of these). Examples of this chemical type are the lighter chlorinated benzenes and volatile aromatic hydrocarbons. At the other extreme, chemicals that are nonvolatile, relatively water-insoluble and recalcitrant are characterized by persistence profile C (**Figure 1**). Such chemicals are strongly sorbed and persist in soils over often extremely long periods. Examples include PCDDs and PCDFs. Most organic compounds fall within these two extremes, exhibiting persistence profiles similar to curve B. Curves A and B are essentially similar in shape with an initial period of rapid loss followed by a much slower dissipation of the chemical. This leads to a 'persistent residual fraction' of the compound forming in the soil, the size of which will be determined largely by the physicochemical properties of the compound and soil. As commented earlier, a loss profile of this shape is often referred to as 'biphasic' and is typical of organic chemicals with intermediate sorption potentials.

## The 'Aging' of Persistent Organic Pollutants Within Soils

As can be seen in **Figure 2**, when a nonionic organic compound enters a soil it undergoes a number of loss processes including leaching, volatilization, chemical and biological degradation, and plant uptake. In addition, sorption- and transport-related processes mean that an increasing component of the soil-borne compound becomes 'irreversibly bound' within the soil matrix, i.e., they cannot be extracted by conventional analytical techniques. The result is that, as the residence times of organic compounds increase within a soil, they become increasingly difficult to extract. This phenomenon has become known as the 'aging effect.' There is also an assumption that there is an analogous decline in a compound's bioavailability as soil residence time increases and extraction efficiency declines, but this may be erroneous (see later).

**Figure 2** demonstrates that this process of declining extractability of soil-bound compounds is, in part, due to the formation of irreversibly bound residues within the soil matrix. The development of irreversibly bound residues is a major contributing factor in the formation of persistent residual fractions of organic chemicals in soils. Addressing this issue is complicated by the fact that the experimental data are highly dependent on the extraction method used and the knowledge that the efficiency of extraction of an organic chemical from a soil will decline with an increase in contact time between a chemical and the soil, i.e., the 'aging effect.'

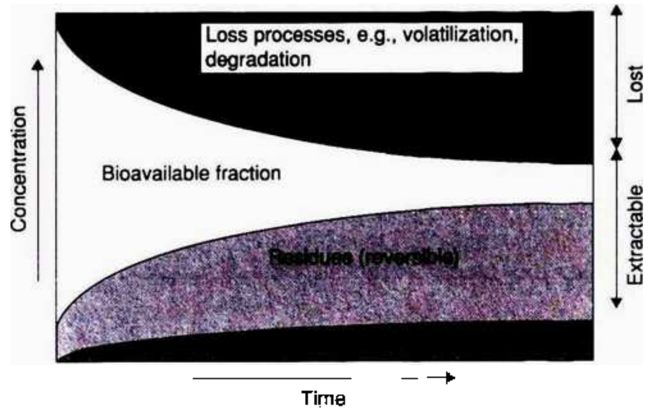


**Figure 2** Dynamics of organic chemical persistence, bioavailability and loss in soils.

## Mechanisms Contributing to the Persistence of Organic Chemicals in Soils

Several groups of mechanisms are thought to be involved in the 'aging effect'. The processes involved can be split into transport-related and sorption-related mechanisms. Transport-related mechanisms are those where sorption/desorption of organic chemicals is controlled by water flow in soils. Due to their particle size, distribution, and management, soils may develop distinct structural arrangements ranging from simple homogeneous granular sands to pronounced columnar and prismatic clay structures. In the latter soil type, swelling and shrinkage may create a network of cracks that rapidly conduct most water while the remainder of the water moves relatively slowly through the soil matrix. Water flow in sandy soils is more uniform, but it is important to recognize that preferential flow may also occur in such soils. Comprehensive reviews of water movement in structured soils and the implications of this for solute transport and fate will not be discussed in detail here but readers requiring more information can find further reading at the end of this article.

### Sorption-Related Mechanisms

The nature and significance of sorption-related mechanisms are less well understood than transport-related mechanisms and much research is still required in this field. The literature still contains many contradictions as to what are the dominant sorption-related mechanisms controlling the persistence of nonionic organic chemicals in soils. Sorption mechanisms can be divided into two distinct classes: diffusive mass transfer and chemisorption. Diffusive mass transfer mechanisms include film diffusion, pore diffusion, retarded intraparticle diffusion, and intraorganic matter diffusion. Film diffusion relates to the resistance encountered by an organic chemical when crossing a thin film of water surrounding soil solids/aggregates, or when moving from the solid phase into the bulk solution or gaseous phases. Film diffusion is generally regarded as insignificant in comparison to other processes discussed here. Pore diffusion is the molecular diffusion of organic chemicals within pores between soil particles inside soil aggregates. The current literature suggests that, similar to film diffusion, pore diffusion rarely contributes to the formation of persistent chemical residual fractions.

Chemisorption occurs when a functional group of an organic chemical interacts with reactive moieties on soil solids through the formation of strong chemical bonds, principally covalent bonds. Such interactions give rise to the formation of a fraction that is irreversibly sorbed, despite exhaustive extraction procedures, and has become known as 'bound residues' in the pesticide science field. However, pesticides are often ionic, in contrast to nonionic compounds such as nonionic persistent organic pollutants, which are less likely to form chemical bonds with soil solids. Thus chemisorption is unlikely to be a major rate-limiting factor governing the persistence and loss of hydrophobic organic chemicals in soils. Where a contaminant has a tendency to ionize or contains polar functional groups, then chemisorption may be important. In addition, biologically mediated degradation processes may alter the structure of nonionic organic chemicals so they are more susceptible to chemisorption. A 1996 study demonstrated that once partial oxidation of a PAH occurs by biologically mediated degradation, the irreversible chemisorption of metabolites can occur and in some cases can be so significant that the process becomes competitive with biodegradation. The significance of chemisorption in the formation of persistent residual fractions has yet to be fully elucidated. Further investigation is clearly necessary because the exact processes involved in persistent residual fraction formation may be important in determining whether the compounds remain chemically and, more importantly, biologically unavailable.

### Sorption Retarded Intraparticle Diffusion

Of the mass transfer mechanisms involved in the formation of persistent residual fractions it is generally agreed that sorption retarded intraparticle diffusion (SRPD) and sorption retarded intraorganic matter diffusion (SROMD) are the two primary rate-limiting factors. SRPD has been defined as 'aqueous-phase diffusion of solute within pores of microporous particles (e.g., sand grains) mediated by retardation resulting from instantaneous sorption to pore walls.' However, SRPD is often used as an umbrella term for a number of similar processes that occur by diffusion of organic compounds through pores. The above definition of SRPD refers to the molecular diffusion of chemicals in pore water that is retarded, chromatographic-like, by local sorption on pore walls. This process is more marked in larger mesopores (diameter >2 μm), and the sorption may or may not be aided by sorption to and within organic matter lining these pores (see later for description of SROMD). As the pore's diameter decreases, the tortuosity of the pore tends to increase, further hindering diffusion processes.

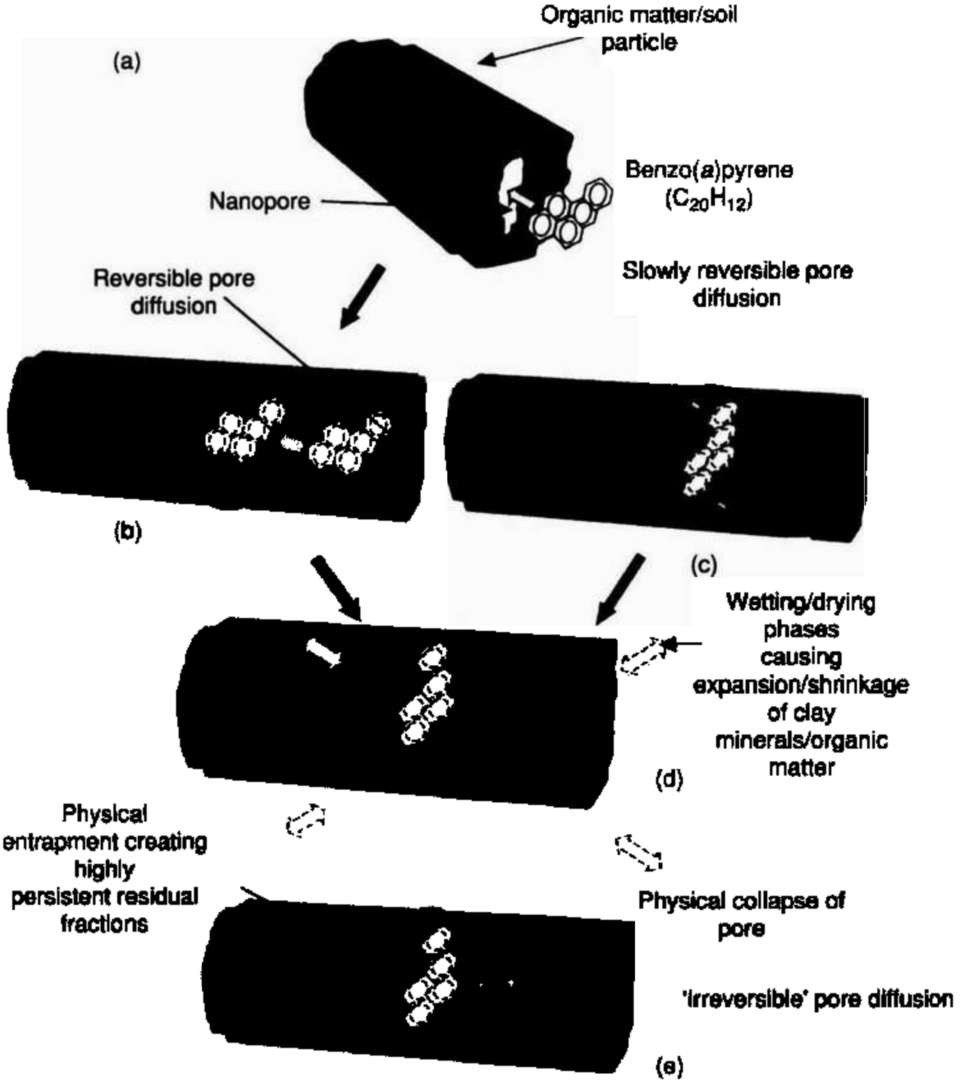


**Figure 3** A conceptual diagram showing the mass-transfer limited processes leading to sorption-retarded pore diffusion. (a)–(e) are discussed in the text.

**Figure 3** graphically illustrates the various stages of diffusion that occur in exceptionally narrow nanopores traversed by a model PAH, benzo(*a*)pyrene. Compounds tend to enter nanopores as they diffuse along gradually narrowing mesopores ($>2\,\mu$m) followed by micropores ($<2\,\mu$m) and eventually end up in nanopores whose diameter may be only a few Angstroms larger than the molecule itself (**Figure 3a**). In extremely narrow pores, diffusion is often exceptionally slow due to the high viscosity of pore water. Polar minerals have one or more layers of water strongly sorbed on their surfaces. This ordered layering of molecules increases the viscosity of the water and greatly slows molecular diffusion within nanopores (**Figure 3b**). The water contained in very narrow pores of a few Angstroms may be 'ice-like' and thereby greatly restrict solute diffusion. The tortuosity of pores becomes an increasingly important factor in the retardation of pore diffusion and tortuosity may only influence desorption when the pore diameter is less than 6 nm. In extremely narrow and tortuous pores, steric hindrance may further retard desorption (**Figure 3c**). When this occurs sorption is only very slowly reversible, contributing to the persistent residual fractions discussed earlier. Natural wetting/drying phases of a field soil will lead to expansion and contraction of clay mineral lattices (**Figure 3d**), which can induce physical collapse of pores and physical entrapment of solute molecules within 'blind alleys' (**Figure 3e**).

**Figure 3** therefore represents a series of steps that progressively restrict solute diffusion and retard desorption leading to persistent residual fractions of organic compounds in soils. These compounds, although chemically unaltered, are usually unavailable to traditional chemical extraction techniques and the assumption is that they are also not bioavailable. However, it is important to remember that, although such compounds are often referred to as irreversibly bound, this may not always be the case. Physical destruction of the soil aggregates they are associated with may release the compounds or at least increase their biological and chemical availability. Acidification dissolves the inorganic oxide cements that hold soil aggregates together and hence may similarly increase rates of pore diffusion and sorption.

### Intraorganic Matter Diffusion

Intraorganic matter diffusion (IOMD) is the other important rate-limiting process involved in the desorption of persistent organic compounds from soils. Soil organic matter can exist as surface coatings, lining pores, or as discrete particles. Organic matter diffusion may be the primary rate-limiting process in soils with high-organic-matter contents but, since slow sorption also occurs in soils with low to zero organic matter, it clearly is not the only rate-limiting process. IOMD is considered to be diffusion solely within the matrix of organic matter (OM), which is thought to resemble a polyelectrolyte polymer mesh, with retardation resulting from reversible sorption on the internal surfaces of organic matter. OM polymers are said to have glassy (condensed, rigid) or rubbery (expanded, flexible) structures (**Figure 4**). Diffusion in polymers occurs by either a place change mechanism, in which movement is accomplished by cooperative interchange of position in polymer segments and the penetrating molecule, or by a defect mechanism where the penetrant may jump between lattice defects, voids, and pores. OM polymers consist of condensed, rigid 'glassy' phases through which diffusion is slow (**Figure 4b**) or more expanded and flexible 'rubbery' structures through which diffusion is faster (**Figure 4a**).

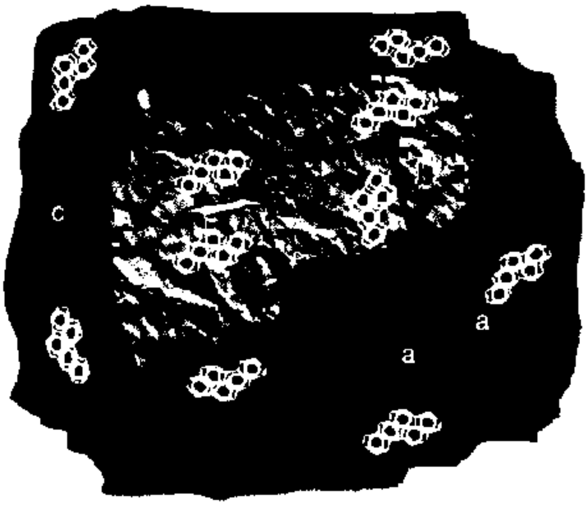


**Figure 4** A conceptual diagram describing the slow sorption processes involved in organic matter diffusion (OMD). (a) Diffusion through a rubbery phase; (b) diffusion through a more condensed glassy phase b; and (c) adsorption in a 'Langmuir site' (see text for further details).

In addition, some organic compounds may undergo 'Langmuir-like' adsorption (**Figure 4c**) in internal voids in the OM matrix, which has been likened to a hole-filling mechanism. These slow diffusion processes lead to the formation of physically protected compounds which, as in the case of SRPD, are often mistakenly labelled as irreversibly bound residues. The OM diffusion and binding processes are in fact transitory since environmental variables may swell or soften the OM matrix, converting it from a glassy to a more rubbery state and accelerating diffusion processes. Increasing temperatures, change in pH, and the presence of cosolvents can all act to increase the diffusivity of tightly bound compounds.

In summary, it is probable that a number of mass transfer mechanisms are responsible for retarded diffusion leading to the formation of persistent residual fractions within most soils. Of these, the two most important are usually SROMD and SRPD. Although much work is needed to determine the exact contribution of each mechanism, it is likely that in most soils it is a synergy of both OMD and SRPD that is primarily responsible for the formation of persistent residual fractions. Residual fractions are often considered both biologically and chemically unavailable. However, as the preceding passages have shown, the mechanisms that lead to their formation are often transitory. This is particularly important when considering biological availability since the very conditions which are known to enhance diffusion rates often occur in the intestines of biological organisms (low pH, mechanical manipulation, and elevated temperatures). Hence, soil ingestion by an organism may have an impact upon a compound's availability and traditional extraction techniques used to determine the risk presented by a persistent residual fraction may be inaccurate.

## The (Bio)availability of Organic Chemicals in Soils

The term 'bioavailability' is often discussed in the literature but rarely defined. For the purposes of this article, it may be defined as 'the fraction of a chemical that is absorbed by an organism.' This is a broad definition and encompasses bioavailability to plants and microorganisms via a variety of potential routes. The amount of compound present in an organism will be a function of the rate of absorption versus the rate of excretion.

POPs in soils are potentially bioavailable to many different organisms. In this article, a very brief overview is provided of the bioavailability of POPs to those organisms that are thought to be most important. These are plants, bacteria, and fungi, and grazing animals. Macroinvertebrates and particularly earthworms are dealt with in more depth as they may represent important routes for POPs to enter the food chain. Organisms such as earthworms that are in constant physical contact with the soil may accumulate significantly higher body burdens of POPs than the surrounding soil. These concentrations may not be great enough to cause either acute or chronic toxicity in the earthworms. However, many species feed on earthworms and will concentrate the lipophilic compounds in the body fat. The result is an increase in POP body burdens as you rise up the trophic level of the food chain. This phenomenon has become known as biomagnification and may lead to symptoms of toxicity in organisms that have little contact with soil POPs.

### Availability to Plants

Generally, translocation of nonionic organic compounds from the roots to shoots is negligible due to the low water solubility, high n-octanol/water partition coefficients and strong associations of many compounds to soil particles/organic matter. Plant uptake of nonionic organic chemicals from soils is usually dominated by vegetative uptake of contaminated vapor from the surrounding air. Although high concentrations of soil-borne nonionic contaminants are rarely available for translocation by plant roots, heavily contaminated soils can influence the concentrations of organics in above-ground vegetation by the soil–air–plant route. In a review of crop plant uptake of organic compounds from soils it was concluded that the vast majority of organic compounds in soil are so strongly sorbed by the soil matrix that they exhibit low bioavailability to crop plants and so are accumulated at very low concentrations in the edible portion of food crops.

### Availability to Soil Microflora

The 'availability' of organic compounds in soils to microorganisms is critical if they are to be broken down and removed from the environment via either natural attenuation or bioremediation. Microbial biodegradation is a major process affecting the persistence of POPs in soils. The extent of POP bioavailability to bacteria in different soils may differ significantly, even under the same optimum growth conditions. Sorption has an important influence on the bioavailability of hydrophobic organic compounds in soils. A reduction in the degradation rates with increasing soil sorption capacity is apparent. Furthermore, a persistent residual fraction of POPs may remain unavailable to soil bacteria after remediation. This can be attributed to the slow migration of an increasingly large fraction of the POPs into the soil organic matter and mineral matrix with time.

As discussed earlier, POP chemical availability falls with increasing residence time in a soil. This reduced chemical availability is often mirrored by an analogous decline in bioavailability. This 'aging effect' may reduce the efficiency of bioremediation of a contaminated soil as a function of time. However, the reduced bioavailability may also reduce the potential toxicological threat of organic compounds in a soil. There is currently a great deal of research devoted to developing an accurate assessment of the bioavailability of organic compounds in soils. When developed, such an assessment of bioavailability will be able to identify both the suitability of a soil for bioremediation and the potential toxicological threat presented by that soil.

### Grazing Animals

Soil ingestion by grazing farm animals has been estimated to be in the range of 1–18% of the dry-matter intake, depending on forage conditions. Thus, ingested soil may represent a potentially major route of uptake for soil-borne organic contaminants by ruminants. It follows that consumption of meat and/or dairy products may be an important exposure pathway to humans. However, the total intake of organic compounds by grazing livestock is relatively low, even when a worst-case scenario is taken into consideration. Furthermore, due to the effect of slow diffusion, discussed earlier, the fraction of a compound 'available' to grazing livestock is likely to be less than that taken up on a weight compound/weight soil basis. It follows that any contemporary risk assessments may overestimate exposure of grazing livestock to soil-borne organic compounds.

### Availability to Earthworms

Earthworms are extremely numerous, especially in pasture land where their biomass can be up to $272\,\mathrm{g\,m^{-2}}$. Add to that the fact that an individual earthworm can consume up to 400 times its own weight in soil over a 12-month period, and their importance in the dynamics and persistence of organic compounds in soils becomes apparent. This high density and activity mean that up to 25% of the surface layers of soil is turned over annually. A relatively high body lipid content means that earthworms accumulate POPs. Earthworms may facilitate faster release of certain sorbed residues, since a large quantity of soil passes through the gut, and is consequently exposed to digestive enzymes that act on soil organic matter. Hence, the assumption that there is an analogous decline in chemical and biological

**Plate 43 Nematodes** Entire body view of a female *Pratylenchus agilis* (× 100 magnification), collected on the Konza Prairie (96 W 35′ 39 N 05′) beneath Scribner's panicum (*Panicum scribnerianum*) and bluegrass (*Poa pratensis*) near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin (2000).

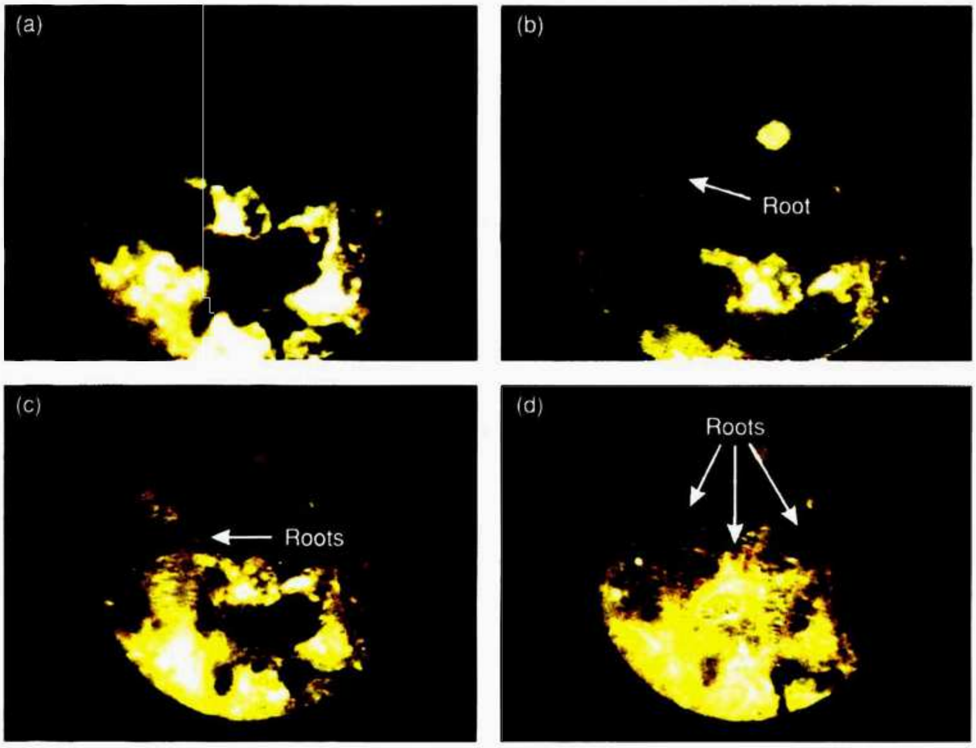


**Plate 44 Nitrogen in Soils/Plant Uptake** Root growth and proliferation in an N-rich organic material added as a spatially discrete patch and comprising lyophilized algal cell organic material patch (C:N ratio of 3.2:1). (a) One day after the patch was added, no roots in the patch zone were observed; (b) 21 days after the patch was added, a single root was seen growing into the N-rich zone. (c) The same image 25 days and (d) 48 days after the patch was added: increased root production and root-hair development in the N-rich zone was clearly observed. After 49 days, the plants were harvested and found to contain 54% of the N orginally added in the algal cell patch in their tissues. Adapted from *Trends in Plant Science*, vol. 5, Hodge A, Robinson D, and Fitter A. Are microorganisms more effective than plants at competing for nitrogen? pp. 304–308. © 2000, with kind permission from Elsevier Science.

**Plate 45 Organic Soils** Frost damage of crop on organic soil.

**Plate 46 Organic Soils** Organic soil used for cranberry production.

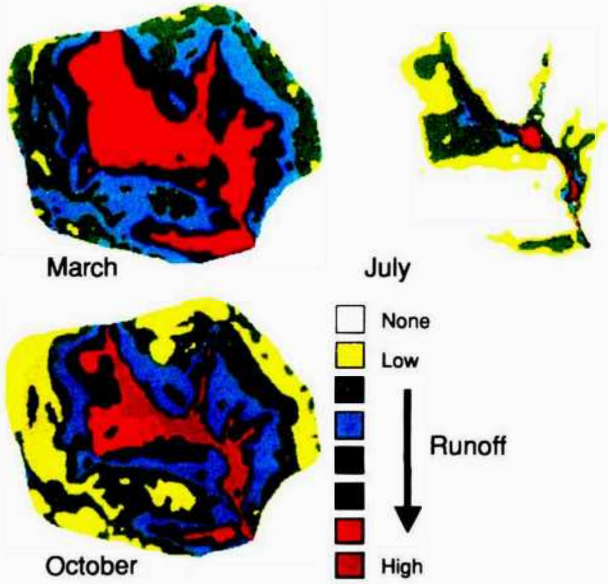


**Plate 47 Overland Flow** Seasonal changes of saturated areas (variable source areas) in a watershed in New York state. Color denotes areas more prone to saturation.

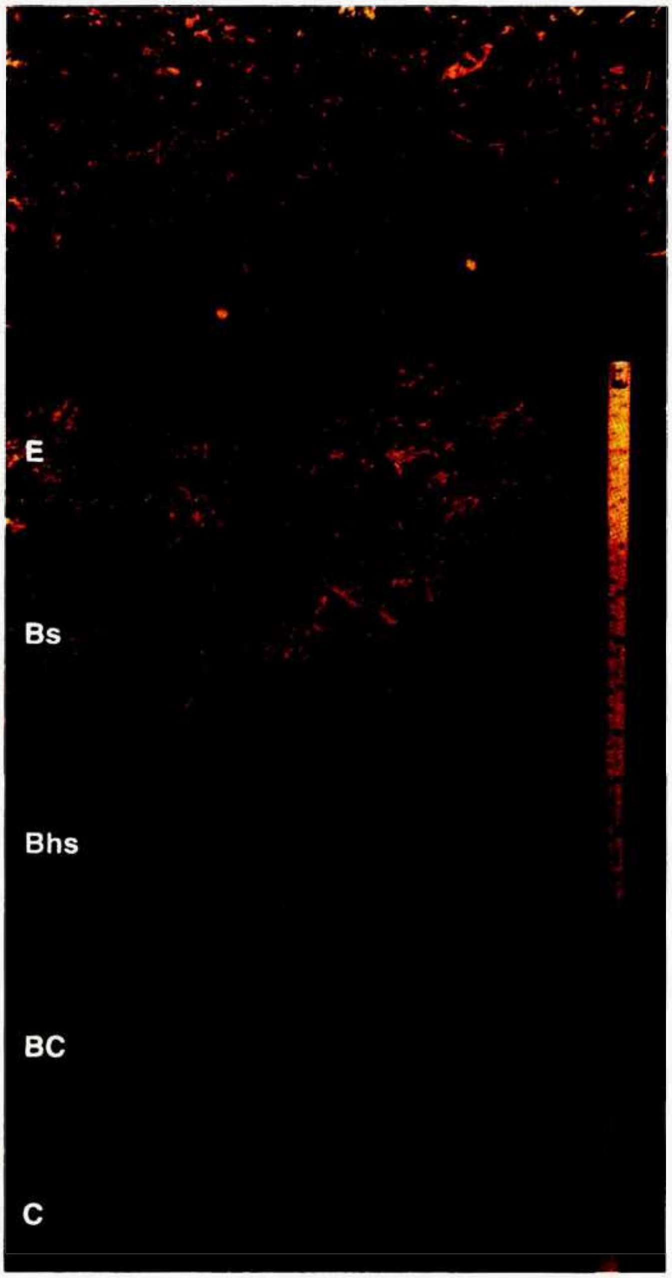


**Plate 48 Pedology/Dynamic** Boreal forest Spodosol showing an inverted sequence: a Bs over a Bhs; Walker Lake, Alaska. The yardstick is 50 cm long.

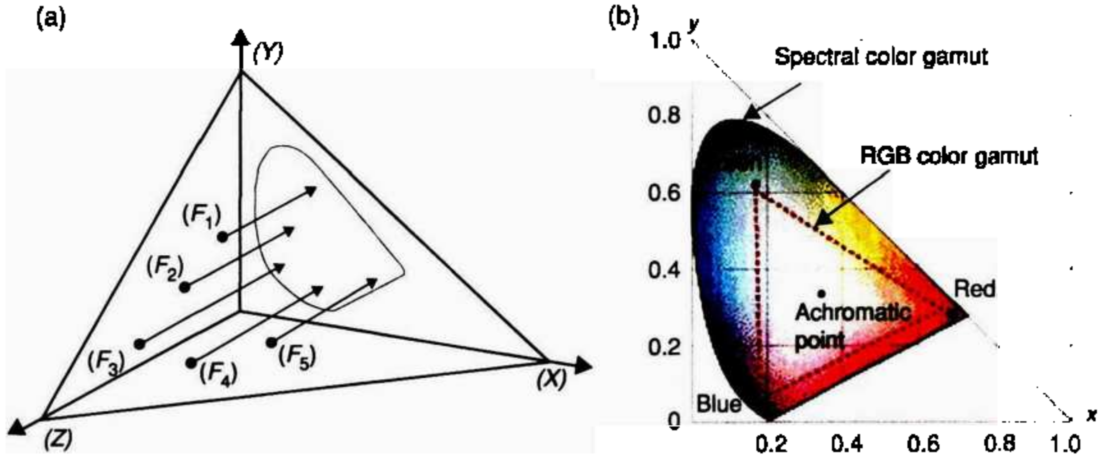


**Plate 49 Pedometrics** (a) Mapping of *XYZ* tristimulus values to (b) an *xy* chromaticity.

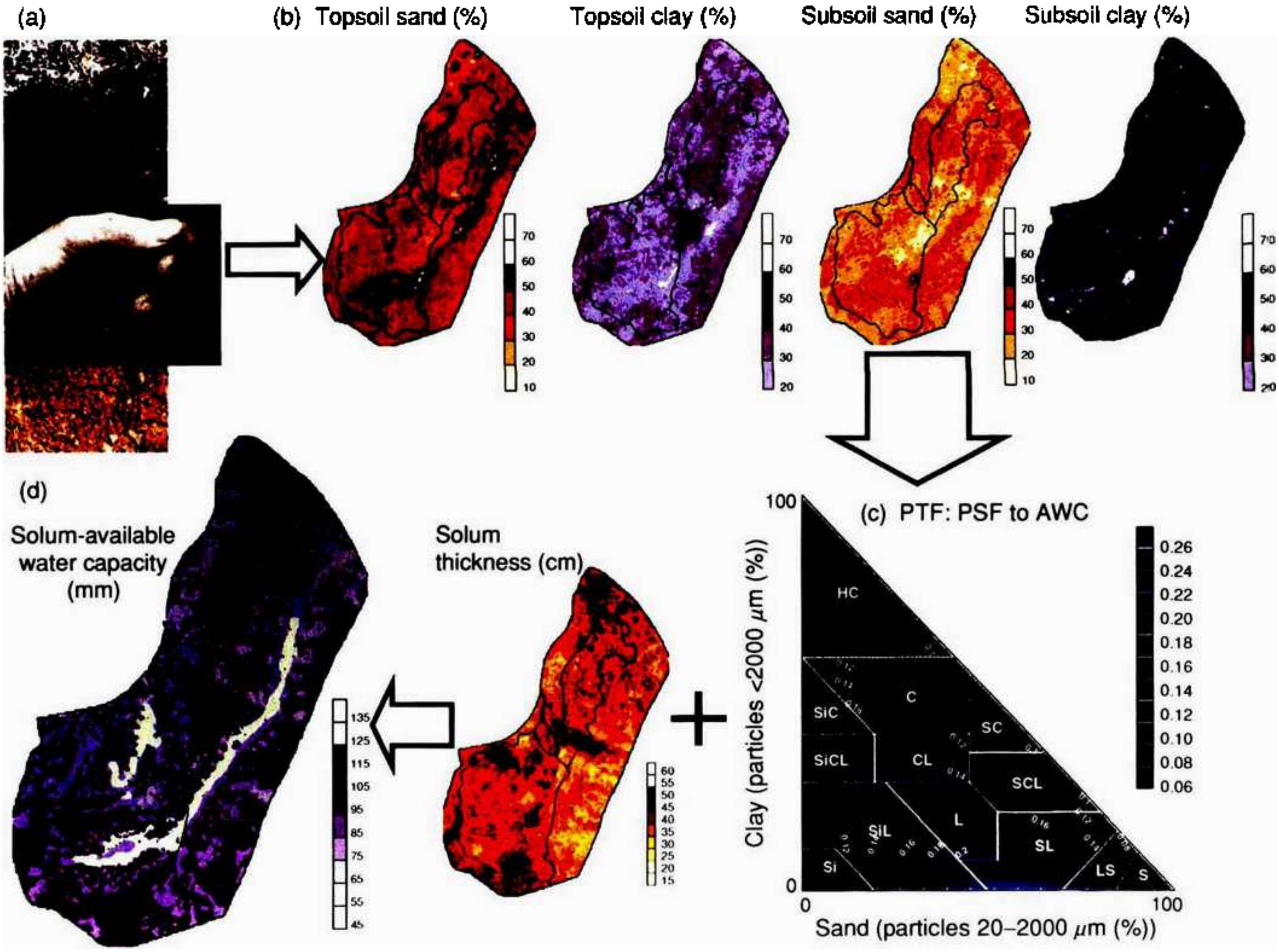


**Plate 50 Pedometrics** A simple example of pedotransfer functions (PTFs): (a) field texture was first determined by 'feel' method for topsoil and subsoil at each sample location; (b) particle-size fractions were then estimated from the textural classes and interpolated on to a fine grid; (c) horizon-available water capacity was in turn estimated through a PTF expressed as a function of particles sizes; (d) a combination of both horizons produces the solum-available water capacity. AWC, available water capacity.

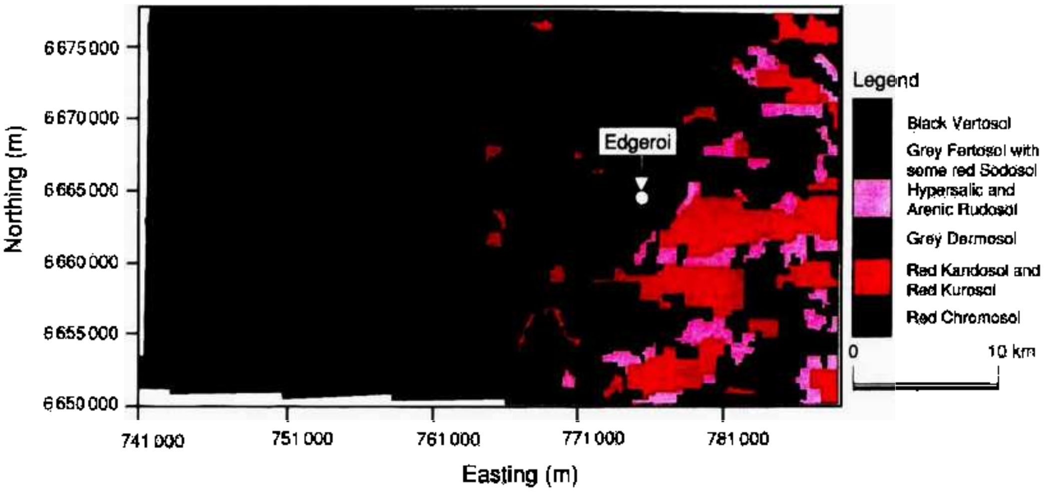


**Plate 51 Pedometrics** Predicted soil classes for the Edgeroi area in accordance with Australian Soil Classification. (Reproduced with permission from McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* 97: 293–327.)

**Plate 52 Polymers and Microorganisms** Glomalin (visualized as immunofluorescence using MAb32B11) on the surface of arbuscular mycorrhizal fungal hyphae on soil aggregates (1–2 mm diameter). Reproduced with permission from Wright SF (unpublished).

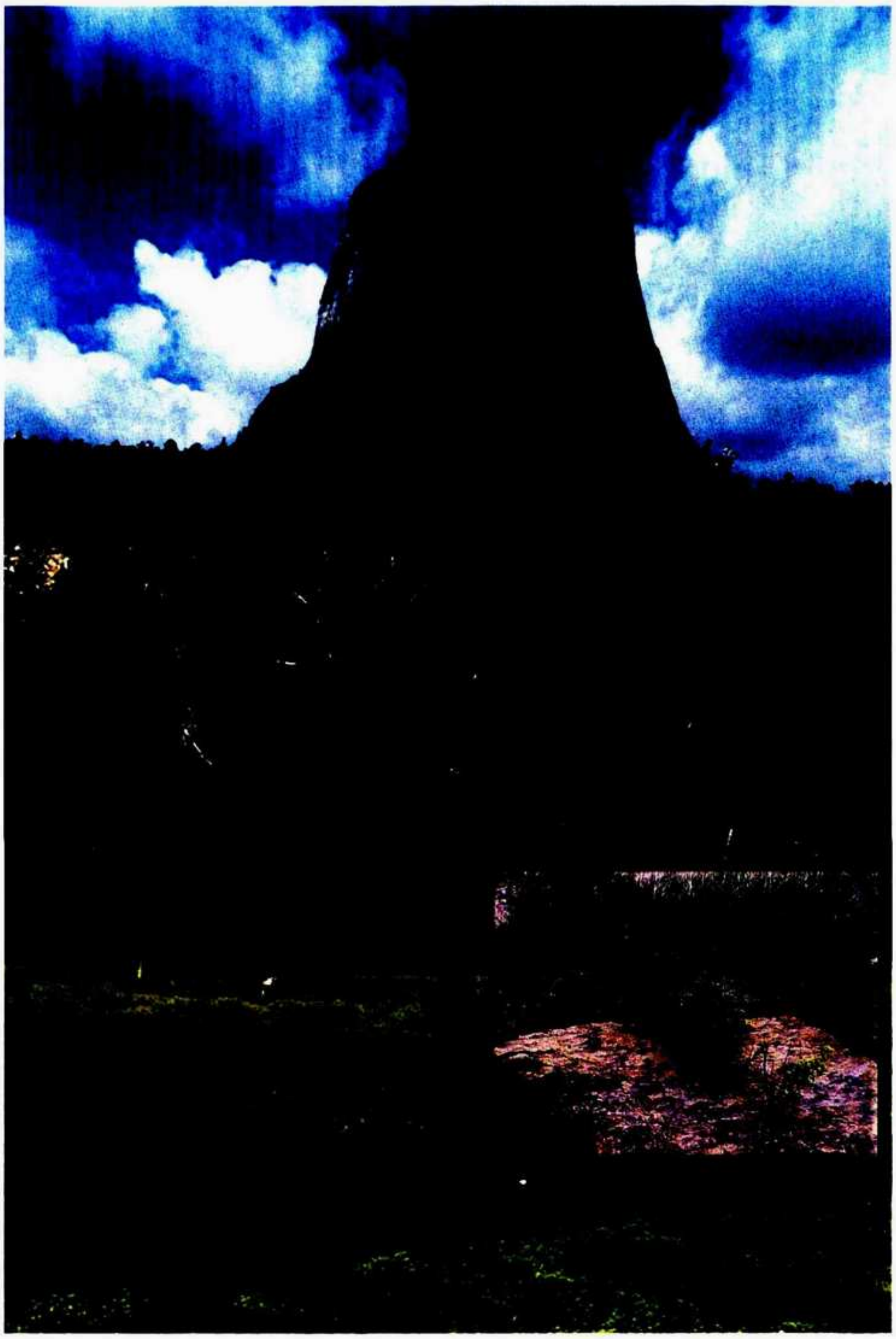


**Plate 53 Range Management** Leafy spurge invasion near Devils Tower, Wyoming, USA. The yellow-green weed dominates the riparian flood zone and can even establish in the harshest of areas (inset, leafy spurge growing out of a sandstone rock).

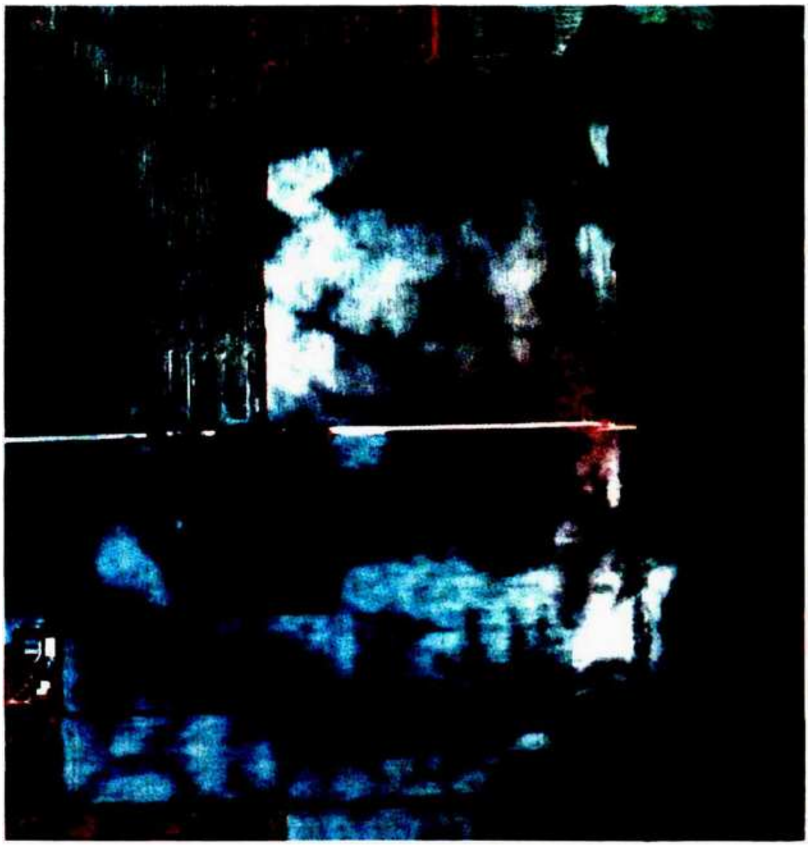


**Plate 55 Remote Sensing/Organic Matter** Aerial color infrared image showing soil patterns over an eastern Indiana farming area. The darker areas are higher in organic matter content and were previously wetland areas. The field in the upper left does not have the same soil patterns, since residues from the previous year cover it. Red shows healthy vegetation such as grasses and bushes; maroon, trees (the color resulting from the leaves, branch spacing, and shadows); white, roads and buildings; dark green to black, areas of high organic matter; light green, low surface soil organic matter content. Muted lined patterns masking the normal colors are corn residues from the previous crop. Courtesy of Emerge.

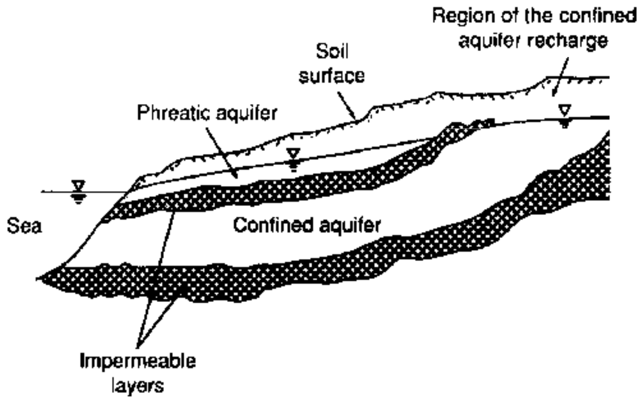


**Figure 1** Schematics of phreatic and confined aquifers.

water for humans. Figure 1 provides a schematic of the major two types of aquifer: (1) phreatic aquifers, and (2) confined aquifers. The bottom of the phreatic aquifer, which is also called the 'free surface aquifer,' comprises impermeable formations, whereas the top is unprotected against the penetration of pollutants through the vadose (unsaturated) zone. Such pollutants usually arrive at the phreatic aquifer with infiltrating runoff water (which provides the aquifer recharge) or water from other sources. A confined aquifer is bounded top and bottom by impermeable formations. However, in some areas, where the confined aquifer is subject to recharge, it is basically a phreatic aquifer.

There are several avenues for groundwater contamination. One common avenue of groundwater contamination is associated with direct contact between surface water and groundwater. An example of this type of groundwater contamination is represented by the role of the Arkansas River in the contamination of the aquifers in western Kansas. The Arkansas River is a draining river of surface water contaminated by pollutants of the big farms in Colorado. This type of water flows downstream through Kansas. In western Kansas, due to excessive pumping of groundwater, the Arkansas River supplies water into the aquifer, and thereby pollutes the groundwater with pollutants diverted from Colorado. In the Tel-Aviv area, Israel, the Ayalon River diverts pollutants from its upstream portion into the aquifer of Tel-Aviv, where the groundwater table is often significantly lower than the water level of the Ayalon River.

However, in most cases, groundwater is subject to contamination due to the arrival of seeping contaminants through the vadose zone of the soil that overlies the aquifer. Sometimes groundwater pollution is transferred by various types of mechanisms into the vadose zone, and generally groundwater contamination is intimately connected with the state of the vadose zone.

## Association of Pollutant Transport with the Water Cycle

Almost every textbook, guidebook, and manual of hydrology incorporates some descriptions and discussions which refer to the global cycle of water and emphasize its importance for properly maintaining the environment. The global water cycle demonstrates how quantities of water in the global scale are almost unchanged, while water is subject to transfer between different parts of the environment. The dielectric coefficient represents the ratio of the potential (and also the field strength) surrounding an electrostatic charge in a specific environment (medium) to the potential surrounding an identical charge in a vacuum. Due to its large dielectric coefficient, water has an extremely high capability to dissolve salts. Therefore, all chemical processes characterizing the different types of organisms take place in a water environment. By employing soaps and detergents, human society has increased the dissolving capability of water even to materials considered 'immiscible in water.' Society also utilizes water to transport solids as suspended and bed load; where suspended load is the flux of suspended solids and bed load is the flux of solids subject to movement (due to momentum transfer from the flowing water) at the bottom of the conduit. In contemporary urban society, water is mainly used as an agent to dissolve and transport solid material. Therefore, nowadays the involvement of the water cycle stages in pollutant transport in and between different parts of the environment is crucial for the preservation of human health and a balanced environment. It is extremely important to calculate budgets and mass flux balances of contaminants in different areas of the environment. In the bible, in the book of Ecclesiastes, the author expressed his astonishment that all rivers go to the sea, but the sea is never full. However, observations indicate that the salinity of the sea increases, and throughout the history of geology scientists have tried (unsuccessfully) to calculate the age of the earth according to the accumulation of salt in seas and oceans. It is not the appropriate place to evaluate why this calculation has not been successful, but simple mass flux balance shows that fresh water flowing into the sea incorporates small concentrations of various salts, and water leaving the sea in the vapor phase is almost free of salts. Therefore, the salinity of the sea must increase, as it does in lakes such as the Dead Sea (which is saturated with sodium chloride) that are not drained into other big water bodies.

Figure 2 provides a general schematic description showing how water and pollutants are transferred

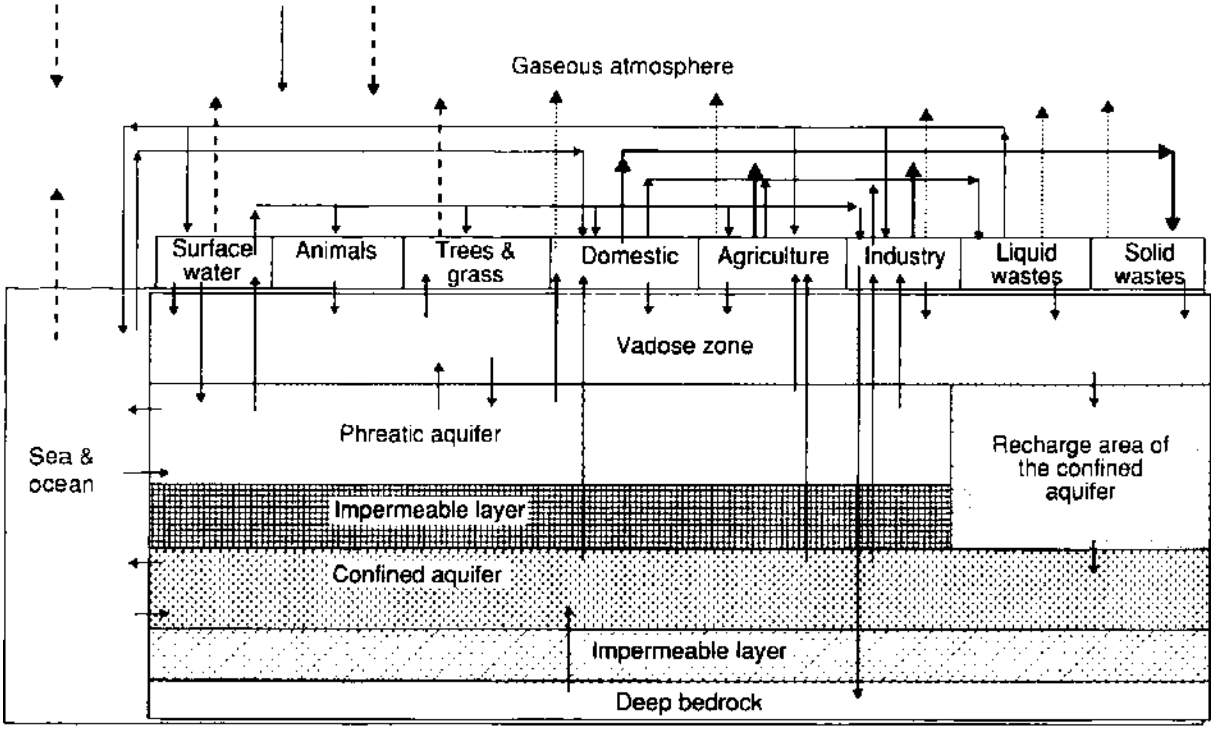


**Figure 2** The water cycle associated with pollutant transport.

between different parts of the environment and different water consumers. This figure shows several parts of the environment that can be considered as storage locations for water: (1) seas and oceans, (2) the gaseous atmosphere, and (3) the various geologic formations. **Figure 2** also shows some consumers of water. Note that it refers to surface water as one of the consumers, though some professionals may prefer to use another definition for surface water. There are about eight major water consumers (again the exact number of consumers is not as important as is their role in pollution production and transport). Thin solid arrows refer to the direction of transport of water with solutes and pollutants; thick solid arrows refer to solid wastes. Dotted arrows refer to transfer of air pollutants; and thick dashed arrows refer to transfer of pure water (vapors and/or precipitation). Therefore, according to **Figure 2**, pure water, as precipitation, is transferred from the atmosphere into the sea and the terrestrial part of the environment, but contaminated water is also transferred, e.g., acid rain.

Regarding consumers, domestic consumption also includes municipal uses, transportation, and all activities of an urban society. The major global water reservoir of seas and oceans provides a source of pure water in the atmosphere through evaporation. Through desalination, the atmosphere provides water for domestic use. On the other hand, salinity transport into groundwater might take place due to sea water intrusion into phreatic and confined aquifers, which is not a rare phenomenon, and may be enhanced in locations of excessive groundwater pumping. Phreatic and confined aquifers comprise the major terrestrial water storage of the environment. This storage is connected with the surface water and supplies water to society (domestic, agricultural, and industrial needs). All water consumers discharge water with higher solute concentrations than they originally obtained. Some consumers produce wastewater that should be transferred into liquid-waste collection sites, treated, and then delivered to seas, surface water storage, or agricultural and industrial uses. Some water used by consumers may directly enter the vadose zone, e.g., agricultural irrigation waters. This water may be subject to evapotranspiration (and thereby solutes of this water are accumulated in vegetation and the vadose zone) and/or infiltration through the vadose zone into the phreatic aquifer. (*See* **Evapotranspiration.**) Such infiltration is associated with the accumulation of some additional quantities of solutes. The phreatic aquifer may supply water with contaminants into the vadose zone by capillary movement, induced by water consumption of trees. Volatile contaminants dissolved in groundwater may be transferred as vapors into the vadose zone gaseous phase. The transfer of pollutant between the phreatic aquifer and the vadose zone is greatly enhanced by temporal, seasonal, and annual fluctuations of the groundwater table elevation.

Various industries apply techniques of deep well injection to get rid of their wastes. As an example, pumping of crude oil is associated with pumping of brines. The latter, in many cases is directly injected into the deep bedrock. In parts of south Florida, practices of deep well injection were used for disposal of industrial wastes.

With regard to the two types of aquifers, the confined aquifer is subject to recharge in restricted areas and discharges water into the sea. Excessive groundwater pumping from the confined aquifer may lead to saltwater intrusion from the sea, as well as brine intrusion from the deep bedrock. The phreatic aquifer is intimately connected with the surface water and through the vadose zone it obtains infiltrating runoff water with solutes accumulated during the infiltration process. The vadose zone is the 'membrane' that separates the phreatic aquifer from the water consumers. The vadose zone changes the solute content (including a variety of solutes, pollutants, suspended and colloidal materials, microbes, parasites, etc.) of the infiltrating water owing to its different soil components, clay minerals, and residual water saturation. During the runoff water infiltration, solutes adsorbed to clay minerals may be released and transferred into the infiltrating water. On the other hand, suspended materials transported with the surface runoff usually cannot penetrate the vadose zone, and they do not arrive at the aquifer. In some cases, the infiltrating water quality may be improved by the adsorption of solutes by clay minerals. However, some of the positive phenomena of water-quality improvement are time-dependent, and therefore they should be carefully evaluated from a quantitative viewpoint.

It is appropriate to assume that every solute whose concentration in water is above its threshold concentration should be considered as a pollutant whose presence is involved in some risks to human health and the environment.

Any part of the environment involved in the water cycle and associated pollutant transport should be preserved in an appropriate risk-free mass budget and flux balance. Figure 2 indicates that special attention should be given to mass balances in the vadose zone and the phreatic aquifer due to the intimate relationships between these two types of strata.

## Simplified Mass Budget and Flux Balances in Phreatic Aquifers

This section applies simplified mass budget and flux balance calculations to a variety of environmental development stages typical of different global locations. **Figure 3** refers to four different scenarios or stages relevant to groundwater contamination within phreatic aquifers. The first stage, which is not emphasized by **Figure 3**, considers the natural environment

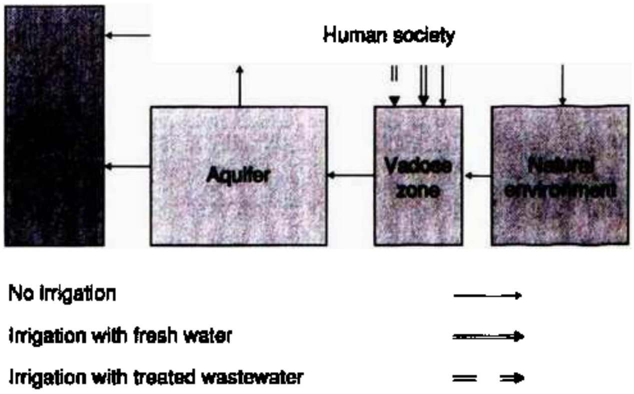


**Figure 3** Different stages (scenarios) of pollutant flux development.

with very few effects from society. The second stage refers to involvement of human society with no (or very limited) irrigation practices. The third stage is associated with freshwater irrigation; and the fourth stage is associated with all types of irrigation by fresh water and treated wastewater. The calculations refer to a single contaminant, which is considered as the limiting pollutant of the environment; but similar calculations can be done with regard to any relevant pollutant.

**Figure 4** describes the involvement of water and pollutant fluxes with the simplified water cycle. The following types of water fluxes and their relevance to pollutant transport are described by **Figure 4**:

1. $Q_{nr}$ is the natural recharge of the phreatic aquifer, namely the net recharge (precipitation minus evapotranspiration). Even in the natural environment, the recharging flow rate accumulates some contaminants, $\Delta C_0$ aboveground and $\Delta C_1$ during its infiltration through the vadose zone;
2. $Q_{so}$ is the groundwater flow rate that is emitted into seas and oceans;
3. $Q_{hs}$ is the flow rate consumed by society. Society always supplies some contaminants into the natural recharge and increases its pollutant concentration by $\Delta C_1$. Human society also increases the pollutant concentration by the amount $\Delta C_4$ before delivering the water into seas and oceans;
4. $Q_{ir}$ represents the irrigation flow rate. It is assumed that people apply groundwater for irrigation purposes. The irrigation flow rate comprises a certain portion of $Q_{hs}$. However, irrigation is associated with evaporation flow rate, $Q_{ev1}$. Evaporation decreases the flux of irrigation water that infiltrates into the aquifer. However, it is assumed that all pollutants present with $Q_{ir}$ arrive at the phreatic aquifer. Furthermore, the reduced infiltrating discharge $(Q_{ir} - Q_{ev1})$ accumulates additional pollutant concentration $\Delta C_1$ during its infiltration through the vadose zone;

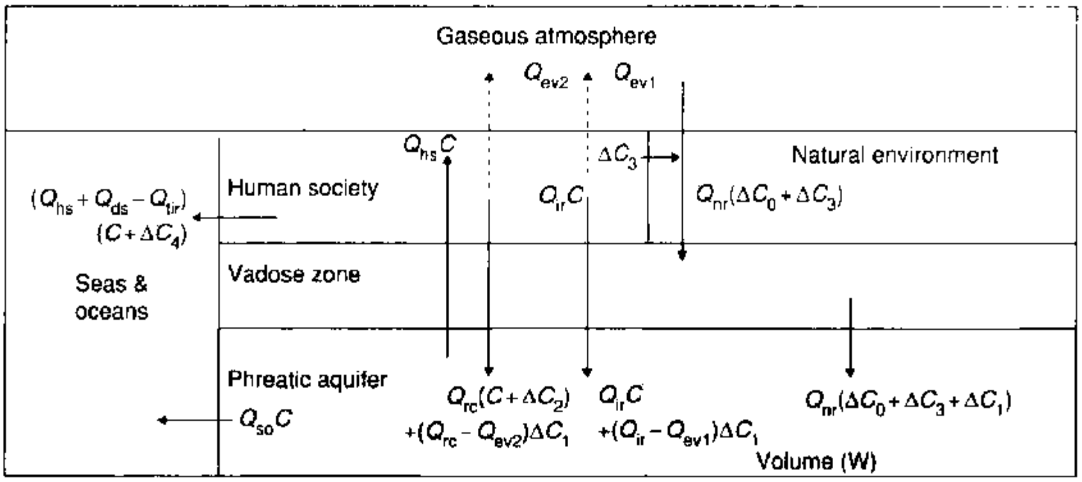


**Figure 4** Considered fluxes of water and pollutants in various scenarios.

5. $Q_{rc}$ is the flow rate of recycled wastewater. Wastewater-recycling decreases the discharge from populations that is delivered into seas and oceans. It is assumed that all recycled wastewater is used for irrigation. Probably the pollutant concentration within this type of water is greater by $\Delta C_2$ than pollutant concentration in groundwater. However, again this type of irrigation is associated with evaporation flow rate, $Q_{ev2}$. Again considerations should be made with regard to the additional pollutant mass flux originating from experiencing wastewater-recycling.

We assume that measures are taken to keep the aquifer total volume *W* unchanged, and we also assume that initially the pollutant concentration is $C_0$. Then all pollutant mass budget and flux balance calculations for the different stages, described by **Figures 3** and **4**, indicate that the pollutant concentration increases exponentially as:

$$C = C_\infty - \Delta C\left[1 - \exp\left(-\frac{Q_{nr} - Q_{ev}}{W}t\right)\right] \qquad [1]$$

where

$$\Delta C = C_\infty - C_0$$

The values of $C_\infty$ and $\Delta C$ are given in **Table 1**, $Q_{ev}$ is the total rate of evaporation. Therefore the time required to obtain steady state conditions, namely when $C$ is smaller than $C_\infty$ by one percent of $\Delta C$ is:

$$(t)_{C=C_\infty-0.01\Delta C} = \frac{4.61(Q_{nr} - Q_{ev})}{W} \qquad [2]$$

**Table 1** summarizes the results of water flux balance calculations and pollutant mass budget and flux balance calculations. In **Table 1**, expressions of the two bottom cells of the fourth column have been obtained by assuming that evapotranspiration associated with irrigation is significantly smaller than the natural recharge of the aquifer. Otherwise, these expressions should be replaced by simple deduction of relevant expressions of the third column. However, it should be noted that only the last term of the bottom cell expression in the fourth column represents the real contribution of wastewater recycling to pollutant concentration in the groundwater, provided that irrigation with treated wastewater replaces irrigation with water pumped from the aquifer. $Q_{tir}$ and $Q_{ev}$ in **Table 1** are values of total irrigation $(Q_{ir} + Q_{rc})$ and total evaporation $(Q_{ev1} + Q_{ev2})$ respectively.

**Table 1** can be useful for initial evaluation of long-term implementation of water resources management practices. This table indicates that irrigation, even with water of drinking quality, increases the pollutant concentration in the aquifer. On the other hand, though the marginal increase in aquifer pollution by wastewater-recycling is greater than irrigation with drinking water, this effect is mitigated if wastewater-recycling is implemented as a substitute for irrigation with drinking water.

Study of **Figure 4** and Table 1 may lead to some useful ideas about the preservation of groundwater quality in phreatic aquifers. First of all, such an objective should be achieved by adequate considerations based on the control of all pollutant inputs, which should be minimized, and pollutant outputs, which should be maximized. **Figure 4** and the results of calculations in **Table 1** indicate that irrigation of all types poses a major threat to groundwater quality, as it circulates pollutants between the groundwater and the human population. Of course, advanced treatment of wastewater to drinking-water quality is appropriate, as it reduces the pollutant input, but such an approach does not necessarily provide sufficient guarantee against the accumulation

**Table 1** Water and pollutant mass budget and flux balance results for typical cases of aquifer exploitation scenarios

| *Stage (scenario)* | *Water balance* | *Pollutant concentration at $t\rightarrow\infty$* | *Increase of $C_\infty$ (value of $\Delta C$)* |
|---|---|---|---|
| Natural environment | $Q_{so} - Q_{nr} = 0$ | $C_\infty = \Delta C_0 + \Delta C_1$ | 0 |
| No irrigation | $Q_{so} + Q_{hs} - Q_{nr} = 0$ | $C_\infty = \Delta C_0 + \Delta C_3 + \Delta C_1$ | $\Delta C_3$ |
| Freshwater irrigation | $Q_{so} + Q_{hs} + Q_{ev1}$<br>$- Q_{nr} - Q_{ir} = 0$ | $C_\infty = \dfrac{Q_{nr}(\Delta C_0 + \Delta C_3 + \Delta C_1) + (Q_{ir} - Q_{ev1})\Delta C_1}{Q_{nr} - Q_{ev1}}$ | $\dfrac{Q_{ev1}}{Q_{nr}}(\Delta C_0 + \Delta C_3 + \Delta C_1)$<br>$+ \dfrac{Q_{ir} - Q_{ev1}}{Q_{nr} - Q_{ev1}}\Delta C_1$ |
| All types of irrigation | $Q_{so} + Q_{hs} + Q_{ev1} + Q_{ev2}$<br>$- Q_{nr} - Q_{ir} - Q_{rc} = 0$ | $C_\infty = \dfrac{Q_{nr}(\Delta C_0 + \Delta C_3 + \Delta C_1) + (Q_{tir} - Q_{ev})\Delta C_1 + \Delta C_1 + Q_{rc}\Delta C_2}{Q_{nr} - Q_{ev}}$ | $\dfrac{Q_{ev2}}{Q_{nr}}(\Delta C_0 + \Delta C_3 + \Delta C_1)$<br>$+ \dfrac{Q_{ev2}}{Q_{nr}}\left(\dfrac{Q_{ir} - Q_{ev1}}{Q_{nr}}\right)\Delta C_1$<br>$+ \dfrac{Q_{rc} - Q_{ev2}}{Q_{nr} - Q_{ev}}\Delta C_1 + \dfrac{Q_{rc}\Delta C_2}{Q_{nr} - Q_{ev}}$ |

of high pollutant concentrations in the phreatic aquifers.

As stated, Figure 4 refers to a single pollutant. In an aquifer system subject to a variety of pollutants, whose presence introduces risks to human health and the environment, similar mass budget and flux balance calculations should be made with regard to each one of these pollutants for the evaluation of long-term processes involved with groundwater quality.

## Penetration of Solutes into the Phreatic Aquifer

**Figure 5** provides a classic description of the distribution of water saturation in the phreatic aquifer and the vadoze zone. This figure does not include the water saturation distribution close to the ground surface, which is subject to significant changes due to precipitation, irrigation, dry seasons, etc.

The region of major interest with regard to solute penetration into the phreatic aquifer is the capillary fringe, which incorporates water subject to pressure below the atmospheric pressure. Due to the capillary forces, water in the capillary fringe is almost stagnant, and does not take part in groundwater horizontal flow. The groundwater table is defined as the level of atmospheric pressure. Therefore, as shown in **Figure 5**, the water level in a well represents the groundwater table. The contaminant usually migrates as a solute with water that infiltrates, as a wetting front through the vadoze zone. During this process, and mainly in the capillary fringe, where water residual saturation is comparatively high, the infiltrating water is subject to mixing with the residual water saturation and redistribution of contaminant concentrations takes place. This mixing is associated with local reduction of the capillary effect and migration of water with pollutants into the aquifer. However, all these processes may significantly increase the pollutant concentration in the capillary fringe. Here, the effect of groundwater table fluctuation may play an important role in preserving the contamination of the capillary fringe at levels similar to those of the aquifer. However, if such fluctuations are minor, then the capillary fringe may be subject to significantly higher pollutant concentrations than those typical of the aquifer.

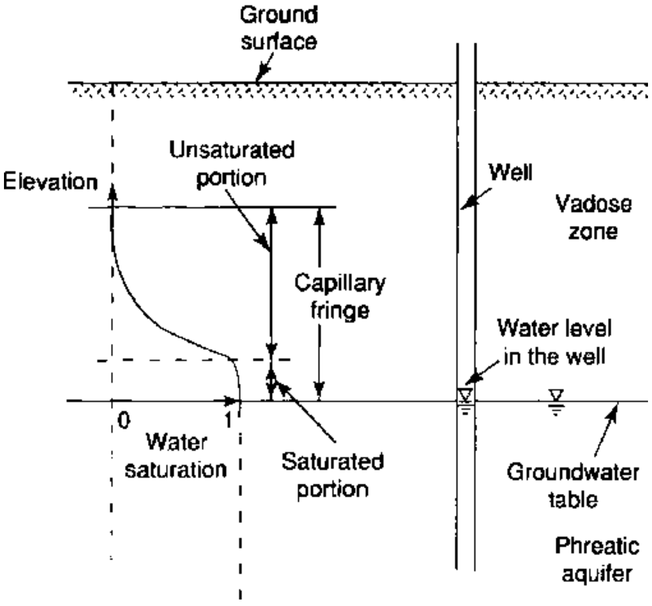


**Figure 5** Water saturation distribution in the phreatic aquifer and the vadose zone.

## Pollutant Transport in the Aquifer

Pollutants that penetrate through the capillary fringe, due to a wetting-front movement or due to groundwater table fluctuations, are subject to transport in the aquifer. The evaluation of pollutant transport in groundwater incorporates at least two basic differential equations, the equation of groundwater flow, namely Darcy law, and the diffusion–advection equation:

$$\mathbf{q} = \mathbf{K} \cdot \mathbf{J} \qquad [3]$$

$$\frac{\partial C}{\partial t} + \mathbf{V} \cdot \nabla C = \nabla \cdot (\mathbf{D}_h \cdot \nabla C) - R \qquad [4]$$

where **q** is the specific discharge vector (Darcy velocity), **K** is the hydraulic conductivity tensor (which depends on the permeability and kinematic viscosity of the fluid, namely water), **J** is the gradient of the piezometric head, $C$ is the pollutant concentration, **V** is the interstitial velocity (the specific discharge divided by the effective porosity), $\mathbf{D}_h$ is the hydrodynamic dispersion tensor (it incorporates effects of diffusion and tortuosity of streamlines), which depends on **V** and properties of the permeable domain, and $R$ represents possible chemical reactions (such as adsorption and decay).

In cases of isotropic porous media and fluid of constant density (no density stratification is significant), the hydraulic conductivity is a scalar quantity. Then eqn [3] represents the specific discharge as a vector originating from the gradient of the piezometric head, which is a scalar quantity. Therefore, all approaches and techniques used for the solution of potential flow problems are also applicable to the description of flow phenomena taking place in the aquifer. However, usually in groundwater flow, the streamlines are almost horizontal, and in such cases the Dupuit approximation can be applied.

**Table 1** Water and pollutant mass budget and flux balance results for typical cases of aquifer exploitation scenarios

| *Stage (scenario)* | *Water balance* | *Pollutant concentration at $t \to \infty$* | *Increase of $C_\infty$ (value of $\Delta C$)* |
|---|---|---|---|
| Natural environment | $Q_{so} - Q_{nr} = 0$ | $C_\infty = \Delta C_0 + \Delta C_1$ | 0 |
| No irrigation | $Q_{so} + Q_{hs} - Q_{nr} = 0$ | $C_\infty = \Delta C_0 + \Delta C_3 + \Delta C_1$ | $\Delta C_3$ |
| Freshwater irrigation | $Q_{so} + Q_{hs} + Q_{ev1} - Q_{nr} - Q_{ir} = 0$ | $C_\infty = \frac{Q_{nr}(\Delta C_0 + \Delta C_3 + \Delta C_1) + (Q_{ir} - Q_{ev1})\Delta C_1}{Q_{nr} - Q_{ev1}}$ | $\frac{Q_{ev1}}{Q_{nr}}(\Delta C_0 + \Delta C_3 + \Delta C_1) + \frac{Q_{ir} - Q_{ev1}}{Q_{nr} - Q_{ev1}}\Delta C_1$ |
| All types of irrigation | $Q_{so} + Q_{hs} + Q_{ev1} + Q_{ev2} - Q_{nr} - Q_{ir} - Q_{rc} = 0$ | $C_\infty = \frac{Q_{nr}(\Delta C_0 + \Delta C_3 + \Delta C_1) + (Q_{tir} - Q_{ev})\Delta C_1 + \Delta C_1 + Q_{rc}\Delta C_2}{Q_{nr} - Q_{ev}}$ | $\frac{Q_{ev2}}{Q_{nr}}(\Delta C_0 + \Delta C_3 + \Delta C_1) + \frac{Q_{ev2}}{Q_{nr}}\left(\frac{Q_{ir} - Q_{ev1}}{Q_{nr}}\right)\Delta C_1 + \frac{Q_{rc} - Q_{ev2}}{Q_{nr} - Q_{ev}}\Delta C_1 + \frac{Q_{rc}\Delta C_2}{Q_{nr} - Q_{ev}}$ |

According to the Dupuit approximation, lines of constant piezometric head (equipotential lines) are vertical. If pollutants have minor effect on the fluid density, the solution of eqns [3] and [4] can be done in a series. This means that after solving the flow with eqn [3] and obtaining a complete description of the streamlines and velocity distribution in the domain, these results are used as an input for the solution of eqn [4]. If the pollutant concentration is comparatively high, then density stratification might be significant, and therefore the velocity and pollutant distributions in the domain should be obtained simultaneously by the simultaneous solution of eqns [3] and [4]. Such cases are typical of saltwater intrusion into aquifers. The solution of eqn [3] is often obtained by analytical means, provided that density stratifications are minor. On the other hand, in most practical cases the solution of eqn [4], under all types of fluid-density distribution in the domain, usually requires the employment of numerical approaches. Only a limited number of analytical solutions to this equation have been developed. However, various methods can be used to obtain approximate first-order descriptions of pollutant transport in the aquifer. **Figure 6** provides a schematic description of the buildup of the contaminant plume at the pollutant penetration site. In some cases, the buildup of this plume can be evaluated by using the approximate approach of boundary layer methods.

It should be noted eqns [3] and [4] refer to an unconsolidated permeable material and a single type of pollutant. If several pollutants are involved, then each pollutant has to be referred to by a particular use of its contaminant transport equation, subject to its particular set of initial and boundary conditions. In some cases also considerations of possible chemical reactions that might occur between the different pollutants have to be taken into account.

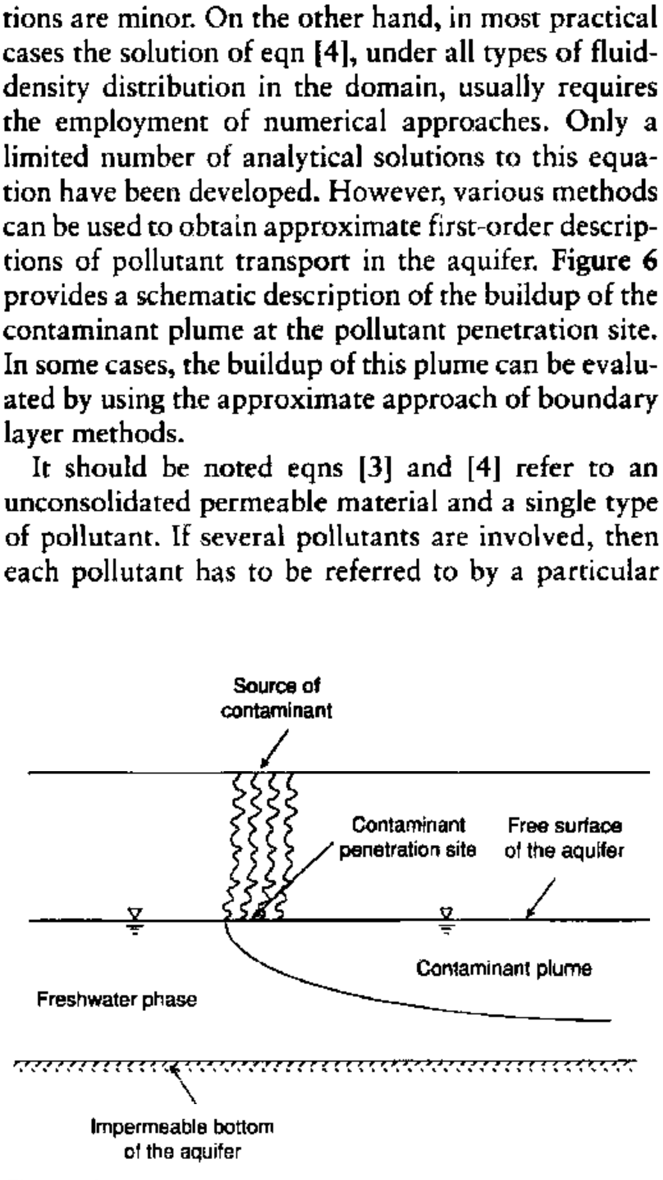


**Figure 6** Buildup of a contaminant plume at a contaminant penetration site. Reproduced from Rubin H and Buddemeier RW (2002) Groundwater contamination downstream of a contaminant penetration site. 1. Extension-expansion of the contaminant plume. *Journal of Environmental Science and Health* A37(10): 1781–1812.

Two major issues that are intimately connected should be noted in connection with the appropriate evaluation of pollutant transport in aquifers: (1) upscaling, namely the relevance of phenomena observed on a small scale, such as in a laboratory, to those predicted at the large field scale; and (2) the effect of the aquifer formation heterogeneity on pollutant transport. Physical quantities characterizing groundwater flow through the aquifer are very little affected by the formation heterogeneity, and flow calculations can be done by using the continuum approach, which leads to eqn [3]. With regard to pollutant transport in groundwater, issues of upscaling and aquifer heterogeneity may impose very serious problems on the simple use of the continuum approach, which leads to eqn [4]. Therefore, a variety of approaches, such as double porosity approximation, and statistical presentation of the medium characteristics and transport parameters are very often used by scientists and engineers.

## Contamination of Groundwater by NAPLS

Since the mid-1980s, much attention has been directed toward the evaluation of groundwater contamination by nonaqueous-phase liquids (NAPLs). This high level of attention has originated from the realization of how common cases of groundwater contamination by NAPLs are, and how many risks to human health and the environment are associated with such a contamination. The term 'NAPL' has substituted the previous term 'immiscible fluids.' This change indicates that, although the miscibility of these fluids in water is very low (in the order of magnitude of parts per million, ppm), such concentrations in water avoid its safe use by human populations, and they also introduce high risks to the environment. 'NAPL' usually represents an organic liquid-phase material. Such a liquid is very often comprised of a mixture of many organic compounds used for a specific task, e.g., fossil fuels that are employed in driving internal combustion engines, or creosote that is used for treating wood in the wood industry. Various types of chlorinated hydrocarbons are NAPLs used for medical purposes, e.g., chloroform, or for degreasing purposes by the steel industry, e.g., trichloroethylene.

Generally, two types of NAPLs are identified: (1) light NAPLs (LNAPLs), whose density is lower than that of water; and (2) dense NAPLs (DNAPLs),

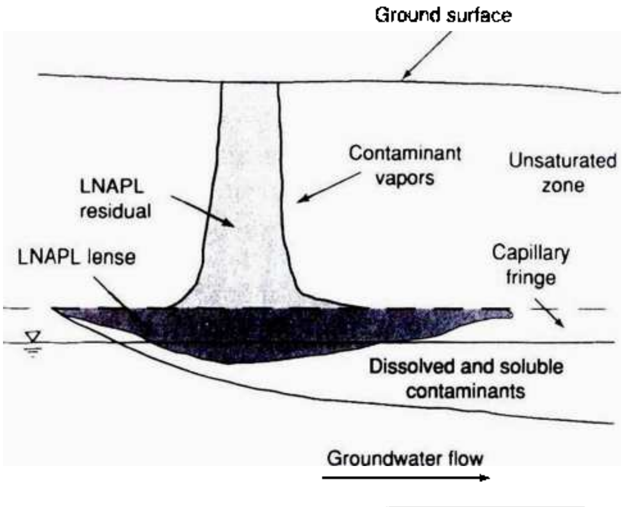


**Figure 7** Contamination of the vadose zone and the phreatic aquifer by light nonaqueous-phase liquids (LNAPL). Reproduced with permission from Rubin H, Narkis N, Carberry JB (1998) Overview of NAPL contamination and reclamation. In: Rubin H, Narkis N, Carberry JB (eds) *Soil and Aquifer Pollution – Non-Aqueous Phase Liquids – Contamination and Reclamation*, pp. 3–17. Springer-Verlag: Berlin, Heidelberg, New York. © Springer-Verlag.

whose density is higher than that of water. This classification is extremely important for the evaluation of NAPL penetration into the aquifer, and its migration with the flowing groundwater. **Figures 7** and **8** provide schematic descriptions referring to LNAPL and DNAPL contamination, respectively.

After the release of the NAPL (LNAPL or DNAPL) at the ground surface, it migrates downward within the vadose zone with limited lateral expansion. During the downward migration, the NAPL plume is subject to partial entrapment and the release of some vapors of volatile compounds into the gaseous phase of the vadose zone. Differences between LNAPL and DNAPL contamination take place after the arrival of the contaminant plume at the capillary fringe and the groundwater table.

As shown in **Figure 7**, at the groundwater table, the LNAPL plume spreads horizontally and creates a lens that floats on top of the groundwater table. If the LNAPL plume originates from the release of a limited and small quantity of NAPL, then possibly the entire mass of NAPL is completely entrapped in the vadose zone without its arrival at the groundwater table. If the released quantity is limited but incorporates a significant volume, then the scenario of migrating LNAPL ends with its complete entrapment within the vadose zone and the floating LNAPL lens on top of the groundwater table.

As shown in **Figure 8**, after its arrival at the groundwater table, the DNAPL plume penetrates into the aquifer and keeps its downward migration. This process is associated with DNAPL entrapment within the phreatic aquifer. If quantities of released DNAPL are large, then the scenario of DNAPL migration ends with complete entrapment of the released DNAPL in the vadose zone, in the phreatic aquifer, and even possible accumulation of the DNAPL in a pool, at the bottom of the aquifer (or on top of impermeable soil lenses that might be located within the phreatic aquifer).

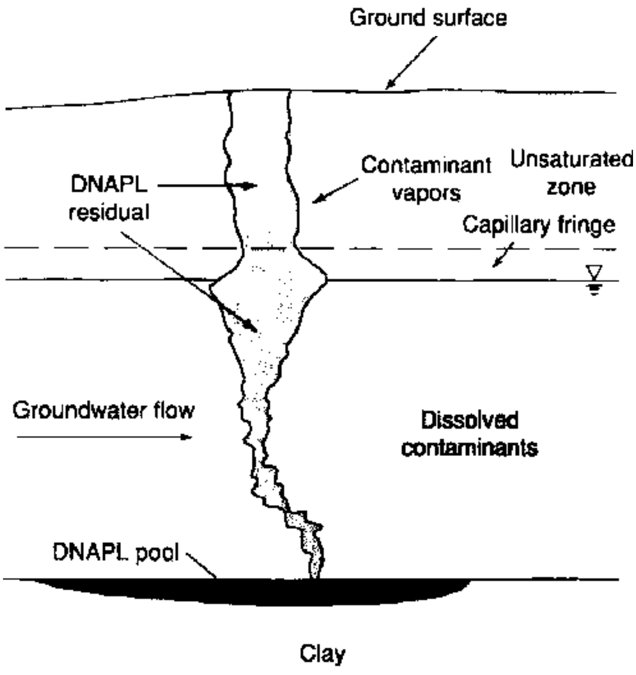


**Figure 8** Contamination of the vadose zone and the phreatic aquifer by dense nonaqueous-phase liquids (DNAPL). Reproduced with permission from Rubin H, Narkis N, Carberry JB (1998) Overview of NAPL contamination and reclamation. In: Rubin H, Narkis N, Carberry JB (eds) *Soil and Aquifer Pollution – Non-Aqueous Phase Liquids – Contamination and Reclamation*, pp. 3–17. Springer-Verlag: Berlin, Heidelberg, New York. © Springer-Verlag.

Scenarios of the NAPL migration in the vadose zone and the phreatic aquifer can be evaluated by using numerical models originating from reference to multiphase flow in porous media. Such models have been used for many years by the oil industry, and have been adapted for the quantification of the environmental issues represented by **Figures 7** and **8**. In the vadose zone, calculations refer to the solid phase and three fluid phases (gaseous, water, and NAPL). In the aquifer, reference is made to the solid phase and two fluid phases (water and NAPL). The mathematical model of multiphase flow in porous media is generally formulated as:

$$\begin{aligned}\varepsilon_\alpha \rho^\alpha \frac{\partial \omega_i^\alpha}{\partial t} + \rho^\alpha \mathbf{q}^\alpha \nabla \omega_i^\alpha - \nabla \cdot (\varepsilon_\alpha \rho^\alpha \mathbf{D}_{h_i}^\alpha \cdot \nabla \omega_i^\alpha) \\ = \sum_{\beta \neq \alpha} E_i^{\alpha\beta} - \omega_i^\alpha E^\alpha \end{aligned} \qquad [5]$$

which is subject to the two following constraints:

$$\sum_i \omega_i^\alpha = 1; \quad \sum_\alpha \epsilon_\alpha = 1 \qquad [6]$$

where $\epsilon_\alpha$ is the volume fraction of the $\alpha$ phase, $\omega_i^\alpha$ is the mass fraction of component $i$ in the $\alpha$ phase, $\mathbf{D}_h^\alpha$ is the hydrodynamic dispersion tensor of the $\alpha$ phase, $\mathbf{q}^\alpha$ is the specific discharge of the $\alpha$ phase, $\rho^\alpha$ is the density of the $\alpha$ phase, and $E_i^{\alpha\beta}$ represents exchange of mass of component $i$ between the $\alpha$ and $\beta$ phases.

Equation [5] refers to each component $i$ (of the total number of relevant compounds), that may comprise portions of the $\alpha$ and $\beta$ phases (solid skeleton, water, gaseous phase, and NAPL). Besides eqns [5] and [6], some constitutive relationships should be applied to calculate the propagation of each fluid phase and specific species in the domain.

However, as implied by preceding paragraphs, scenarios of multiphase flow involved in contamination of the vadose zone and the phreatic aquifer usually take place for a limited time, until all quantities of the NAPL are entrapped within the porous medium; then the only mobile fluid phase present in the aquifer is water. Under such conditions the effect of the immobile, entrapped NAPL on flow parameters of groundwater are represented by the decrease in the water saturation and the reduced permeability of the formation. These effects lead to the following modified Darcy law:

$$\mathbf{q} = \nabla \cdot \left[ g \left( \mathbf{k} \frac{k_{rw}}{\nu} \right) \cdot \mathbf{J} \right] \qquad [7]$$

where $g$ is the gravitational acceleration, $\mathbf{k}$ is the permeability tensor, $k_{rw}$ is the relative water permeability (which depends on the water saturation), and $\nu$ is the kinematic viscosity of the water.

The modified diffusion–advection equation, which is applied to simulation of dissolved constituents transport with the flowing groundwater, is given by:

$$\phi S_w \frac{\partial C}{\partial t} + \mathbf{q} \cdot \nabla C = \nabla \cdot (\phi S_w \mathbf{D}_h \cdot \nabla C) + R + E^{wn} \qquad [8]$$

where $\phi$ is the porosity, $S_w$ is the water saturation and $E^{wn}$ represents interphase mass transfer, namely mass transfer from the entrapped NAPL into the flowing water phase, due to NAPL dissolution.

The chemical reaction term, $R$, of eqn [8] is mainly attributed to adsorption and desorption, as well as biological degradation of the organic compound that is used as a substrate for the development of the biomass. The interphase mass transfer term, $E^{wn}$, of eqn [8] is considered as the major contributor to groundwater contamination due to the presence of the entrapped NAPL in the aquifer. It can also be useful for aquifer cleanup procedures such as pump-and-treat.

Most cases of NAPL contamination in aquifers are attributed to petroleum hydrocarbons, namely fossil fuels. Obviously this phenomenon originates from the widespread use of these types of NAPLs. However, most types of olefins that incorporate the major compounds of fossil fuels do not impose high risks on the environment. Fossil fuel compounds of high risks include various fuel additives and BTEX (benzene, toluene, ethylbenzene, and xylene). Due to the widespread use of fossil fuels, entrapment of LNAPL in the vadose zone and aquifers is a quite common phenomenon. Usually, such an entrapment with comparatively high residual NAPL saturation is limited to the top layers of the aquifer and regions of the capillary fringe. Therefore, the entrapped LNAPL supplies dissolve constituents into the flowing groundwater, as well as volatile organic compounds into the gaseous phase of the vadose zone.

DNAPL contamination of groundwater is less common than LNAPL contamination, and entrapment of DNAPL in aquifers is quite rare. However, the constituents of DNAPLs are usually chlorinated hydrocarbons, which pose a high risk to human health and the environment. Their threshold concentrations are of minute values, of the order of magnitude of several parts per billion. Therefore, the penetration of water solution, which incorporates several parts per million of these compounds, into groundwater may cause a very significant contamination of the aquifer. Furthermore, the chlorinated hydrocarbons are volatile organic compounds (VOCs). The migration of such compounds with groundwater may be associated with the release of poisonous vapors into the vadose zone. Such a phenomenon is somehow mitigated by the immobile water saturation of the capillary fringe. But intensive fluctuations of the groundwater table may significantly reduce the efficiency of this vapor barrier.

## Summary

There is an intimate connection between the pollution of groundwater and the vadose zone. Special attention is given here to calculations of overall mass budgets and flux balance of pollutants for the evaluation of phreatic aquifer contamination due to its use as a source of water for human populations. When water shortage becomes a significant barrier for the development of various societies, it is appropriate to evaluate efficient approaches of increasing the number of water resources and avoiding the deterioration of groundwater quality.

Various methods and means are used for the evaluation of the fate of constituents that are dissolved in water and subject to transport with the water cycle, leading to groundwater pollution. The classification of NAPLs to LNAPLs and DNAPLs is extremely important, as these two types of NAPL have different phenomenological effects on the pollutant penetration into the aquifer. There are also significant differences between their potential risk to human health and the environment, their physical properties, rate of biodegradation, and the possible source of organic contaminants in the liquid and vapor phase in the vadose zone.

## List of Technical Nomenclature

| | |
|---|---|
| $\alpha, \beta$ | indices referring to the particular fluid phase (water, gas, NAPL) |
| $\Delta C$ | difference between final and initial concentration (kg m$^{-3}$) |
| $\Delta C_0$ | pollutant accumulated within $Q_{nr}$ in the natural environment (kg m$^{-3}$) |
| $\Delta C_1$ | pollutant accumulated within infiltrating water in the vadose zone (kg m$^{-3}$) |
| $\Delta C_2$ | pollutant accumulated within complete water-recycling (kg m$^{-3}$) |
| $\Delta C_3$ | pollutant accumulated within $Q_{nr}$ by activities of the human population (kg m$^{-3}$) |
| $\Delta C_4$ | pollutant accumulated within wastewater released into seas and oceans (kg m$^{-3}$) |
| $\epsilon$ | volume fraction |
| $\nu$ | kinematic viscosity (m$^2$ s$^{-1}$) |
| $\rho$ | density (kg m$^{-3}$) |
| $\varphi$ | porosity |
| $\omega$ | mass fraction |
| C | pollutant concentration (kg m$^{-3}$) |
| $C_0$ | initial pollutant concentration (kg m$^{-3}$) |
| $C_\infty$ | final pollutant concentration (kg m$^{-3}$) |
| $E$ | rate of pollutant interphase transfer (kg m$^{-3}$ s$^{-1}$) |
| $D_h$ | hydrodynamic dispersion tensor (m$^2$ s$^{-1}$) |
| K | hydraulic conductivity tensor (m s$^{-1}$) |
| k | permeability tensor (m$^2$) |
| $k_{rw}$ | relative soil permeability for water |
| $Q_{ev}$ | rate of evapotranspiration involved with all types of irrigation (m$^3$ s$^{-1}$) |
| $Q_{ev1}$ | rate of evapotranspiration involved with regular irrigation (m$^3$ s$^{-1}$) |
| $Q_{ev2}$ | rate of evapotranspiration involved with wastewater-recycling irrigation (m$^3$ s$^{-1}$) |
| $Q_{hs}$ | rate of water consumption by the human population (m$^3$ s$^{-1}$) |
| $Q_{ir}$ | rate of water used for regular irrigation (m$^3$ s$^{-1}$) |
| $Q_{nr}$ | rate of water supply into the aquifer by natural recharge (m$^3$ s$^{-1}$) |
| $Q_{rc}$ | rate of water used for irrigation with recycled wastewater (m$^3$ s$^{-1}$) |
| $Q_{so}$ | rate of water supplied by the aquifer into seas and oceans (m$^3$ s$^{-1}$) |
| $R$ | rate of pollutant formation and/or disappearance by chemical reactions (kg m$^{-3}$ s$^{-1}$) |
| $S_w$ | water saturation |
| $t$ | time (s) |
| $W$ | volume of the aquifer (m$^3$) |

*See also:* **Evapotranspiration**; **Groundwater and Aquifers**; **Nitrogen in Soils:** Nitrates

## Further Reading

Bedient PB, Rifai SH, and Newell CJ (1996) *Groundwater Contamination.* Englewood Cliffs, NJ: Prentice-Hall.

Domenico PA and Schwartz (1997) *Physical and Chemical Hydrology*, 2nd edn. New York: John Wiley & Sons, Inc.

Rubin H and Atkinson JA (2001) *Environmental Fluid Mechanics.* New York: Marcel Dekker.

Rubin H, Narkis N, and Carberry J (eds) (1998) *Soil and Aquifer Pollution (Non-Aqueous-Phase-Liquids – Contamination and Reclamation).* Heidelberg, Germany: Springer-Verlag.

Rubin H, Nachtnebel HP, Fuerst J, and Shamir U (eds) (2002) *Water Resources Quality – Preserving the Quality of our Water Resources.* Heidelberg, Germany: Springer-Verlag.

Sun NZ (1996) *Mathematical Modeling of Groundwater Pollution.* New York: Springer-Verlag.

# Industrial

**S P McGrath**, Rothamstead Research, Harpenden, UK

## Introduction

The term 'pollution' is often used, but seldom defined. A search of the literature reveals that it is often used synonymously with the term 'contamination.' Is there a distinction between pollutants and contaminants? For example, the US Environmental Protection Agency (EPA) defines pollution as "Generally, the presence of a substance in the environment that because of its chemical composition or quantity prevents the functioning of natural processes and produces undesirable environmental and health effects." Under the US Clean Water Act, pollution has been defined as the man-made or man-induced alteration of the physical, biological, chemical, and radiological integrity of water and other media. The word 'integrity' again seems to infer some harm has occurred. However, the US EPA also gives a definition of the term 'contamination' as "Introduction of microorganisms, chemicals, toxic substances, wastes, or wastewater in a concentration that makes the medium unfit for its next intended use." Once more, some damage is inferred, but here microorganisms are specified in addition to chemicals. It would be useful if microorganisms or presence and absence of harm gave a distinction between these terms, but many of the published definitions and recent usage show they are in fact considered synonymous. So although a few authors have tried to distinguish between pollution and contamination, here, the focus will be on the addition of chemical substances to soil that affect its use or affect other parts of the environment through soil.

Time and rate of addition of a substance to soil are important, as substances can be degraded, leached, or volatilized, thereby being removed from soils. Soils can immobilize pollutants, and it is only when the inputs far outweigh these loss pathways that the soil itself will become polluted. Of course, soils may in turn become the source of a pollutant rather than a sink when losses to water or volatilization take place. Another important concept relating to which chemicals become pollutants in soil is the idea of residence time. Because some potential pollutants arriving in soils can be rapidly degraded, leached, or volatilized, they may never become a serious problem. Examples would be some of the easily degraded organic compounds, which disappear quickly, but on the other hand some recalcitrant organic and inorganic pollutants build up in soil because there is little loss by any pathway.

The importance or magnitude of pollution depends on factors such as the rate of addition of the pollutants per unit area. Each route or pathway that pollutants follow to reach a site needs to be summed. Pathways may include wastewater from industry, solid wastes, emission to the air and subsequent deposition, and leaks and spills (**Figure 1**). Concentrations and total loads in the pathways minus the loss rates due to the processes mentioned above and

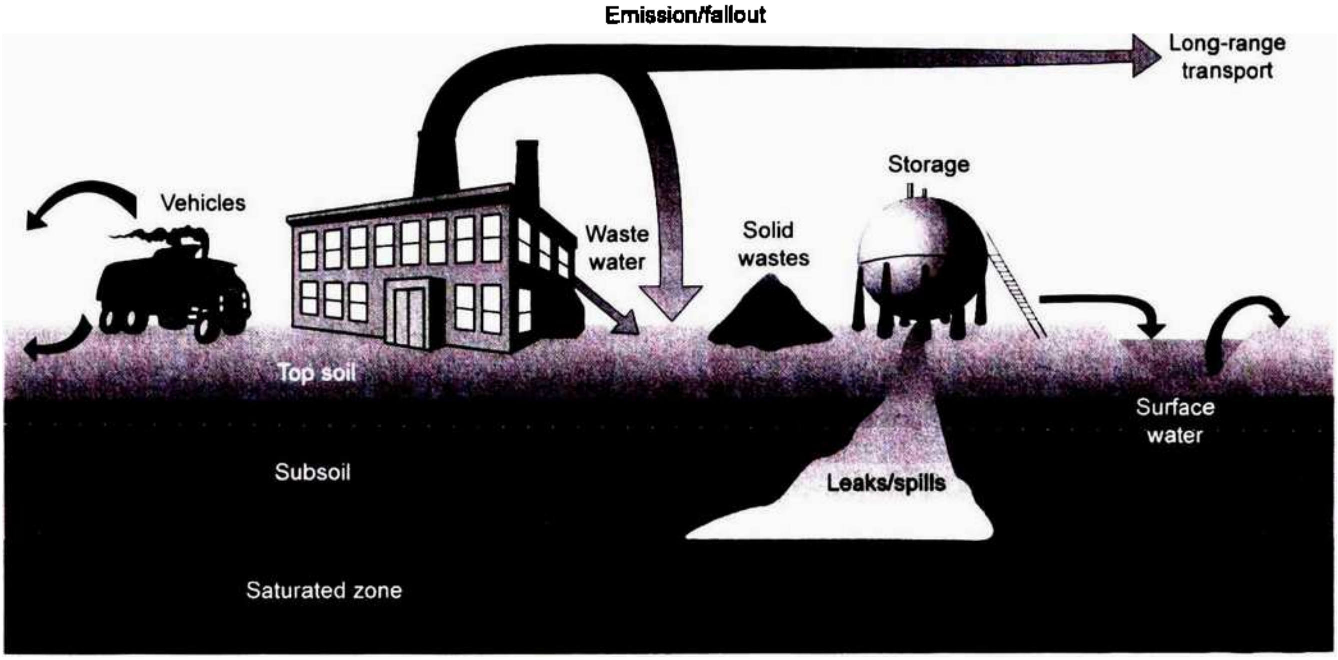


**Figure 1** Pathways of industrial pollutants in the soil environment.

**Table 1** Estimated number of contaminated sites in Europe and the United States

| *Country* | *Contaminated sites (estimate of total)* | *Contaminated sites (in critical condition)* |
|---|---|---|
| Austria | 3300 | – |
| Belgium | 8300 | 2000 |
| Denmark | 20 000 | 3600 |
| Germany | >100 000 | 10 000 |
| Italy | 5600 | 2600 |
| Netherlands | 110 000 | 4000 |
| European Union | 1500 000 | 300 000 |
| Norway | 25 000 | – |
| United Kingdom | 100 000 | – |
| United States | 37 600[a] | >1300[b] |

Source Adriano DC (2001) *Trace Elements in Terrestrial Environments. Biogeochemistry, Bioavailability and Risks of Metals.* New York: Springer; Eijsackers HJP and Hamers T (eds) (1993) *Integrated Soil and Sediment Research.* Dordrecht. Kluwer; Overcash M (1996) European soil remediation research: 1992–1994. *Critical Reviews in Environmental Science and Technology* 26. 337–368; European Environment Agency (2001) *Indicator Fact Sheet Signals 2001 – Soil.*

[a] Reported to the US EPA in 1993.

[b] National priority list ('Superfund sites').

time of exposure all affect the buildup of the pollution and hence its importance.

Because of the difficulty of determining whether sites are polluted, there are only estimates in relatively few countries of the extent of what is normally referred to as 'contaminated land' (**Table 1**). For example, in the UK it is estimated that 300 000 ha of land are affected to some extent by contamination left by industrial activities. In the European Union 300 000 sites have already been identified as potentially or definitely contaminated, and the best estimate is that there are 1.5 million contaminated areas.

Excluding pesticides and radionuclides, the most common pollutants found at US Superfund sites include: acetone, arsenic, barium, benzene, 2-butanone, cadmium, carbon tetrachloride, chloroform, chromium, cyanide, 1,1-dichloroethene, 1,2-dichloroethane, lead, mercury, methylene chloride, napthalene, nickel, pentachlorophenol, polychlorinated biphenyls (PCBs), polycyclic aromatic hydrocarbons, tetrachloroethylene, toluene, trichloroethylene, vinyl chloride, xylene, and zinc. Common radionuclides at Superfund sites are: americium-241, cesium-137, cobalt-60, iodine, plutonium, radium, radon, strontium, technetium-99, thorium, tritium, and uranium.

## Origins of Industrial Pollution

Many different industries and processes emit pollutants into the environment. Use of agricultural chemicals, such as nitrate, and pesticides will not be discussed here. Activities that are nonprimary or 'commercial' also result in emissions and pollution of soil, examples being transport and dredging of waterways, where pollutants are often deposited on the sides and banks of roads or along the banks of rivers. Industrial emissions may also move through air or water to sewage-treatment facilities, where they are often concentrated in the sewage sludge solids that are frequently utilized on land (*See* **Waste Disposal on Land:** Municipal). However, there are also domestic inputs of some pollutants to sewage.

The European Environment Agency considers that most cases of soil contamination arise from the following industries: chemicals, metals, energy, mining, oil, electronics, glass, ceramics, stone, textile, leather, wood, paper, food, trade, and traffic. Some of the activities likely to cause pollution of soils are shown in **Table 2**. Some of these arise from point sources such as factories, smelters, waste pipes, ditches, or mines. These are stationary sources and may be small or large. One feature of point sources is that the concentrations of pollutants can be high and may build up rapidly. They may also fall off rapidly with distance from the source. Nonpoint sources of pollution are not easily associated with or as regulated as point sources and normally originate from such activities as agricultural practices, urban runoff, and transportation.

## Types of Pollutant

Many thousands of pollutants have been dispersed in the environment by human activities, some for a long time. They are broadly classified as inorganic and organic pollutants. However, depending on the source and other local emissions, multiple or 'copollution' of soils may occur. An example would be the land around a factory that used multiple chemicals in its processes or land that has been used by different industries over time.

**Table 2** Activities likely to cause pollution in soils

| *Activity* | *Examples* | *Likely inorganic pollutants* | *Likely organic pollutants* |
|---|---|---|---|
| *Mineral extraction/processing* | | | |
| Coal | Deep or open cast mines, crushing | Acidity, pyrites, sulfur | Dust |
| Metals | Ferrous and nonferrous, deep or open cast, crushing, waste lagoons | All metals and associated anions | |
| Other minerals | Asbestos, chalk, china clay, gypsum | Sulfides, sulfates | |
| *Waste disposal* | | | |
| Municipal | Domestic waste, sewage sludge | Many, including leachates | Many, including leachates |
| Industrial | Unused raw materials, by-products, spent catalysts | Many, including leachates | Many, including leachates |
| *Industries* | | | |
| Brick works | | Fluoride, sulfur dioxide | |
| Chemical works | Battery, electronics, and pigments manufacture, electroplating | Acidity, alkalinity, As, asbestos, Cd, Cr, Cu, cyanide, Pb, Hg, Ni, sulfate, Zn | Dioxins, halogenated hydrocarbons, pesticides, phenols, PCBs, solvents |
| Gasworks | Coal gasification | Sulfur, cyanides | Phenols, tar, PAHs |
| Iron and steel | Blast furnaces | Metal wastes, iron and steel slag | |
| Nonferrous smelting/ processing | Lead, zinc, copper smelters | Pb, Zn, Cu, Hg, Cd, Ni, acidity | |
| Oil refineries | Petroleum production | Lead compounds | Tar, oils, phenols, sulfate |
| Power generation | Coal, gas, nuclear plant | Fuel ash, fly ash, various radionuclides | PCBs, PAHs |
| Recycling/scrap | Batteries, cars, domestic appliances | Acidity, asbestos, Pb, Ni, Cd | Oils, paint, phenols, PCBs |
| Sewage treatment | Various processes | Metals, phosphate | Various |
| Tanneries | | Chromium | Dyes |
| *Military bases* | | | |
| Air/army/navy bases | Munitions, anticorrosion or antifouling chemicals, preservatives | Various | Various, fuels |
| Defense establishments | | Various | Various |
| Firing ranges | | Pb, various | Various |
| Transportation | | | |
| Airports | Runoff, workshops | | Fuel, oil, deicing compounds |
| Canals, waterways, docks | Dredged sediments, dumped waste | Acidity, alkalinity, antifouling compounds, metal wastes | Contaminated sediments |
| Railways | Railways, borders and sidings/yards | Metal wastes, asbestos | Oil, paint |
| Roads | Borders, central reservation, runoff areas | Pb, Cd, Zn | PAHs |
| *Energy supply* | | | |
| Pipelines | Oil pipelines | | Oil leaks |
| Electricity supply network | Cables and towers, substations, transformers | Zn, Cu | PCBs |
| Oil/petroleum storage | | Pb | Tar, petroleum, oils, phenols |

PCBs, polychlorinated biphenyls; PAHs, polycyclic aromatic hydrocarbons.

### Inorganic Pollutants

Materials that are not based on carbon compounds are termed 'inorganic pollutants.' It is important to realize that some elements are in fact essential nutrients required by organisms, and that they can be deficient in some soils or diets; examples include Zn, B, Cu, Se. A consequence of this is that additions of nutrients will at first be beneficial for the growth and reproduction of organisms. It is only when they become excessive that negative impacts will occur.

This typically occurs because the soil's ability to hold these substances is overwhelmed, along with the ability of the organisms in the ecosystem to cope with the oversupply.

### Organic Pollutants

Organic molecules in the environment are usually classified depending on whether they are persistent. For example, solvents may almost completely sterilize the soil if added in large quantities, but they typically volatilize in a relatively short time and the soil may then recover. Some compounds used in industry, such as surfactants and plasticizers, can have disruptive effects on the endocrine system. However, these normally act through the water pathway or directly in packaged and processed food and are not thought to persist in soils. The term 'persistent organic pollutants' (POPs) refers to those which do not degrade, leach out, or volatilize from soils readily. This group can, therefore, build up to concentrations in soil that may inhibit biological processes. Characteristics that make a compound persistent in soil include its low aqueous solubility, low vapor pressure, and a high octanol to water partition coefficient (*See* **Pollutants: Persistent Organic (POPs)**). These characteristics usually mean that only small amounts of POPs are taken up by plant roots because of their low solubility and tendency to bind to organic matter in soil. However, grazing animals and some invertebrates such as earthworms may mobilize POPs from soils in their digestive systems when they ingest soil.

### Examples of Pollutants

Many industrial and commercial sources emit pollutants to air and water or direct to soil. Some are short-lived because they are degraded and others leach rapidly in environments with excess rainfall over evapotranspiration. However, others bind to organic matter, clay, and oxides or form insoluble compounds in surface soils and so tend to remain and accumulate in soil with time. Those that have this tendency are metals, metalloids, persistent organic compounds, asbestos, coal tar, fuel ash, and some radionuclides (*See* **Pollutants: Persistent Organic (POPs)**). Some of the industrial sources of these pollutants are shown in **Table 3**. It is important that any 'new' substances being used in industrial processes or mobilized in the environment are considered also for their persistence and effects in soil.

## Legislative Controls on Pollution

The ultimate environmental control method is to prevent the dispersion of pollutants in the environment. This has resulted in complete bans on the production of some compounds perceived to be very harmful. For example, manufacture of PCBs in the USA was banned by the EPA in 1977, though it continued under considerable restrictions in other countries until the mid-1980s. Even after production of any substance ceases, it will continue to be present in products, and thus present a disposal hazard, for a considerable period, e.g., PCBs in the insulation of electrical transformers. Legislation needs to be coordinated such that removing a toxic pollutant from air (e.g., by stopping incineration of certain materials) does not lead to transfer of the pollutants to soil or water. On the other hand, legislation for some classes of pollutants is based on containment or controls on the rate of emission to air or water, or on the purposeful additions of materials to land.

### Source–Pathway–Receptor Principle

As mentioned in the Introduction, the term 'pollution' infers that some harm has already been done or will take place. Recent legislation and schemes to determine whether land is polluted or contaminated take a risk assessment approach to evaluate potential harm. The source–pathway–receptor principle requires that all three aspects are established at a site and that there is significant harm. Source is an origin of pollution, which could be wastewater or contaminated soil in the present context. The pathway is a means by which the pollutant can reach and affect a receptor. A receptor is an organism or group of organisms, an ecological system, or a piece of property which is being or could be harmed by a pollutant. A 'pollutant linkage' has to be established, which means the relationship between the pollutant, pathway, and receptor. In some cases there may be more than one pathway (for example, dust and water contamination). Another valuable aspect of the pollutant linkage idea is that the breaking of this linkage can be a very important way of stopping pollution or remediating contaminated sites. Prevention can be brought about by stopping the source or putting a barrier in the pathway to the receptor (*See* **Remediation of Polluted Soils**).

Also, the bioavailability of the pollutant to an organism is an important consideration in the pathway: because a pollutant is present in soil (or another medium) does not necessarily mean that it is likely to transfer and affect an organism. In soils, binding of pollutants by organic matter or soil minerals may render them sparingly or nonbioavailable (*See* **Pollutants: Persistent Organic (POPs)**). An example is chromium; two forms occur in the environment, with valency states III and VI. In soils however,

**Table 3** Pollutants likely to occur in soils in contaminated environments

| *Pollutants* | *Industrial sources* |
|---|---|
| *Metals/metalloids* | |
| Antimony | Electronics, metal smelting |
| Arsenic | Insecticide manufacture and use, wood preservation |
| Beryllium | Nuclear, munitions, and electronics |
| Cadmium | Pigments, alloys, electroplating, batteries, mining, and smelting industries |
| Chromium | Plating, tanning, inks, wood treatment, pigments, and dyes |
| Copper | Mining and smelting, wood treatment, electronics, plumbing, copper and brass manufacture |
| Lead | Batteries, tetraethyl lead production, mining, smelting, plumbing |
| Mercury | Chemical plant, mining and smelting, refining gold and silver, electrolysis industry, chlor-alkali plant, paper mills |
| Nickel | Plating and metal works, batteries, iron and steel industries, electronics |
| Selenium | Smelting, paint, pigment, electronics |
| Silver | Mining, photographic, electroplating, silver manufacture |
| Thallium | Cement industry, electronics |
| Zinc | Mining and smelting, plating, galvanizing, runoff from galvanized surfaces, print and paper |
| *Other inorganic pollutants* | |
| Fluoride | Cement and brick production, aluminum smelting, phosphate fertilizer manufacture |
| Cyanides | Coal gasification, metallurgical processes, solvent/separation of gold |
| Sulfur | Sulfides, residues from coal processing |
| *Organic pollutants* | |
| *Aromatic compounds* | |
| Benzene | Tar, gas, or coke works, petroleum refineries |
| Ethyl benzene | Tar, gas, or coke works, petroleum refineries |
| Toluene | Tar, gas, or coke works, petroleum refineries |
| Xylene | Tar, gas, or coke works, petroleum refineries |
| Phenols | Tar, gas, or coke works, petroleum refineries |
| *Polycyclic aromatic hydrocarbons (PAHs)* | Combustion of coal hydrocarbons and wastes |
| *Chlorinated organic compounds*[a] | Pesticide manufacture, deodorants, timber and textile peservatives, transformer insulation |
| *Dioxins and furans* | Sewage sludge application, waste disposal, waste incineration |
| *Other pollutants* | |
| Asbestos | Insulation in buildings, boats, other transportation, power stations |
| Coal tar | Tar, gas, or coke works, petroleum refineries |
| Fuel ash | Coal-fired power generation |
| Protons | Oxidation of sulfur dioxide and metal sulfides |
| Radionuclides | Power generation, weapons production, phosphate manufacture |

[a] Aliphatic chlorinated compounds, chlorobenzenes, chlorophenols, polychlorinated biphenyls.

chromium is usually reduced to chromium III, which is insoluble, and as a result little is bioavailable. In some circumstances in aquatic systems, chromium VI may occur, which is bioavailable and very toxic.

## Persistent, Bioaccumulative, and Toxic Pollutants

One way of prioritizing pollutants is to choose those that are known to be persistent, bioaccumulative, and toxic (PBT). These are substances that can build up in the soil and the food chain to levels that are harmful to human and ecosystem health. Other features related to their chemical characteristics are that they may travel long distances, and transfer easily between air, water, and soil, and therefore reside for a long time in the environment. In addition to some pesticides, the US EPA's first priority list of PBT pollutants includes the following industrial pollutants: benzoapyrene, hexachlorobenzene, alkyl lead, mercury and its compounds, octachlorostyrene, PCBs, dioxins, and furans. Because of their properties, PBTs have attracted many voluntary or legal controls on their production, use, and entry into the marketplace to decrease their emission in the environment.

Statistics show that the production of some potential pollutants (for example, nonferrous metals) has increased continually over the last decade. However, due to the legislative controls discussed above, and because of controls on emissions of chemicals to the environment, the point and nonpoint (diffuse) sources of pollution are decreasing in the developed countries. However, it should be remembered that, in many rapidly developing countries without strong environmental controls, pollution is actually increasing.

*See also*: ; **Pollutants:** Persistent Organic (POPs); **Remediation of Polluted Soils**; **Waste Disposal on Land:** Municipal

## Further Reading

Adriano DC (2001) *Trace Elements in Terrestrial Environments. Biogeochemistry, Bioavailability and Risks of Metals*. New York: Springer.

Bridges EM (1987) *Surveying Derelict Land*. Oxford: Clarendon Press.

European Environment Agency (2001) *Environmental Signals 2001*. Copenhagen: EEA.

Jones KC (ed.) (1991) *Organic Contaminants in the Environment. Environmental Pathways and Effects*. London: Elsevier Applied Science.

Knox AS, Gamerdinger AP, Adriano DC, Kolka RK, and Kaplan DI (1999) Sources and practices contributing to soil contamination. In: *Bioremediation of Contaminated Soils*, pp. 53–87. Agronomy Monograph No. 37. Madison, Wisconsin: American Society of Agronomy.

UK Department of the Environment, Transport and the Regions (2000) *Environmental Protection Act 1990: part IIA, Contaminated Land*. DETR Circular 02/2000. London: DETR.

---

# POLYMERS AND MICROORGANISMS

**M C Rillig**, University of Montana, Missoula, MT, USA

## Introduction

Extracellular polymers, organic macromolecules found outside the cell or hyphal wall, are a common feature in eukaryotic and prokaryotic soil microbes. This summary deals with four main groups of extracellular microbial polymers, believed to be of particular significance in soil biology and the interaction of soil microbes with their environment: exopolysaccharides (EPS), hydrophobins (a group of small fungal proteins), glomalin (a glycoproteinaceous substance produced by arbuscular mycorrhizal fungi, AMF), and extracellular lipid compounds (for example, glycolipids such as rhamnolipids). Extracellular polymers are present in the soil in complex mixtures (spatially and temporally), being produced by a variety of microbes and being subject to modifications after entering the soil environment. They fulfill a range of functions in soil, from biofilm formation and modification of the habitat to increasing solubility of carbon substrates. In addition to the groups discussed here, extracellular enzymes are clearly a crucial component of the soil environment for microbes with absorptive nutrition (such as soil fungi and bacteria), but these are examined in detail elsewhere (*See* **Enzymes in Soils**). Polymeric compounds secreted by microbes in soil are also pertinent in the context of formation and composition of organic matter (*See* **Organic Matter:** Principles and Processes), and in the process of soil aggregation (*See* **Aggregation:** Microbial Aspects).

## Extracellular Polymers: Definition and Overview

Commonly microbes in soil are embedded in a matrix of extracellular polymeric substances; the production of these substances is a common feature of archaea, Eubacteria, and eukaryotic microorganisms (e.g., algae, fungi). As a result of the production of these substances, biofilms are formed. Biofilms are mostly discussed in the aquatic microbial ecology literature, because they are uniquely suited to the challenges of inhabiting an aquatic habitat. Clearly, biofilms with their exopolymer matrix occur in soil as well, and can be seen by electron microscopical techniques (see **Figure 1**). Biofilms are accumulations of microbes (can be mixed communities), extracellular polymeric substances, multivalent cations, and various biogenic particles, as well as dissolved compounds.

Extracellular polymeric substances have been defined as "substances of biological origin that participate in the formation of microbial aggregates." Other definitions highlight the role these substances play in cohesion of microbes and adhesion (binding to the substratum). However, these more structural roles are not the only functions of extracellular polymers (see **Table 1**).

Extracellular polymers include organic macromolecules (formed by polymerization of similar or identical building blocks) such as polysaccharides, proteins, nucleic acids, and lipids. These polymers may contain nonpolymeric substituents (which can sometimes greatly influence the properties of the polymer). For example, proteins can be glycosylated or there can be acetyl or succinyl groups in polysaccharides.

These polymers are found by definition outside the cell wall, where they may have arrived through active

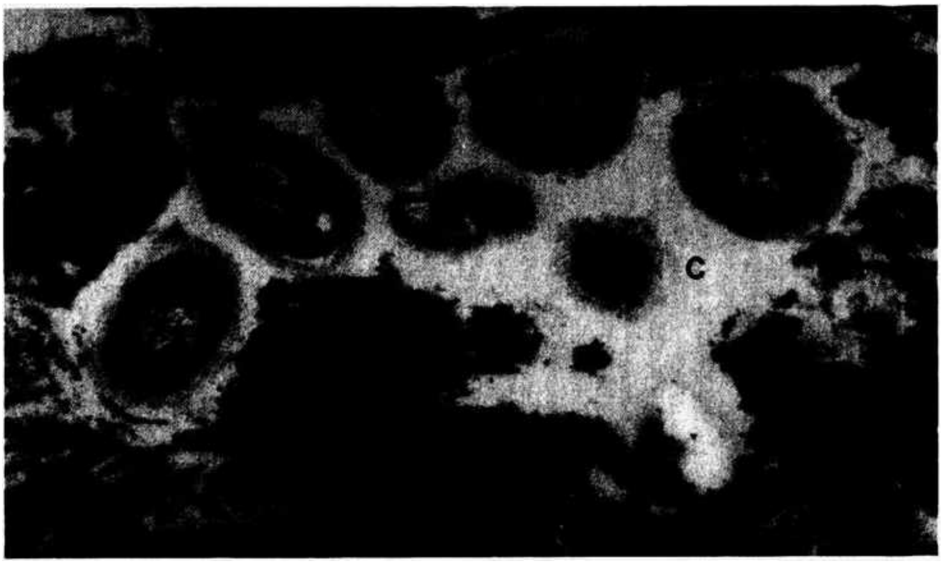


**Figure 1** Electron micrograph showing a colony of bacteria (B) surrounded with exopolymer material (C, capsule material) in the rhizosphere of clover. Lanthanum stain, ×30 000. Reproduced with permission from Emerson WW *et al.* (1978) *Modification of Soil Structure*. Copyright 1978. © John Wiley & Sons Limited.

**Table 1** Examples of the variety of functions of microbial exopolymers in soil

| *Exopolymer* | *Function* |
|---|---|
| Exo- and ectoenzymes | Crucial component of absorptive nutrition of microbes |
| Hydrophobins | Lowering water surface tension, allowing hyphal emergence into air |
| Hydrophobins, some exopolysaccharides | Microbial attachment/adhesion to surfaces |
| Glomalin, many other exopolymers | Stabilization/formation of soil aggregates: involvement in the modification of the 'macro'-environment; importance in soil carbon storage via stabilization of labile carbon inside aggregates |
| Some rhamnolipids | Biocidal (antiviral, antibacterial, antifungal) activity |
| Exopolysaccharides, many other exopolymers | Biofilm formation: aggregation of microbes (increases potential for communication, synergism, gene transfer); leading to enhanced sorption of nutrients (inorganic, organic); retention of exoenzymes |
| Exopolysaccharides | Enhanced desiccation resistance |
| Some rhamnolipids | Surfactant activity (enhanced degradation of hydrocarbons; biosurfactants) |
| Some exopolysaccharides | Symbiotic signal (e.g., *Rhizobium*); also pathogen recognition |
| Attached exopolymers | Modification of microbial cell/hyphal wall properties (e.g., hydrophobicity, charge) with potential consequences for microbial transport, nutrient absorption |
| Exopolysaccharides, other exopolymers | Protection from toxic compounds, e.g., metals, biocides |
| Attached expolymers | Food web processes: protection from grazers, predation (e.g., amebae) |

secretion processes (for example, through the vesicle-mediated pathway at the growing tip of fungal hyphae), sloughing off of cell-surface material, or lysis of microbial cells. Additionally, organic compounds from the environment (e.g., from lysis and hydrolysis) can become part of the polymer matrix.

Once extracellular, polymers are subject to various modifications by enzymatic processes. Various polymer-degrading enzymes such as hydrolases can be present in the matrix. Polymer-breakdown products can sometimes be reutilized by microbes. Due to these processes within a biofilm, the composition of extracellular polymers can be variable both spatially and temporally.

Extracellular polymers can also be broadly distinguished into those that are attached (bound) and nonattached. Bound polymers include capsules and sheaths of bacteria, and loosely bound polymers, while soluble polymers include slime and colloids, as well as other water-soluble macromolecules.

Historically, extracellular polymers were studied by extraction methods. Fractions were defined operationally using the extraction method of choice, e.g., centrifugation, boiling, ultrasonication, filtration, or solubility in various solvents (e.g., NaCl, NaOH, organic solvents). This work was mostly performed using laboratory pure cultures of microbes. In addition, there are now available techniques for the *in situ*, nondestructive study of these polymers, including infrared spectroscopy (e.g., attenuated total reflectance/Fourier transform infrared spectroscopy (ATR/FT-IR), detects signals of surface chemical groups), nuclear magnetic resonance (NMR), and confocal laser scanning microscopy (CLSM). These methods have greatly increased the understanding of the three-dimensional structure and composition of extracellular polymer matrices.

In summary, extracellular polymers occur in complex mixtures of compounds in the environment (to which frequently more than one species or group of microbes can contribute); they can arrive at a certain location by a variety of different mechanisms and may not have been formed where they are found; these mixtures are then subject to processing and modification in the environment as a function of time and space, and the structure of the polymer matrix can also change in response to environmental conditions. It is very important to keep this complexity in mind in the following discussion of selected components of these extracellular polymer mixtures.

## Hydrophobins: Secreted Fungal Proteins

### Structure and Properties

Hydrophobins are relatively recently discovered, small (ca. 100 amino acids), secreted, moderately hydrophobic proteins produced by fungi in the *Ascomycota* and *Basidiomycota*. Their occurrence in other Eumycotan phyla (*Chytridiomycota*, *Zygomycota*) is not yet clearly established. Based on hydropathy patterns (hydropathy is a measure of the hydrophobic character of an amino acid) and solubility characteristics, hydrophobins have been divided into class I and class II (the latter having been found only in the *Ascomycota* so far). Sequence homology of hydrophobins is very low (especially among class I hydrophobins), but all proteins share eight conserved cysteine residues; these cysteine residues form intramolecular disulfide bridges. Some hydrophobins, like the intensively studied SC3 (isolated from *Schizophyllum commune*, a basidiomycete fungus), are glycosylated. Several fungi have more than one hydrophobin gene. For example, *S. commune* contains at least four hydrophobin genes, which are expressed at different fungal life-history stages (e.g., monokaryotic versus dikaryotic mycelium). Six hydrophobin genes have been identified in *Cladosporium fulvum*. Complementation studies showed that some hydrophobins could only partially substitute for each other, suggesting that hydrophobins may have evolved to fulfill different functions.

All hydrophobins so far studied self-assemble at hydrophilic–hydrophobic interfaces, such as a water–air boundary. This self-assembly is accompanied by conformational changes of hydrophobin monomers that form the amphipathic film or membrane.

Virtually all studies performed with hydrophobins appear to have been conducted under *in vitro* conditions to study molecular and biochemical aspects of their production; hence there is little information about the amount and behavior of hydrophobins found in soils or their persistence in the soil environment.

### Functions

Hydrophobins are believed to be essential in several functions in the fungal life-cycle and development. These proteins, by virtue of their ability to self-assemble into amphipathic membranes, allow hyphae to break the water–air barrier (overcoming surface tension), coat air-exposed fungal structures with a hydrophobic film, and allow hyphae to attach to hydrophobic surfaces. Hydrophobins have also been found to alter hyphal wall composition.

**Hydrophobins allow fungi to escape their aqueous environment** This is a necessary step for the formation of fungal fruiting bodies. However, mycelium growth in soil is also confronted with air–water boundaries as hyphae cross pore spaces between aggregates. The importance of secreted hydrophobins such as SC3 in hyphal growth across a water–air barrier lies in the lowering of water surface tension (Figure 2). A strain of *S. commune* with the *SC3* gene disrupted produces few aerial hyphae, but by adding SC3 protein to the culture medium, water surface tension is decreased to wild-type levels, and formation of aerial hyphae is possible. The time course of decrease in water surface tension of a culture of *S. commune* also correlates with the production and secretion of SC3 into the medium. The mechanistic details of emergence of hyphae as mediated by a hydrophobin-membrane are not known.

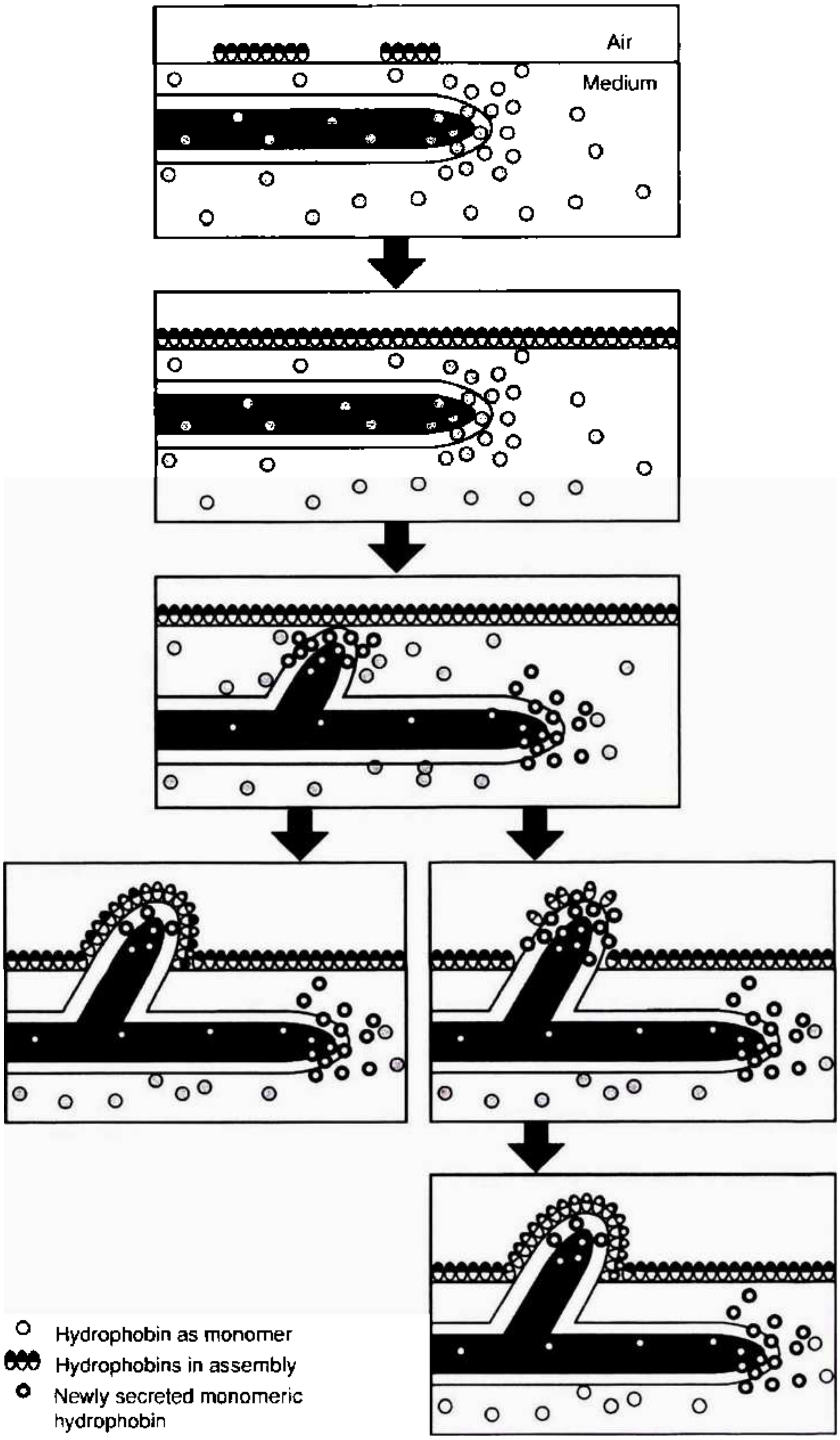


**Figure 2** Functioning of fungal hydrophobins at the air–water interface. Monomeric hydrophobins are secreted into the aqueous medium. These spontaneously assemble at the air–water interface, where they greatly decrease the surface tension, allowing hyphae to emerge. In the process of emergence, it is unclear whether hyphae just stretch the hydrophobin film that has formed at the interface (bottom left) or if the film is ruptured as the hypha emerges (bottom right). Hydrophobins coat the hypha exposed to the air. Reproduced from Wösten HAB (2001) Hydrophobins: multipurpose proteins. *Annual Reviews of Microbiology* 55: 625–646. With permission, from the *Annual Review of Microbiology*, Volume 55 © 2001 by Annual Reviews www.annualreviews.org.

# Glomalin: An Abundant Glycoproteinaceous Substance Produced by Arbuscular Mycorrhizal Fungi

## Properties

Glomalin is a relatively recently discovered glycoproteinaceous substance (of as yet unknown amino acid sequence and structure) produced apparently only by AMF. AMF are obligate biotrophs (in the phylum *Glomeromycota*) living in symbiosis with the roots of ca. two-thirds of higher plants; they are ubiquitous in soils of terrestrial ecosystems (*See* **Mycorrhizal Fungi**).

Glomalin is, like many substances found in soil (e.g., humic acid), operationally defined, and hence definitions are subject to change once structural details become known. Glomalin is the protein (or the proteins, since it is somewhat unlikely that all AMF species produce exactly the same substance) that reacts with a monoclonal antibody (MAb32B11) raised against crushed spores of the AMF species *Glomus intraradices*, and as the protein (as determined with a Bradford assay) that is extracted with $50\,\mathrm{mmol\,l^{-1}}$ sodium citrate (pH 8.0) in successive cycles of autoclaving (121°C) until the supernatant is clear. Several fractions of glomalin are distinguished, based on their immunoreactivity (i.e., if the protein can be detected using an enzyme-linked immunosorbent assay (ELISA) with MAb32B11) and ease of extraction from soil (easily extractable glomalin is extracted in an initial round of autoclaving for only 30 min with $20\,\mathrm{mmol\,l^{-1}}$ sodium citrate, pH 7.0).

Defined in this way, glomalin occurs in very large amounts in soil, typically in the range of several to $10\,\mathrm{mg\,g^{-1}}$ soil (in one case $60\,\mathrm{mg\,cm^{-3}}$ soil was found). Glomalin decomposes slower than the AMF hyphae producing it, and the estimated turnover time for glomalin in soil is on the order of several decades (derived from chronosequence studies and C-14 carbon dating). Glomalin is also produced in sterile *in vitro* cultures of AMF, but most of the work has been done in pot cultures or soil.

Glomalin appears to have N-linked oligosaccharide side chains, and the extract also contains Fe in variable percentages (it is not known in what way Fe is bound in the molecule). In its native state, glomalin is assumed to be insoluble and hydrophobic. Evidence (e.g., from NMR studies) is mounting that glomalin is a compound distinct from humic acid; its $^1$H NMR spectrum is also distinctly different from tannic acid. The pathway of release of glomalin into soil is not clear; glomalin could be actively secreted at the hyphal tip, or it could be sloughed off hyphae as they grow through soil.

It is not known what factors control glomalin production. However, in an experiment with AMF *in vitro* cultures, glomalin concentrations were significantly higher when AMF hyphal growth was reduced compared with control, perhaps suggesting that glomalin is not constitutively produced but that its production is subject to environmental controls. There is also evidence that AMF species differ in glomalin production rates. Glomalin concentrations in soil are responsive to several disturbances, including various agricultural practices (e.g., crop rotations, tillage, etc.), and factors of global change (e.g., elevated atmospheric $CO_2$, warming).

## Functions

**Soil aggregation** Possibly the most important function of glomalin from the ecosystem perspective, and the best-documented overall, relates to its role in soil aggregation. While the mechanisms are not clearly understood, there is a strong and consistent correlation of glomalin concentrations in soil (or soil aggregates) with soil-aggregate water stability (a measure of resistance of aggregates to disintegration when sieved in water). The correlation is frequently strongest using the immunoreactive easily extractable glomalin (IR-EEG) fraction of glomalin. The relationship of glomalin concentration with soil aggregate water stability is curvilinear such that at higher glomalin concentrations no additional aggregate stability is achieved (**Figure 3**). Microscopic observation shows that aggregates are irregularly coated with glomalin (as identified with MAb32B11; see **Figure 4**), suggesting that a possible mode of action involves sealing off

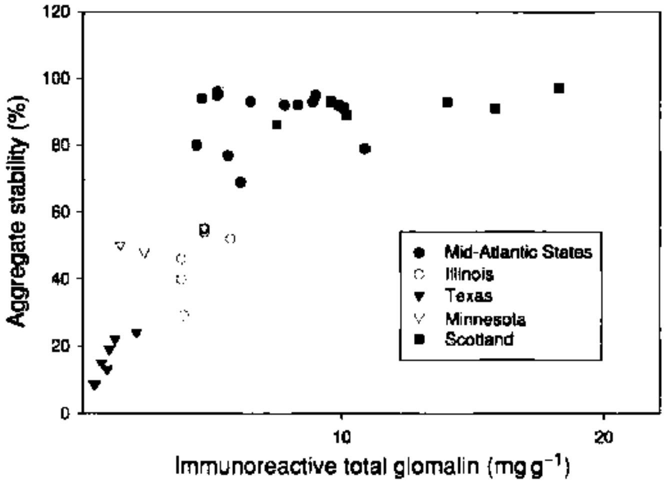


**Figure 3** Curvilinear relationship of glomalin concentrations with soil aggregate water stability (1- to 2-mm aggregates) across a variety of soils. Reproduced with permission from Wright SF and Upadhyaya A (1998) A survey of soil for aggregate stability and glomalin, a glycoprotein produced by hyphae of arbuscular mycorrhizal fungi. *Plant and Soil* 198: 97–107 (Figure 3). With kind permission of Kluwer Academic Publishers.

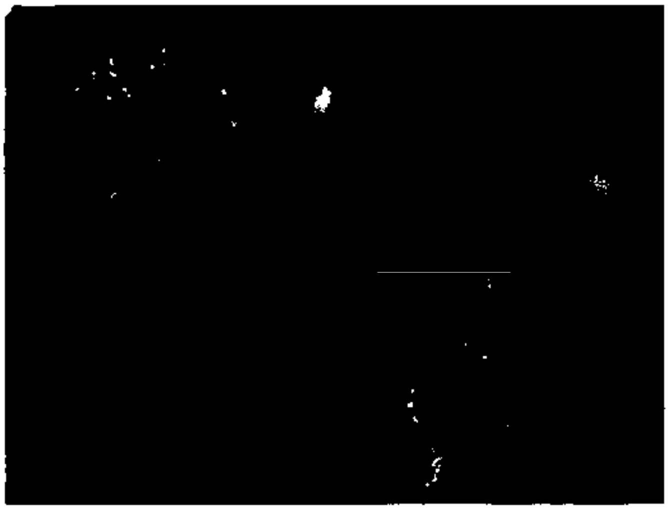


**Figure 4** (see color plate 52) Glomalin (visualized as immunofluorescence using MAb32B11) on the surface of arbuscular mycorrhizal fungal hyphae on soil aggregates (1–2 mm diameter). Reproduced with permission from Wright SF (unpublished).

aggregates to prevent intrusion of water into the aggregate interior. Additionally, glomalin may make AMF hyphae 'sticky' so that the enmeshment mechanism of hyphae could be more efficient. Importantly, the relatively slow turnover of glomalin in soil may explain why aggregates formed by AMF are typically stable for a longer period than aggregates formed by other fungi, for example saprobes.

**Role in the biology of AMF and symbiosis** It is remarkable that plants and their AMF symbionts invest a significant amount of carbon into glomalin, begging the question what might be physiological roles of the protein. AMF hyphae are less palatable to grazing collembolans than hyphae of other, for example, saprobic fungi; it is hypothesized that glomalin may play a role in the decreased palatability of AMF hyphae in soil. The growth of AMF may be enhanced in aggregated soil compared with soil with no macropore spaces, although this is difficult to show experimentally (due to the presence of confounding factors when comparing aggregated and nonaggregated soil). The investment into glomalin may then serve to increase hyphal growth rate via the effect of the substance on soil aggregation. Results consistent with this idea have been obtained in an artificial bead-culture system. Other hypothesized functions include the adsorption of metals to the protein, and a role in the Fe economy of the rhizosphere.

## Exopolysaccharides

EPS are extracellular polysaccharides that occur widely among prokaryotes; bacterial EPS have been studied in greatest detail. Also most microalgae and many filamentous fungi have EPS. EPS are abundant extracellular products that either accumulate on the surface of the cell (for example as a capsule) or that are released into the cell's surroundings. EPS originate from cytoplasmic sugars, and in some cases (e.g., succinoglycan in *Rhizobium*; alginate in *Pseudomonas*) details of the mechanisms of assembly, polymerization, and export are known. In *Rhizobium*, over 20 genes are known to be involved in EPS synthesis.

### Structure

EPS are primarily composed of carbohydrates, but, in addition to sugars, there may be organic and inorganic components. Many or most are polyanionic in nature. Organic substituents include various organic acids (acetate, pyruvate, etc.) and amino acids. Inorganic substituents include sulfate, phosphate, and also cations (due to the polyanionic nature). For example, some alginates bind $Ca^{2+}$. Hetero- and homopolysaccharides are distinguished, depending on whether the monomers in the polymer are different or all the same. EPS can also be further distinguished based on linkage types and branching.

### Functions

**EPS mediate attachment to roots** Work with associative diazotrophs (N-fixing bacteria), for example *Alcaligenes*, has shown that EPS are important in attachment of bacteria to roots. Mutant strains of *Alcaligenes faecalis* that produce either no or reduced levels of EPS, or that strongly overproduce EPS are not able to attach to rice roots as well as wild-type bacteria (the difference being several orders of magnitude of cells per grams of root).

**EPS as symbiotic signals: *Rhizobium*** EPS are known to play a role in the development of root nodule symbioses between *Rhizobium* and legumes. A mutant of *R. meliloti* that is defective in the production of the EPS succinoglycan (a heteropolysaccharide) is defective in nodule colonization, although (noninfected) pseudonodules are formed. This indicates that succinoglycan is not the initiating signal, but that it is required for successful nodule invasion. Invasion is blocked where the infection thread (which normally carries rhizobia into the nodule cells of the plant) penetrates beyond the epidermal cell layer.

**Desiccation resistance** Bacteria in soil generally exist within a matrix mainly composed of EPS. This EPS matrix may protect bacteria from drying and from fluctuations in water potential. EPS production may hence serve to alter their microenvironment to enhance survival of desiccation in soils that experience drought conditions (complementing other mechanisms, such as synthesis of compatible solutes or uptake of inorganic solutes such as $K^+$). Slowing the rate of drying within the microbe's microenvironment may increase the time for adjustment of metabolism. Additionally, increased water retention and diffusion rates in EPS may help soil bacteria to maintain physiological functions at low water potentials. At various soil water potentials, glucose diffuses faster in EPS or in EPS-amended clay than in pure clay, due to the higher volumetric water content of EPS and their water-saturated porosity.

In experiments using a sand matrix, EPS produced by a soil *Pseudomonas* species has held several times its weight in water at low water potentials (e.g., between 3 and 10 g water per gram of solid EPS at −3 MPa), suggesting that EPS may serve as a water reservoir for cells in the vicinity. EPS-amended sand also holds more water at all water potentials than unamended sand, and the presence of EPS slows the drying rate of the sand. Cultures of the bacteria exposed to desiccation while growing in sand contain more EPS, suggesting that resources are allocated to EPS production during drought. Also, the structure of the bacterial EPS is altered in response to desiccation (as shown by cryoscanning electron microscopy), such that the EPS structures appear less porous and more massive under desiccation.

**Soil aggregation** Numerous products and activities of microbes in soil contribute to the complex process of soil aggregation (*See* **Aggregation:** Microbial Aspects). Organic binding agents that are involved in stabilizing aggregates can be categorized into transient, temporary, and persistent binding agents based on their persistence in soil. Transient binding agents are decomposed rapidly (and produced rapidly) by microorganisms in soil; the most important of these are polysaccharides, including EPS of microbial origin. Evidence for the involvement of polysaccharides in soil aggregation comes from many sources, spanning the spectrum of soil ecological approaches. Correlative studies have related the concentration of certain fractions of extracted polysaccharide (e.g., NaOH or hot water extraction) with aggregate stability; however, as is the case with many other soil extractions, large amounts of other components of soil organic matter may be coextracted (for example humic acid). Some studies have used the ratio of the monosaccharides mannose to xylose (M/X) to assess the relative origin of carbohydrates in soil (after hydrolysis). The proportion of carbohydrates of microbial origin is assumed to increase with increasing M/X ratios. Other approaches include selective oxidation of polysaccharides (using periodate), electron-microscopical observations of aggregates (showing attachment of clay particles to EPS material), and controlled environment studies with defined microbial EPS and clays.

*Alginate: a model system for EPS biology* The acidic EPS alginate has been shown to be involved in biofilm formation of the Gram bacterium *Pseudomonas aeruginosa*. Alginate biosynthesis has become a model system in the study of EPS production (perhaps not least due to the biomedical context of cystic fibrosis research). Alginate is composed of β-1,4-linked D-mannuronic acid and α-L-guluronic acid. The monomers can be arranged in a homopolymeric or heteropolymeric fashion. In addition bacterial alginates are frequently O-acetylated to different degrees. The polymeric structure and O-acetylation determine the physicochemical properties of this EPS; for example, increased O-acetylation increases the water-binding capacity of alginate. Bacterial alginate has a high molecular mass and is negatively charged (as a consequence it is hydrated and viscous). Alginate is produced in the environment, in human infections, and industrial systems. It also functions as a virulence factor of plant pathogenic *Pseudomonas* spp., and is involved in the encystment process of *Azotobacter* spp.

A considerable amount of detail is available for the biosynthesis pathway of alginate. More than 20 genes are involved in alginate biosynthesis in *P. aeruginosa* (all located on the bacterial chromosome). Biosynthesis starts with fructose-6-phosphate and proceeds through the phases of activation of precursor, oxidation of precursor (to guanosine diphosphate (GDP)-mannuronate), polymerization, modification (epimerization, O-acetylation), and export. The initial steps in the pathway are common in general carbohydrate metabolism; a critical shunt and kinetic control point in alginate synthesis is GDP-mannose dehydrogenase (*algD*), the *algD* gene being tightly controlled. The export mechanism of alginate is not clearly established yet, but modification of alginate takes place in the periplasmic space: O-acetylation and epimerization. It is not clear how these reactions are controlled, but their rates are important in determining the properties of alginate.

Alginate biosynthesis is triggered by a variety of environmental factors. These include high osmolarity, nutrient limitation, presence of toxic compounds (e.g., copper), and attachment of bacteria to a solid surface. High osmolarity and ethanol (dehydrating agent) enhance transcription of alginate biosynthesis genes (and to a limited extent, alginate production) in *P. aeruginosa* and *P. syringae*. AlgU, a regulatory element (sigma factor) involved in the regulation of biosynthesis of the EPS alginate in *P. fluorescens*, is essential in providing desiccation tolerance *in vitro* and in soil. ATR/FT-IR and CLSM were used to study the role of alginate and alginate structure in biofilm formation by *P. aeruginosa* on surfaces. Alginate appears to be required for formation of thicker, three-dimensional biofilms.

## Extracellular Lipid Compounds

Lipids are typically discussed in the context of membrane composition (e.g., phospholipids) or intracellular storage compounds; however, extracellular lipids are produced by both prokaryotes and eukaryotes. Simple extracellular lipids (e.g., free fatty acids, and various oxidation products of microbes utilizing alkanes) are not discussed here; the focus is on complex extracellular lipids. These contain long-chain acyl moieties attached to a nonlipid component; the latter is frequently a carbohydrate, a protein, or a peptide.

The combination of lipid and nonlipid components in the molecule confers hydrophilic and hydrophobic properties, and hence some of these compounds can be active as surfactants (also referred to as biosurfactants). The emulsifying ability of some complex extracellular lipids explains probably most of their biological role; however, some of these lipid compounds also alter the hydrophobicity of the microbial cell, which also influences (positively or negatively) uptake of, for example, hydrocarbons. Microbially produced surfactants offer the advantage of being potentially less toxic and more biodegradable than some synthetic surfactants, explaining the heightened interest in this compound in biotechnology. The efficiency of biosurfactants in soil remediation has been studied in the case of soil contamination by metals (e.g., $Cd^{2+}$, $Zn^{2+}$, $Pb^{2+}$), polychlorinated biphenyls (PCBs), phenantrene, and polyaromatic hydrocarbons (PAHs).

Glycolipids are lipids with a sugar moiety; the most common groups are sophorolipids (glycolipids containing the disaccharide sophorose; e.g., in *Candida* yeasts), trehaloselipids (containing the disaccharide trehalose; e.g., in *Arthrobacter*, *Rhodococcus*), and rhamnolipids (containing the sugar rhamnose). Rhamnolipids have been extensively studied with respect to their surfactant ability (mostly in the context of bioremediation of metal or hydrocarbon contaminated soils and subsoils), for example in strains of the soil bacterium *P. aeruginosa*.

In addition to having surfactant function (allowing the bacteria to grow on hydrocarbons), rhamnolipids have some bactericidal (against Gram-negative) and antiviral activity. Rhamnolipid from *P. aeruginosa* has zoosporicidal activity against phytophathogenic fungal species such as *Pythium*, *Phytophthora*, and *Plasmopara*. Proposed mechanisms of action include disruption of the fungal plasma membrane by the rhamnolipid. This points to a potential use of glycolipids in biological control of agricultural pests.

*See also:* **Aggregation:** Microbial Aspects; **Enzymes in Soils**; **Mycorrhizal Fungi**; **Organic Matter:** Principles and Processes

## Further Reading

Boulton CA (1989) Extracellular microbial lipids. In: Ratledge C and Wilkinson SG (eds) *Microbial Lipids*, vol. 2, pp. 669–694. London: Academic Press.

Chenu C and Roberson EB (1996) Diffusion of glucose in microbial extracellular polysaccharide as affected by water potential. *Soil Biology and Biochemistry* 28: 877–884.

Foster RC, Rovira AD, and Cock TW (1983) *Ultrastructure of the Root–Soil Interface*. St Paul, MN: The American Phytopathological Society.

Gacesa P (1998) Bacterial alginate biosynthesis – recent progress and future prospects. *Microbiology* 144: 1133–1143.

Leigh JA and Walker GC (1994) Exopolysaccharides of *Rhizobium*: synthesis, regulation and symbiotic function. *Trends in Genetics* 10: 63–67.

Roberson EB and Firestone MK (1992) Relationship between desiccation and exopolysaccharide production in a soil *Pseudomonas* sp. *Applied and Environmental Microbiology* 58: 1284–1291.

Tisdall JM and Oades JM (1982) Organic matter and water-stable aggregates in soils. *Journal of Soil Science* 33: 141–163.

Wingender J, Neu TR, and Flemming HC (1999) *Microbial Extracellular Polymeric Substances*. Berlin: Springer-Verlag.

Wösten HAB (2001) Hydrophobins: multipurpose proteins. *Annual Reviews of Microbiology* 55: 625–646.

Wright SF and Upadhyaya A (1998) A survey of soil for aggregate stability and glomalin, a glycoprotein produced by hyphae of arbuscular mycorrhizal fungi. *Plant and Soil* 198: 97–107.

## Poorly Crystalline Alluminosilicates *See* Amorphous Materials

# POROSITY AND PORE-SIZE DISTRIBUTION

**J R Nimmo**, US Geological Survey, Menlo Park, CA, USA

Published by Elsevier Ltd.

A soil's porosity and pore-size distribution characterize its pore space, that portion of the soil's volume that is not occupied by solid material. The basic character of the pore space governs critical aspects of almost everything that occurs in the soil: the movement of water, air, and other fluids; the transport and the reaction of chemicals; and the residence of roots and other biota. By convention the definition of pore space excludes fluid pockets that are totally enclosed within solid material – vesicles or vugs, for example, that have no exchange with the pore space that has continuity to the boundaries of the medium. Thus we consider a single, contiguous pore space within the body of soil. In general it has fluid pathways that are tortuous, variably constricted, and usually highly connected among themselves. **Figure 1** shows an example of a two-dimensional cross section of soil pore space.

The pore space is often considered in terms of individual pores – an artificial concept that makes possible many quantifications of its essential character. Though many alternatives could serve as a basis for the definition of pores and their sizes, in soil science and hydrology these are best conceptualized, measured, and applied with respect to the fluids that occupy and move within the pore space.

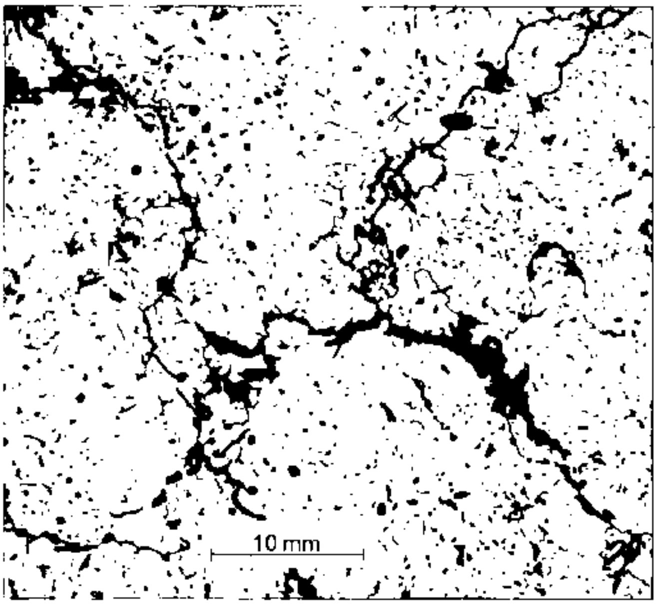


**Figure 1** Cross section of a typical soil with pore space in black. Study of this figure would lead to an underestimate of porosity, because pores smaller than about 0.1 mm do not appear. (Adapted from Lafeber D (1965) The graphical representation of planar pore patterns in soils. *Australian Journal of Soil Research* 3: 143–164, with permission.)

## Porosity

Porosity, $\phi$, is the fraction of the total soil volume that is not occupied by solid material. Thus it is a single-number quantification of the amount of space available to fluid within a specific body of soil. With the assumption that soil is a continuum, porosity can be considered as a function of position. Being simply a fraction of total volume, $\phi$ can range from 0 to 1, typically falling between 0.3 and 0.7 for soils.

### Porosity in Natural Soils

The porosity of a soil depends on several factors, including: (1) packing (bulk) density, (2) the breadth of the particle-size distribution (polydisperse versus monodisperse), (3) the shape of particles, and (4) cementing materials. Mathematically considering an idealized soil of packed uniform spheres, $\phi$ must fall between 0.26 and 0.48, depending on the packing. Spheres randomly thrown together will have $\phi$ near the middle of this range, typically 0.30–0.35. A sand with grains nearly uniform in size (monodisperse) packs to about the same porosity as spheres. In a polydisperse sand, the fitting of small grains within the pores between large ones can reduce $\phi$, conceivably below the 0.26 uniform-sphere minimum. **Figure 2** illustrates this concept. The particular sort of arrangement required to reduce $\phi$ to 0.26 or less is highly improbable, however, so $\phi$ also typically falls within 0.30–0.35 for polydisperse sands. Particles more irregular in shape tend to have larger gaps between their nontouching surfaces, thus forming media of greater porosity. In porous rock such as sandstone, cementation or welding of particles not only creates pores that are different in shape from those of particulate media, but also reduces the porosity as solid material takes up space that would otherwise be pore space. Porosity in such a case can easily be less than 0.3, even approaching 0. Cementing material can also have the opposite effect. In many soils, clay and organic substances cement particles together into aggregates. An individual aggregate might have a 0.35 porosity within it, but the medium as a whole has additional pore space in the form of gaps between aggregates, so that $\phi$ can be 0.5 or greater. Observed porosities can be as great as 0.8–0.9 in a peat soil (extremely high in organic matter).

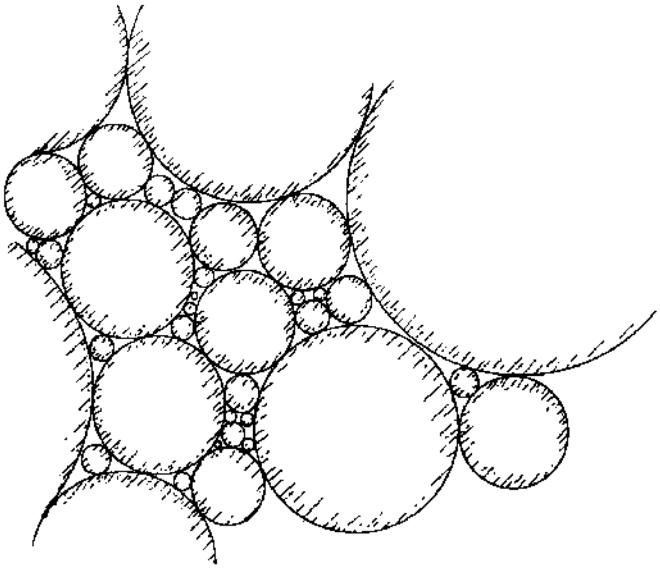


**Figure 2** Dense packing of polydisperse spheres. (Adapted from Hillel D (1980) *Fundamentals of Soil Physics*. San Diego: Academic Press, with permission.)

Porosity is often considered to partition into two components, most commonly called textural and structural porosity. The textural component is the value the porosity would be deduced to have if the arrangement of the particles were random, as described above for granular material without cementing. That is, the textural porosity might be approximately 0.3 in a granular medium. (The low end of the 0.30–0.35 range is preferred, because the high end probably reflects nonrandom structural influences in real sands.) The structural component, often considered to represent macropores, is arithmetically the difference between the textural porosity and the total porosity.

The texture of the medium relates in a general way to the pore-size distribution, as large particles give rise to large pores between them, and therefore is a major influence on the retention curve. Additionally, the structure of the medium, especially the degree to which it has aggregation, shrinkage cracks, wormholes, etc., substantially influences the retention curve.

### Measurement of Porosity

The technology of thin sections or of tomographic imaging can produce a visualization of pore space and solid material in a cross-sectional plane, as in **Figure 1**. The summed area of pore space divided by total area gives the areal porosity over that plane. An analogous procedure can be followed along a line through the sample, to yield a linear porosity. If the medium is isotropic, either of these would numerically equal the volumetric porosity as defined above, which is more usually of interest.

The volume of water contained in a saturated sample of known volume can indicate porosity. The mass of saturated material less the oven-dry mass of the solids, divided by the density of water, gives the volume of water. This divided by the original sample volume gives porosity.

An analogous method is to determine the volume of gas in the pore space of a completely dry sample. Of course not only the sampling but also the drying of the soil must be done so as not to compress the soil or otherwise alter its porosity. A pycnometer can measure the air volume in the pore space. A gas-tight chamber encloses the sample so that the internal gas-occupied volume can be perturbed by a known amount while the gas pressure is measured. This is typically done with a small piston attached by a tube connection. The Boyle law indicates the total gas volume from the change in pressure resulting from the volume change. This total gas volume minus the volume within the piston, connectors, gaps at the chamber walls, and any other space not occupied by soil gives the total pore volume to be divided by the sample volume.

To avoid the need to saturate with water or air, measurements of particle density, $\rho_p$, and bulk density, $\rho_b$, are used. From the definitions of $\rho_b$ as the solid mass per volume of soil and $\rho_p$ as the solid mass per solid volume, their ratio $\rho_b/\rho_p$ is the complement of $\phi$, so that:

$$\phi = 1 - \rho_b/\rho_p \qquad [1]$$

Often the critical source of error is in the determination of soil bulk volume, measured on an uncompacted sample or from the volume of the excavation from which the soil sample originated.

## Pores and Pore-Size Distribution

### The Nature of a Pore

That soil has pores is obvious, but the precise delineation of an individual pore readily leads to ambiguity and contradiction. Pore space can be partitioned into individual pores only by an arbitrary criterion, which often is not explicitly stated when pores or their sizes are discussed. Because of this inherent arbitrariness, some scientists argue that the concepts of pore and pore size should be avoided. Much valuable theory of the behavior of the soil–water–air system, however, has been built on these concepts, defined using widely, if not universally, accepted criteria.

A particularly useful criterion considers the pore space as a collection of channels through which fluid can flow. The effective width of such a channel varies along its length. Wide portions can be considered individual pore bodies, and narrow portions can be considered the pore openings that separate them. These are often described with anatomical metaphors, the wide part being the 'belly' or 'waist', and the constrictive part being the 'neck' or 'throat.' In a medium dominated by textural pore space such as sand, pore bodies are the intergranular spaces of dimensions typically slightly less than those of the adjacent particles. At another extreme, a wormhole, if it is essentially uniform in diameter along its length, might be considered a single pore. The boundaries of such a pore are of three types: (1) interface with solid, (2) constriction – a plane through the locally narrowest portion of pore space, or (3) interface with another pore (e.g., a crack or wormhole) or a hydraulically distinct region of space (e.g., the land surface).

Pores conceptualized by this cellular approach are especially relevant to hydraulic behavior, as has been recognized for more than 70 years. The initial application was to Haines jumps, shown in **Figure 3**, still considered the basic phenomena of capillary hysteresis. The pore openings, which control the matric pressure $P$ at which pores empty, are smaller than the pore bodies, which control the $P$ at which pores fill. As the medium dries and $P$ decreases, water retreats gradually as the air–water interface becomes more curved. At the narrowest part of the pore opening, this interface can no longer increase curvature by gradual amounts, so it retreats suddenly to narrower channels elsewhere. An analogous phenomenon occurs during wetting, when the decreasing interface curvature cannot be supported by the diameter of the pore at its maximum width. The volume that empties and fills in this way is essentially an individual pore. Not all pore space is subject to Haines jumps – water remains in crevices and in films (not seen in **Figure 3**) coating solid surfaces. Various models and theories treat this space in different ways. By the definition above it is part of a pore in addition to the volume affected by the Haines jump.

Pores can be classified according to their origin, function, or other attributes. A textural/structural

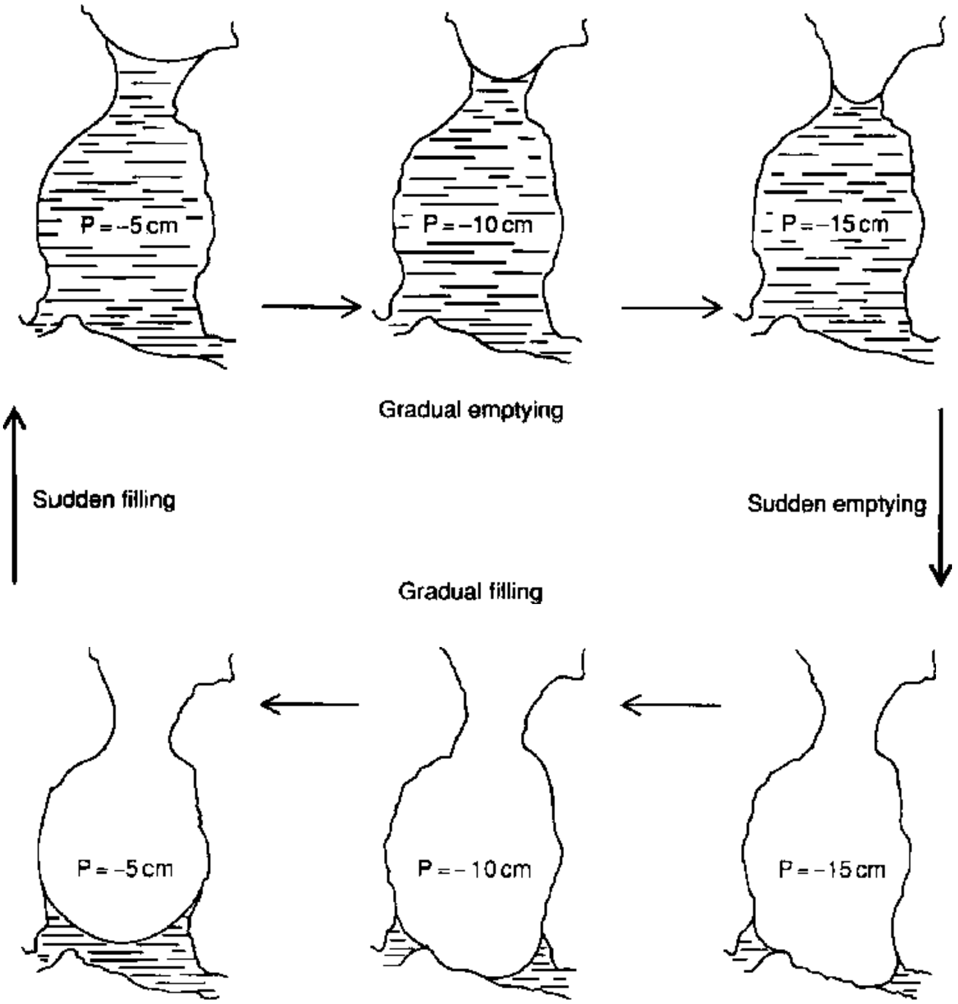


**Figure 3** Dynamics of a Haines jump. (Adapted from Miller EE and Miller RD (1956) Physical theory for capillary flow phenomena. *Journal of Applied Physics* 27: 324–332, with permission.)

distinction is possible, analogous to porosity. Hierarchical classifications are also possible. Intragranular or dead-end pores (if not entirely enclosed within solid) might empty or fill with water, but without contributing directly to fluid movement through the medium. Intergranular pores are the major portion of textural porosity, as discussed above. Intraaggregate pores may be essentially equivalent to intergranular pores within an aggregate. Interaggregate pores, including shrink–swell cracks, are common types of macropores. Biogenic pores, for example the channels left by decayed roots and the tunnels made by burrowing animals, are another common type of macropore in soils.

Specific pore sizes are usually specified by an effective radius of the pore body or neck. This relates to the radii of curvature of the air–water interface at which Haines jumps occur, and by capillarity, to the matric pressures at which these occur. Alternative indicators of size include the cross-sectional area or the volume of a pore, and the hydraulic radius, the ratio of the cross-sectional area to circumference, or of pore volume to specific surface.

The pore-size distribution is the relative abundance of each pore size in a representative volume of soil This can be symbolized with a function $f(r)$, proportional to the combined volume of pores of effective radius $r$. Figure 4 shows examples. As with porosity, the continuum assumption makes valid the concept that $f(r)$ varies from point to point within a body of soil. Also like porosity, $f(r)$ may be taken to comprise textural and structural components.

## Measurement

The most obvious and straightforward measurements of pore size are with geometric analysis of images of individual pores. This can be done using various types of microscopy on thin sections or other flat soil surfaces, or tomographs. Dimensions of pore bodies and necks can be measured manually or by computer analysis of digitized images. The lengths of segments of solid or pore along one-dimensional transects can serve similar purposes. As in the case of porosity, isotropy is required for assuming the equality of lineal, areal, and volumetric pore size distribution. For pore size, however, a more important problem is that, when working with fewer than three dimensions, one doesn't know what *part* of pore the selected slice intersects; because it does not in general go through the widest part, it underestimates the pore diameter. Correction techniques are necessary to estimate true pore body and opening sizes.

Three-dimensional analysis is possible with impregnation techniques. In these, the soil pore space is filled with a resin or other liquid that solidifies. After it has done so, the medium is broken up and individual blobs of solid resin, actually casts of the pores, are analyzed as particles.

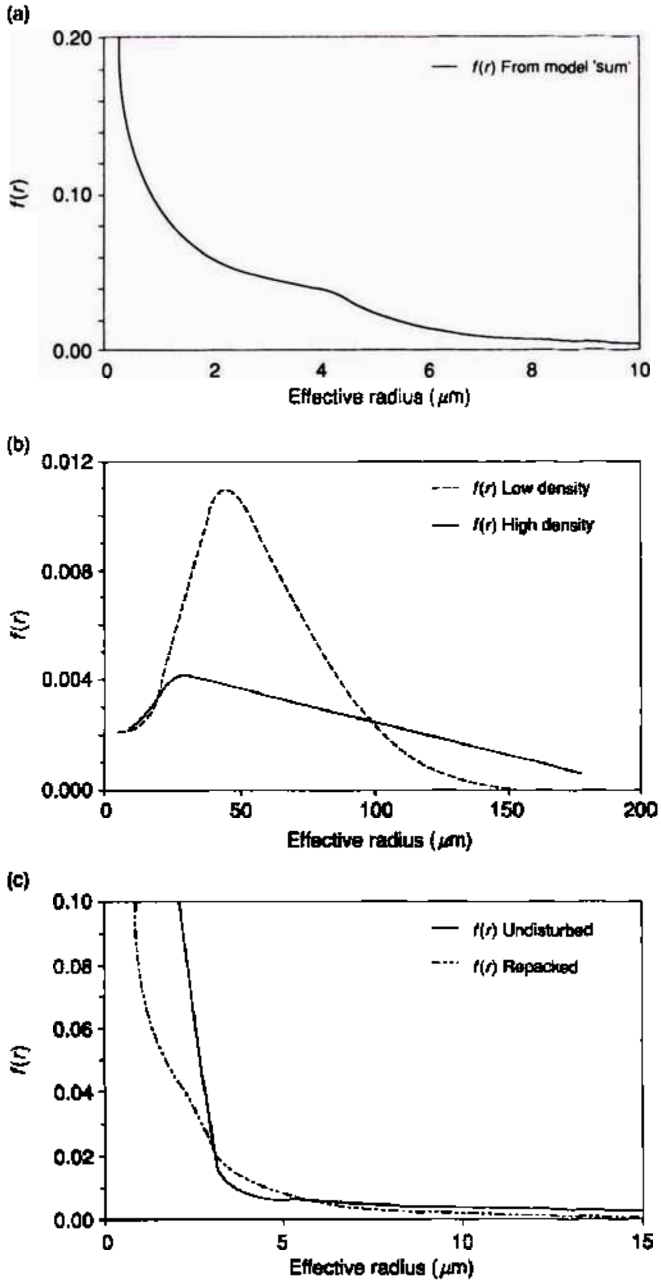


**Figure 4** Pore-size distributions based on measured water retention. (a) Loamy soil. (Reproduced with permission from Schofield (1935).) (b) Silty sand at two packing densities. (Data from Croney D and Coleman JD (1954) Soil structure in relation to soil suction (pF). *Journal of Soil Science* 5: 75–84.) (c) Paleosol of sandy loam texture from 42 m depth, as a minimally disturbed core sample, and after disaggregation and repacking to the original density. (Data from Perkins KS (2003) Measurement of sedimentary interbed hydraulic properties and their hydrologic influence near the Idaho Nuclear Technology and Engineering Center at the Idaho National Engineering and Environmental Laboratory. *Water-Resources Investigations Report* 03-4048. Idaho Falls, Idaho: US Geological Survey; and Schofield RK (2002) The pF of the water in soil. *Transactions of the Third International Congress of Soil Science*, pp. 38–48. London: T Murby.)

Image-based techniques can be prohibitively tedious because enough pores must be analyzed to give an adequate statistical representation. They can give a wealth of information, however, on related aspects such as pore shape and connectivity.

More common than imaging methods are those based on effective capillary size. These use data from fluid behavior in an unsaturated medium, usually the emptying or filling of pores during soil drying or wetting (**Figure** 3), that is, the water-retention curve $\theta(P)$, where $\theta$ is the volumetric water content. Because large pores fill or empty at $P$ near 0, a medium that has many large pores will have a retention curve that drops rapidly to low $\theta$ at high matric potentials. Conversely, one with very fine pores will retain much water even at highly negative $P$, thus having a flatter retention curve. By the capillary hypothesis, the $P$ at which a pore empties (or fills) corresponds to the pore opening (or body) size according to:

$$r = \frac{-2\sigma\cos\alpha}{P} \qquad [2]$$

where $\sigma$ is the surface tension and $\alpha$ is the contact angle. This formula can convert a measured $\theta(P)$ into an equivalent $\theta(r)$ curve. This curve is actually a cumulative pore-size distribution; the water content on a drying $\theta(r)$ indicates the combined volume of all pores with opening radius less than $r$. Applying the fundamental theorem of calculus, the direct pore-size distribution is simply the derivative:

$$f(r) = \frac{d\theta}{dr} \qquad [3]$$

Mercury porosimetry is analogous to the water retention-based method, but uses air as the wetting fluid, corresponding to the water, and mercury as the nonwetting fluid, corresponding to the air in the water–air system. Mercury is forced into the pores of dry soil incrementally, so that the relation between mercury content and mercury pressure is obtained. Applying eqn [2] with the appropriate values of surface tension and contact angle for mercury leads to a pore-size distribution estimate as for the extraction of water from a water–air system.

The different measurement techniques do not give exactly the same results. Emptying and filling depend on more than capillarity – different methods are affected differently by unintended influences. Contact angles, for example, are dynamic, are not likely to be at handbook values in soil, and may deviate quite differently for water and for mercury. The swelling of clays can be a major influence on a water-retention curve, but should have no effect with mercury.

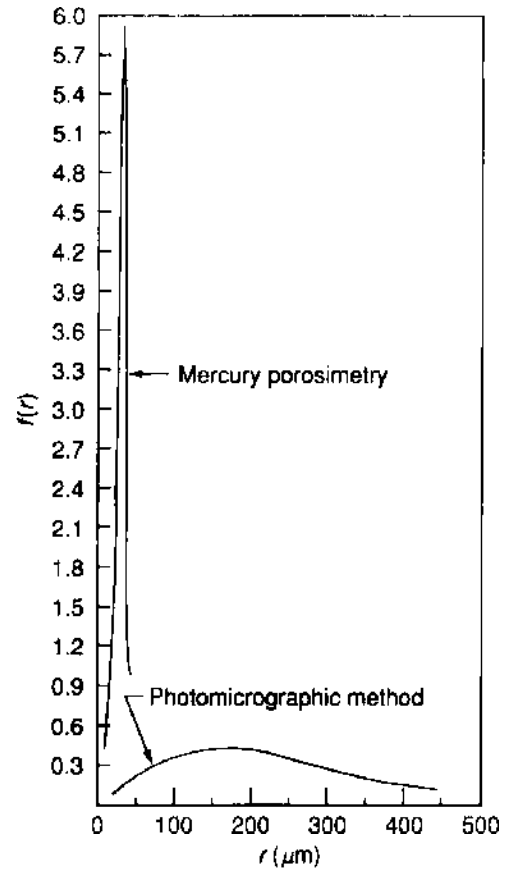


**Figure 5** Pore-size distributions obtained for an artificial medium using an imaging method and mercury intrusion. (Reproduced with permission from Dullien FAL and Batra VK (1970) Determination of the structure of porous media. *Industrial Engineering Chemistry* 62: 25–53.)

Imaging techniques are subject to some entirely different influences, such as the chance and subjectivity involved in assessing pore bodies and openings. **Figure** 5 shows an example of substantially different results from different methods. In soil environmental applications, use of the water-retention curve is most common. Data are more commonly available for this than for any other method. One can also expect a water-based method to give better results for a water-based application (e.g., hydraulic conductivity).

### Representation

The basic representation of pore-size distribution is as the function $f(r)$ (**Figure** 4). Corresponding graphs of the cumulative size distribution can be used equivalently. Often a specific functional form or other representational model of pore-size distribution is useful. A normal or lognormal can be fitted to the data, for example. Structural features may give the soil a bimodal pore-size distribution, leading to several distinctive effects on water flow. Bimodal, trimodal, or other multimodal forms are possible, and can be represented by superpositions of the normal or lognormal forms. Self-similar or fractal forms of the distribution function are also used, giving a power-law form of the cumulative pore-size distribution.

### Typical Features of a Pore-Size Distribution

With the equivalent-capillary concept taken as broadly as possible, the smallest possible pore is limited to the order of a few molecular diameters, approximately 1 nm. Fluid behavior is then likely to be dominated by interaction with the solid material and is not necessarily describable by capillary laws or even standard thermodynamics. The smallest pores measured by water retention or mercury intrusion are typically 50 or 100 times larger than this, though with greater cost and effort, greater magnitudes of pressure and hence smaller measured $r$ can be achieved. As the imputed pore size approaches zero, in effect thin films are associated with effective radii of filled pores. Because $r$ relates inversely to $P$, the whole dry portion of the retention curve is contained in a very small region near $r = 0$, where the apparent number of pores becomes large. In the extreme case, when $\theta$ is held to remain artificially finite at an assumed residual water content as $P$ goes to negative infinity, $f(r)$ at $r = 0$ is a delta function.

There is no real upper limit to pore size, though instability will cause the capillary hypothesis to break down for $r$ of a few millimeters. The experimental limit is typically approximately 0.5 mm.

At either extreme, where capillary phenomena lose dominance, eqn [2] no longer applies, though it still can give $r$-values corresponding to $P$. These may be useful for translating one property to another (discussed below) even though they are invalid in terms of the capillary analog.

At intermediate values of $r$, sometimes $f(r)$ has a pronounced peak, as in Figures 4b and 5. Such a peak is assumed to exist in many applications and interpretations of pore-size distributions, for example normal or lognormal representations, though it does not always exist in the measurable range of $r$. A peak in $f(r)$ corresponds to an inflection point in $\theta(r)$. The inverse relation of $r$ to $P$ means that $\theta(P)$ is much more likely to have an inflection point than $\theta(r)$; if $\theta(r)$ has an inflection point, then $\theta(P)$ is mathematically required to have an inflection point, but the converse is not true. The common case of an S-shaped retention curve (with inflection point in the measured range, and normally a distinct air-entry effect) will have a defined peak in the pore-size distribution only if the middle portion of the retention curve has a slope that differs markedly from that of the two end segments. This is frequently true for monodisperse media and for repacked samples, in which pores larger than a certain $r$ have been destroyed. For many soils, perhaps most, $f(r)$ does not have a peak except at $r = 0$ or so close to 0 that it cannot be measured. This is especially likely for media that are macroporous, polydisperse, or that show an air-entry effect at essentially $P = 0$. Retention curves represented by a fractal or power-law model are also in this category. Lognormal models or other representations that have a defined peak can often still be applied but with the peak outside the measurable range. Some models (e.g., of hysteresis and hydraulic conductivity) may be unaffected by the lack of a measurable peak even though they were derived assuming the pore-size distribution resembles some form of normal distribution.

## Significance to Soil and Water Behavior

### Significance of Porosity

One obvious significance of porosity is that it is an upper limit of volumetric water content. It is similarly essential to the definition of degree of saturation, $\theta$ divided by porosity. Another significance is that, within the pore space, the complement to $\theta$ is gas content. That is, volumetric gas content equals the difference between $\theta$ and porosity.

The magnitude of porosity roughly indicates complexity of structure, being greater for greater complexity, as noted above for aggregated soil. The spatial variability of porosity is also important, greater variability correlating with greater heterogeneity of the soil. Significant spatial variability on a small scale also implies greater structural complexity, independent of the magnitude of porosity.

### Significance of Pore-Size Distribution

A major importance of a soil's pore-size distribution is that it relates to other soil properties in a complex and useful way, as discussed below. It indicates complexity of structure in far more detail than porosity alone. The spatial variation of pore size is an important characteristic of the medium. The pore-size distribution of different parts of soil is the fundamental basis for the concept of aggregates, for example. By some definitions, pore size can permit essential distinctions between micropores and macropores (and mesopores, where that term is used for intermediate-size pores).

The relation of pore-size to particle-size distribution in a randomly structured medium is likely to be monotone: bigger pores are associated with bigger particles. With the nonrandom structure of most soils, the relation to particle size is complex. Bigger pores can be associated with bigger particles, but also with smaller particles such as clays that promote aggregation and hence the existence of large interaggregate pores.

With more intricate conceptualizations, subdivisions of a pore-size distribution can represent more detailed properties of soils, for example the hysteresis of soil water retention. **Figure 6** shows an example in which one graph represents pore-opening and pore-body size distributions. The added information of what fraction of pores with a given opening size have a particular range of body sizes is taken to represent the essential quantitative basis of soil water hysteresis.

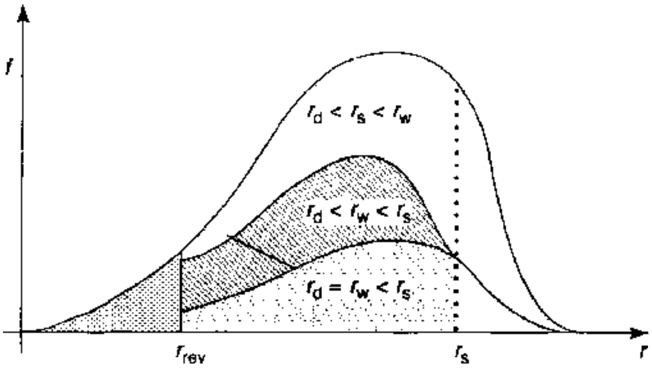


**Figure 6** Hypothetical pore-size distribution with subregions distinguished on the basis of pore body and opening radii and drying–wetting history. (Adapted from Nimmo JR (1992) Semiempirical model of soil water hysteresis. *Soil Science Society of America Journal* 56: 1723–1730.)

### Pore Dynamics

Natural and artificial soil processes create and destroy pores, and induce changes in their size and other attributes. **Table 1** lists some of the common effects on pore-size distribution caused by routine processes. In general, if the soil is compacted by uniform stress, such as a weight imposed at the land surface, it normally loses large pores and gains small pores. **Figure 4b** shows results of this type. Disturbance from irregular stresses, as during digging or repacking, has a variety of effects on pore size, often with the net effect of a decrease in the number of large pores and an increase in the number of small pores, as shown in **Figure 4c**. Several types of process can create pores. Though small intergranular pores are seldom completely closed, some processes can close macropores, in effect destroying them.

### Applications to Soil Transport Properties

Because the pores it characterizes are conduits and repositories for water, the pore-size distribution relates directly to soil hydraulic properties. In addition to water retention, as described above, because the characterized pores are also fluid conduits, it connects with predictions of hydraulic conductivity

**Table 1** Possible effects of routine soil processes on pore-size distribution

| *Soil process* | *Effect* |
|---|---|
| Shrinkage | Enlarged macropores<br>New macropores created<br>Within an aggregate, intraaggregate pores decrease in size, or increase in size if clay particles shrink |
| Swelling | Decrease in the size of macropores<br>Closed macropores<br>Within an aggregate, intraaggregate pores increase in size, or decrease in size if clay particles expand |
| Mechanical compression | Decrease in the size of macropores<br>Closed macropores<br>Aggregates break up, reducing the number of intraaggregate pores, thereby reducing the fraction of the pore space represented by the smallest pores |
| Disturbance from digging or plowing | Existing macropores destroyed<br>Interclod macropores created<br>Aggregates break up, reducing the number of intraaggregate pores, also the fraction of the pore space represented by the smallest pores |
| Biological activity | New macropores created<br>Enlarged macropores, as by ongoing traffic of ants or burrowing mammals<br>Decrease in the size of macropores, e.g., if they are affected by compression resulting from the expansion of a nearby root<br>Increased aggregation, promoting the creation of interaggregate macropores, and possibly smaller intergranular pores within aggregates<br>Constricted or obstructed pores, e.g., through growth of microorganisms |
| Chemical activity | Constricted or obstructed pores through formation of precipitates<br>Enlarged pores through dissolution of precipitates<br>Increased or decreased interparticle cohesion, with complex effects on pore size and structure |

$K$. The same principles can be applied to the transport of fluids other than water, including gases, though water flow is the most common application.

By analogy to laminar flow in tubes as quantified by the Poiseuille law, the conductance of a single pore can be inferred to be proportional to the fourth power of its effective radius. This makes its hydraulic conductivity proportional to the square of its effective radius. An estimated $f(r)$ distribution tells the relative abundance of each conduit size, thus providing the ingredients needed to predict $K$.

A $K$ prediction based on the capillary hypothesis assumes pores are filled if smaller than a certain radius determined by $\theta$. The portion of $f(r)$ representing the filled pores is relevant to unsaturated $K$. The simplest possibility is to integrate $f(r)$ weighted by $r^2$ over the domain represented by filled pores, giving a number proportional to $K$ for the corresponding $\theta$. Multiplying this integration by a separately determined matching factor gives the actual predicted $K$.

Many specific models in the published literature are based on these ideas. They differ in how they treat matters such as pore length, connectedness, tortuosity, and the distinction between pore opening and body dimensions. Popular models include those of Mualem and Burdine, which have been analytically combined with widely used empirical formulas for retention curves.

Considering solute transport, the pore-size distribution affects advection as it does hydraulic conductivity. Additionally it is important to dispersion, expected to be greater for a broader pore-size distribution. Adsorptive reactions with other soil constituents are related in a complex way. The smaller pores are associated with longer residence times and greater relative surface area, but most solutes may go quickly through the large pores with minimal opportunity to react. Interchange between fast-transporting and slow-transporting portions of the pore space is a vital and much-investigated aspect of solute transport in soils. Sometimes the terms 'mobile' and 'immobile' are used in this context, but of course the distinction is not as sharp as these terms imply.

For particle transport, many aspects are essentially the same as for $K$ and solute transport. Additionally, the phenomenon of straining depends critically on the proportion of pores smaller than a given particle size. This is the dominant factor in some particle-transport applications.

Pore-size distribution relates indirectly to heat transport. It correlates in some way with the degree of aggregation, the moisture relations, the number of points of particle contact, and other factors affecting thermal conductivity.

## Summary

The characterization of pore space is a vital and fruitful aspect of soil investigation. Liquid, solid, and gas constituents of the soil govern the form and development of pores, whose character in turn profoundly influences the nature and behavior of the soil.

Soil porosity is fairly well standardized in definition and measurement techniques. Pore size, however, is not obviously definable, much less measurable. Yet it is central to topics such as macropores, aggregation, fractures, soil matrix, and solute mobility. Pore size plays a key role in various proposed means of quantifying soil structure. It also has a major practical role in the prediction of hydraulic properties. New pore-size concepts, measurement techniques, and relations to transport phenomena are likely to remain a major emphasis in the study of soil.

*See also:* **Aggregation:** Physical Aspects; **Capillarity; Fractal Analysis; Hysteresis; Isotropy and Anisotropy; Macropores and Macropore Flow, Kinematic Wave Approach; Spatial Variation, Soil Properties; Structure; Swelling and Shrinking; Texture**

## Further Reading

Adamson AW and Gast AP (1997) *Physical Chemistry of Surfaces*, 6th edn. New York: John Wiley.

Baver LD (1938) Soil permeability in relation to non-capillary porosity. *Soil Science Society of America Proceedings* 3: 52–56.

Burdine NT, Gournay LS, and Reichertz PP (1950) Pore size distribution of petroleum reservoir rocks. *Petroleum Transactions – AIME* 189: 195–204.

Childs EC and Collis-George N (1950) The permeability of porous materials. *Proceedings of the Royal Society of London*. In: *Series A: Mathematical and Physical Sciences*, pp. 392–405.

Dullien FAL and Batra VK (1970) Determination of the structure of porous media. *Industrial Engineering Chemistry* 62(10): 25–53.

Haines WB (1930) Studies in the physical properties of soil. V. The hysteresis effect in capillary properties, and the modes of moisture distribution associated therewith. *Journal of Agricultural Science* 20: 97–116.

Kosugi K (1994) Three-parameter lognormal distribution model for soil water retention. *Water Resources Research* 30(4): 891–901.

Marshall TJ and Holmes JW (1988) *Soil Physics*, 2nd edn. Cambridge, UK: Cambridge University Press.

Miller EE and Miller RD (1956) Physical theory for capillary flow phenomena. *Journal of Applied Physics* 27(4): 324–332.

Mualem Y (1974) A conceptual model of hysteresis. *Water Resources Research* 10(3): 514–520.

Mualem Y (1976) A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research* 12(3): 513–522.

Nimmo JR (1997) Modeling structural influences on soil water retention. *Soil Science Society of America Journal* 61(3): 712–719.

Nimmo JR and Akstin KC (1988) Hydraulic conductivity of a sandy soil at low water content after compaction by various methods. *Soil Science Society of America Journal* 52(2): 303–310.

Raducu D, Vignozzi N, Pagliai M, and Petcu G (2002) Soil structure of tilled horizons influenced by management practices and implement geometry. *Advances in Geoecology* 35: 149–162.

Ragab R, Feyen J, and Hillel D (1982) Effect of the method for determining pore size distribution on prediction of the hydraulic conductivity function and of infiltration. *Soil Science* 134(6): 141–145.

# POTASSIUM IN SOILS

**P M Huang**, University of Saskatchewan, Saskatoon, Canada
**J M Zhou and J C Xie**, The Institute of Soil Science, Academia Sinica, Nanjing, People's Republic of China
**M K Wang**, National Taiwan University, Taipei, Taiwan

## Introduction

Potassium (K) is a major constituent of soils and plants. The common range of K content in the soil is 0.4–30 g K kg$^{-1}$. Of the major nutrient elements, K is generally the most abundant in the soil. Of this total K content, 98% is bound in the mineral form, whereas 2% is in soil solution and exchangeable phases.

Among the mineral cations essential for plants, K is the largest in size. Therefore, the number of oxygen ions coordinating K in mineral structure is high. The strength of each K-O is, thus, relatively weak. Potassium has a polarizability of 0.088 nm$^3$, which is higher than for $Ca^{2+}$, $Li^+$, $Mg^{2+}$, and $Na^+$, but lower than for $Ba^{2+}$, $Cs^+$, $NH_4^+$, and $Rb^+$ ions. Ions with higher polarizability are preferentially selected in ion exchange reactions.

The vitality of K in plant growth and yield formation has been known since the work of von Liebig, published in 1840. The availability of K to plants is related in many ways to the structural and surface chemistry of soil minerals. The large amount of K in soil minerals, fixation of applied K, and ion selectivity and dynamics of soil K have led researchers to study the relationships between K uptake by plants and the chemistry of soil K for decades.

## Forms of Soil Potassium

Soil K exists in four forms in soils: solution, exchangeable, fixed, and structural (**Figure 1**). Solution and exchangeable K generally account for relatively small proportions of the total K in soils. Potassium-bearing micas and feldspars are the major K reserves of the soil. Upon reaction with weathered micas, vermiculites, allophane, and zeolites, K is also present in fixed form in soils. In the immediate vicinity of the fertilizer zone, K can react with aluminum hydroxides and acid phosphate solution to form taranakite. However, the nucleation and crystallization of taranakite can be perturbed in soils with high Fe content, especially under reduced and acidic conditions. In acid soils, K could also be coprecipitated with Al and sulfate to form alunite. The order of the soil K forms in terms of bioavailability is: solution > exchangeable > fixed > structural.

### Solution K

Plants and microorganisms directly take up soil solution K. This form of K is also most subject to leaching in soils. The concentration of soil solution K varies from 2 to 5 mg K l$^{-1}$ for normal agricultural soils of humid regions and is an order of magnitude higher in arid-region soils. The amount of soil solution K is generally too low to meet the requirement of K by a crop during a growing season. Soil solution K is in dynamic equilibrium with other forms of soil K and is influenced by the soil moisture content, the concentrations of cations in the solution and on the exchanger phase, and the concentration of solution inorganic and organic ligands.

### Exchangeable K

Exchangeable K is the portion of soil K that is electrostatically bound as an outer-sphere complex to the surfaces of soil minerals and organic matter. It is readily exchanged with other cations and bioavailable. The replenishment of a K-depleted soil solution is affected by the release of exchangeable K from soils.

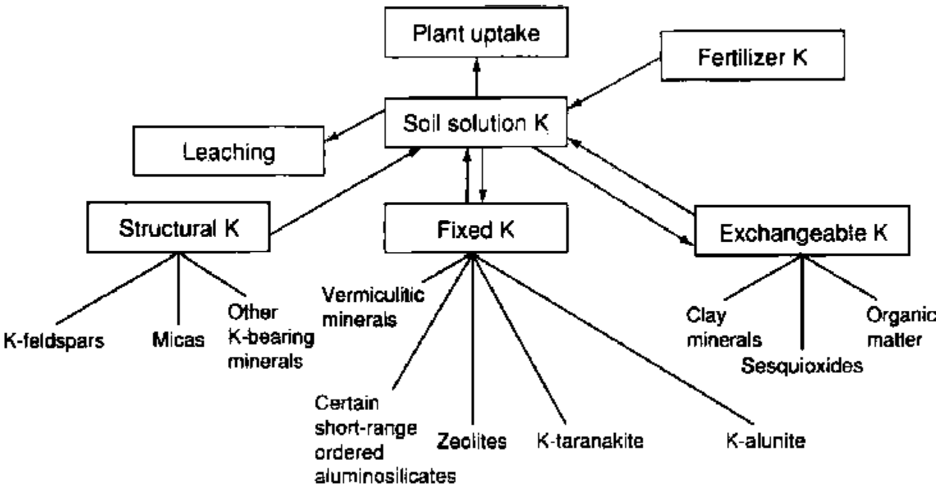


**Figure 1** Interrelationship of various forms of soil K. Reproduced with permission from Sparks DL and Huang PM (1985) Physical chemistry of soil potassium. In: Munson RD (ed.) *Potassium in Agriculture*. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

### Fixed K

Fixed K (nonexchangeable K) is different from structural K in that it is not bound within the crystal structures of soil minerals. It is held in the interlayer space of weathered micas, vermiculites, and hydroxy-interlayered vermiculites. Potassium becomes fixed because the binding forces between K and the mineral surfaces are greater than the hydration forces between individual K ions. This results in a partial collapse of the crystal structures and the K ions are physically trapped to varying degrees, making K release a slow, diffusion-controlled process. Fixed K can be found in a series of soil components and is also present in wedge zones of weathered micas and vermiculites. Cations with a size similar to $K^+$, e.g., $NH_4^+$ and $H_3O^+$, can exchange K from wedge zones of interlayers. Large hydrated cations, such as $Ca^{2+}$ and $Mg^{2+}$, cannot fit into the wedge zones and, thus, cannot release fixed K. Fixed K is moderately to sparingly available to plants. Release of fixed K to the exchangeable form can take place when levels of exchangeable and solution K are decreased by plant removal, microbial activity, and leaching. The ability of plants to take up nonexchangeable K appears to be related to the plant species.

### Structural K

Most of the total soil K is in the form of structural K, mainly as K-bearing primary minerals such as feldspars and micas. The total soil K is present in a series of particle-size fractions. The distribution of K in feldspars and micas in a wide range of particle-size fractions of moderately weathered soil profiles is illustrated in Table 1. The K-feldspars and K-micas of soils serve as an important reserve of available K, although structural K is generally assumed to be only slowly available to plants. The availability of structural K depends on the dynamics of K release from micas and K-feldspars, which are, in turn, related to the crystal structure and atomic bonding of these K-bearing primary minerals, the degree of weathering of these minerals, and the level of K in other forms. A substantial amount of K is released from the sand fractions of Delaware soils that are high in K-feldspars. This, along with the large quantities of mineral K in these and other Atlantic coastal plain soils, helps to explain the often observed lack of crop response to K fertilization on these soils.

## Ion Selectivity and Fixation of Potassium in Soils

### Ion Selectivity

Potassium selectivity of soils is affected by their mineralogical and chemical properties.

Mineralogical properties The layer charge, particle size, configuration of interlayers and wedge zones, position of wedge zones in interlayers, hydroxy interlayers, and variable-charge minerals are important mineralogical properties that affect K ion selectivity of soils.

The site of negative charge and the charge density of minerals may affect $K^+$ selectivity. The high $K^+$ selectivity of clay-size muscovite is attributed to the negative charge arising in the tetrahedral sheet. The $K^+/Ca^{2+}$ cation exchange selectivity (CES) values, expressed as the equivalent ratio of exchangeable $K^+$ to $Ca^{2+}$, range from 0.1 for montmorillonite to 1.7 for muscovite. This corresponds to the order of

**Table 1** Distribution of feldspar K in various size fractions of a Haverhill soil profile (Aridic Haploboroll) in Saskatchewan, Canada

| Horizon | Particle size (μm) | Total K (g kg$^{-1}$) | Feldspar K (g kg$^{-1}$) | Mica K (g kg$^{-1}$) | % fraction of K<br>Feldspar | <br>Mica |
|---|---|---|---|---|---|---|
| Ap | <0.2 | 21.1 | – | 21.1 | – | 100 |
| | 0.2–2 | 26.1 | 0.8 | 5.3 | 3 | 97 |
| | 2–5 | 20.1 | 6.1 | 14.0 | 30 | 70 |
| | 5–20 | 17.3 | 8.7 | 8.6 | 50 | 50 |
| | 20–50 | 15.4 | 9.7 | 5.7 | 63 | 37 |
| | 50–500 | 12.7 | 7.6 | 5.1 | 60 | 40 |
| | 500–2000 | 17.3 | 17.3 | – | 100 | – |
| | >2000$^{a}$ | 16.9 | 6.8 | 10.1 | 40 | 60 |
| Bm | <0.2 | 18.3 | – | 18.3 | – | 100 |
| | 0.2–2 | 29.9 | 0.0 | 29.9 | – | 100 |
| | 2–5 | 24.6 | 3.6 | 21.0 | 15 | 85 |
| | 5–20 | 20.4 | 7.7 | 12.7 | 38 | 62 |
| | 20–50 | 14.7 | 8.5 | 6.2 | 58 | 42 |
| | 50–500 | 14.2 | 8.4 | 5.8 | 59 | 41 |
| | 500–2000 | 18.2 | 18.2 | – | 100 | – |
| | <2000 | 17.6 | 6.5 | 11.1 | 38 | 63 |
| Ck | <0.2 | 16.8 | – | 16.8 | – | 100 |
| | 0.2–2 | 27.4 | 2.1 | 25.3 | 8 | 92 |
| | 2–5 | 26.2 | 8.7 | 17.5 | 33 | 67 |
| | 5–20 | 22.7 | 9.2 | 13.5 | 41 | 59 |
| | 20–50 | 15.6 | 8.3 | 7.3 | 53 | 47 |
| | 50–500 | 11.9 | 9.2 | 2.7 | 77 | 23 |
| | 500–2000 | 20.7 | 20.7 | – | 100 | – |
| | <2000 | 17.7 | 6.9 | 10.8 | 39 | 61 |

[a] Whole mineral fraction.

Reproduced with permission from Somasiri S, Lee SY, and Huang PM (1971) Influence of certain pedogenic factors on potassium reserves of selected Canadian Prairie soils. *Soil Science Society of America Proceedings* 35: 500–505.

increasing layer charge and surface charge density of the minerals.

Potassium selectivity of soils varies with particle size. It increases with particle size in some soils. Potassium selectivity may decrease with particle size at later stages of weathering. Much of the K selectivity apparently occurs in the interlayer regions.

Potassium selectivity is influenced by the configuration of interlayers and the position of the wedge zone in the interlayer. Instead of large hydrated ions such as Ca$^{2+}$ and Mg$^{2+}$, K$^{+}$ is selected due to the space limitations for diffusion of these large hydrated cations in the wedge zones. If the wedge zone is near the edge of the particle, it can be effective in only a small amount of K selection. If the wedge zone is deep within the particle, much more K can be selected. Wedge zones increase as more interlayers are opened and decrease as the interlayers are opened all the way through the particle.

Potassium selectivity is enhanced by the presence of hydroxy-Al and hydroxy-Fe(III) polymers in the interlayer space. These interlayer materials act as props to hinder the collapse of the layers about the K ions. Potassium ions are able to move more easily in propped-open interlayer spaces than in collapsed interlayers (**Figure 2**). Besides the propping effect, the preferential occupation by the hydroxy-Al polymers of the interlayer exchange sites that would normally adsorb Ca may also affect K selectivity. Due to their size, the polymers would not be adsorbed at wedge zones but rather in a more expanded interlayer position (**Figure 3**). Consequently, sites that would normally adsorb Ca$^{2+}$ ions would be occupied by Al polymers, whereas wedge sites would be relatively unaffected by the adsorption of hydroxy-Al

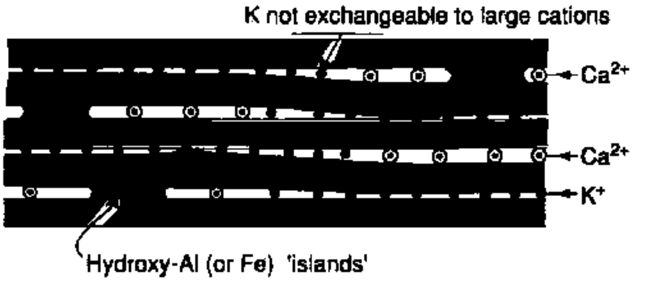


**Figure 2** Proposed model of an expandable layer silicate with interlayers indicating effect on K fixation. Reproduced with permission from Rich CI (1968) Mineralogy of soil potassium. In: Kilmer VJ, Younts SE, and Brady NC (eds) *The Role of Potassium in Agriculture*. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

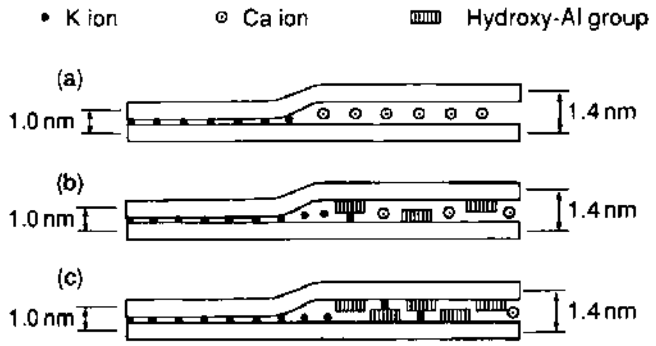


**Figure 3** Model of the effects of hydroxy-Al interlayers on K exchange. (a) In the absence of hydroxy-Al interlayers; (b) the propping effect and preferential occupation of Ca-adsorbing sites by hydroxy-Al interlayers; (c) the retarding effect of hydroxy-Al interlayers on the entry of large hydrated Ca ion. Reproduced with permission from Huang PM (1987) Aluminum and the fate of nutrients and toxic substances in terrestrial and freshwater environments. In: Singh M (ed.) *Encyclopedia of Systems and Control*. Oxford: Pergamon Press.

polymers. Hydroxy-Al interlayers may also exert a retarding effect on the entry of the more hydrated $Ca^{2+}$ ions. This would result in the increase of the selectivity toward the less hydrated K ions.

Besides layer silicates, variable-charge minerals also contribute to adsorption and selectivity of K ions. The contribution of these minerals to K selectivity is pH-dependent. Further, both cationic and anionic environments have a profound influence on the adsorption of $K^+$ and $Ca^{2+}$ in metal oxides such as birnessite (Mn oxide). This is of significance in understanding the chemistry of K in variable-charge soils.

**Chemical properties** Besides mineralogical properties, chemical properties also affect K selectivity of soils. The hydrated ion size and valence greatly influence K selectivity. In soils containing partially weathered micas and vermiculites where wedge zones exist, $NH_4OAc$ removes more exchangeable K than does $Mg(OAc)_2$, although $Mg^{2+}$ is usually a better replacer of $K^+$ than $NH_4^+$. The wedge zones exclude the hydrated $Mg^{2+}$ due to steric hindrance and preferentially adsorb cations that can fit into the wedge sites. $MgCl_2$ can remove more $K^+$ than can $Mg(OAc)_2$ because of the depression effect of soil pH by $MgCl_2$ and the formation of $H_3O^+$ by exchange or hydrolysis of exchangeable and polymeric Al. The $H_3O^+$ ion is similar in size to K and is thus able to fit into the wedge zone and exchange $K^+$. A combination of $NH_4OAc$ and $MgCl_2$ is more effective in removing exchangeable $K^+$ than either electrolyte alone. This is attributable to $Mg^{2+}$ opening up the interlayer space and facilitating the exchange of $K^+$ by $NH_4^+$.

Ions do not migrate rapidly enough to account for the reaction rate involving $H_3O^+$ ions. The lifetime of an individual $H_3O^+$ is exceedingly short, $\sim 10^{-13}$ s., since all of the protons are undergoing rapid exchange. As long as there is a continuous film of $H_3O^+$ to a $K^+$ ion, the exchange of $H_3O^+$ for K could proceed rapidly. Potassium exchange of soils is enhanced by low pH. The effect of pH on ion selectivity varies with the associated cations. pH has little effect for $NH_4^+$. $NH_4^+$-saturation collapses the interlayer structure and, thus, excludes $H_2O$ which is the medium for transport of protons to and from the exchange sites. By contrast, the effect of pH for $Li^+$ and $Mg^{2+}$ is pronounced. The increased efficiency of $Li^+$ and $Mg^{2+}$ with a decrease in pH is attributed to an expansion of the interlayer structure, which facilitates the entry of water and protons and the exchange of $H_3O^+$ for K.

Potassium selection in the wedge zones is influenced by the concentration of $K^+$ in the soil solution. There is K selection at the wedge sites at low K concentration. As the $K^+$ concentration increases, the silicate layer collapses at the wedge sites, resulting in an entrapment of ions deeper in the interlayer space. A high $K^+$ concentration is necessary to initiate collapse of the interlayer space of vermiculite and to enhance K selectivity in the absence of wedge zones.

Temperature also affects K selectivity of soils. A rise in temperature decreases K selectivity of Delaware soils (loamy sand, Typic Quartipsamments; sandy loam, Typic Hapludults; silt loam, Typic Hapludults; clay minerals include vermiculite, chloritized vermiculite, mica, kaolinite, chlorite, and quartz). This indicates that K sorption decreases with increasing temperature. The Delaware soils contain wedge zones that would selectively exclude large hydrated cations such as $Ca^{2+}$ and $Mg^{2+}$ and preferentially select smaller cations such as $NH_4^+$ and $K^+$. As the temperature increases, the hydration of $Ca^{2+}$ and $Mg^{2+}$ is decreased relative to that of $K^+$. Therefore, the size is not as important in $K^+$ selectivity. Furthermore, increased temperature could collapse the interlayer space of hydroxy-interlayered vermiculite and, thus, decrease $K^+$ selectivity of the soils.

## Fixation of Potassium

Potassium fixation in soils has been extensively studied. The extent of K fixation is related to the type of clay minerals and their charge density, the degree of hydroxy interlayering, the concentration of $K^+$ ions and competing ions such as $NH_4^+$, the moisture content, and solution pH. The expansible 2:1 layer silicates such as weathered micas, vermiculites, and smectites are the major clay minerals responsible for K fixation in soils. Weathered micas and vermiculites fix K under moist as well as dry conditions, whereas smectites fix K only under dry conditions. The amount of K fixation by smectites is very small unless their charge density is high. Some soil smectites have a higher charge density and likely have wedge sites near mica-like zones where K selectivity is high and K fixation can take place.

The principal mineral responsible for K fixation in acid soil is probably dioctahedral vermiculite. Counteracting the effect of high-charge density on K fixation in many acid soils is the presence of hydroxy-Al and –Fe(III) polymers in the interlayer space. The polymers act as props between the unit silicate layers and inhibit collapse of the layers about the K ions. The presence of hydroxy-interlayer polymers in vermiculite increases the Gapon coefficient $K_G$ (K/Ca) from $5.7 \times 10^{-2}$ to $11.1 \times 10^{-2}$ l (mmol)$^{1/2}$. On the other hand, K fixation is reduced markedly.

Since the pioneer research on the effect of freezing and thawing on K fixation and release, the significance of wetting and drying and freezing and thawing has been recognized by numerous researchers. The degree of K fixation or release on wetting or drying depends on the type of colloids and the level of $K^+$ in the soil solution.

Potassium fixation in soil and clay systems is influenced by pH. The pH effect on K fixation is related to: (1) the concentration of $H_3O^+$ and the competing cations and (2) the extent of interlaying and the resultant charge neutralization and propping effect. A marked increase in K fixation in soils was observed when pH was raised to about 9 or 10 with $Na_2CO_3$. There was no fixation at pH values up to 2.5; the amount of K fixation increased very rapidly between pH 2.5 and 5.5. Fixation of K increased more slowly above pH 5.5. At low pH, the lack of K fixation is attributed to a high concentration of $H_3O^+$ and its ability to replace $K^+$. At pH >5.5, $Al^{3+}$ cations precipitate as hydroxy polycations, which increase in the number of OH groups as pH increases until a discrete gibbsite phase is formed. At pH about 8, $Al^{3+}$ does not neutralize the negative charge on the clay and cannot prevent K fixation. The increase in K fixation between pH 5.5 and 7.0 is attributed to the decreased number of hydroxy-interlayer material that decreases K fixation.

## Potassium Release from Soils

### Primary Minerals

Micas and feldspars are the main K reserves of soils (**Figure 1**). The release of K from micas and feldspars replenishes the K in the labile pool. Therefore, these K-bearing primary minerals have a significant impact on long-term soil K fertility.

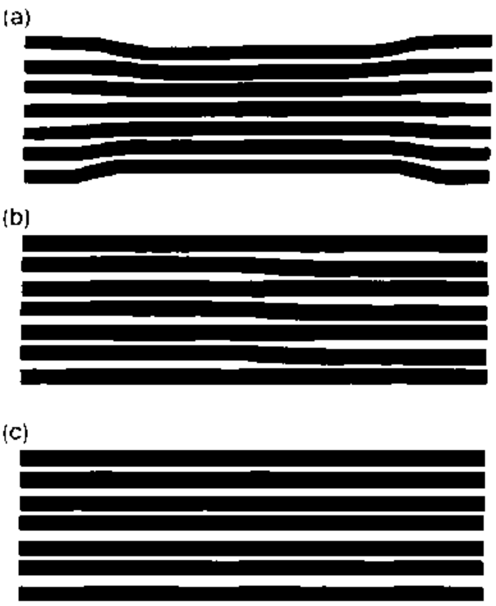


**Figure 4** Diagram of a weathered mica particle containing about 50% expanded (vermiculite) layer. (a) 'Frayed edge' and mica core; (b) alternate layers open halfway through interlayer; and (c) regularly interstratified mica-vermiculite (no wedge zone). Reproduced from Rich CI (1972) Potassium in soil minerals. *Proceedings of the 9th Colloquium of the International Potash Institute* 9: 15–31.

**Micas** Potassium release from micas may proceed by two processes: the transformation of micas to expansible 2:1 layer silicates by exchanging the $K^+$ with hydrated cations and the dissolution of micas resulting in the formation of weathering products. Both types of reaction occur in soil environments.

The transformation of micas to 2:1 expansible layer silicates through $K^+$ release proceeds by first, edge weathering and second, layer weathering. These weathering mechanisms are depicted in **Figure 4**. In models (a) and (b), the expanded interlayers are not continuous but have internal terminations forming 'wedge zones.' In model (c), interlayer spaces are either expanded all the way through the particle or nonexpanded uniformly through each layer. Edge weathering is most common in larger mica particles; in contrast, layer weathering is most common in smaller particles.

The mineralogical factors influence the release of $K^+$ from micas by both cation exchange and dissolution reactions. These include tetrahedral rotation and cell dimensions, tetrahedral tilting, hydroxyl orientation, chemical composition, particle size, structural imperfections, and layer charge alterations and related reactions. Ionic factors also substantially influence the release of K from micas. The hydronium ion, $H_3O^+$, is one of the dominant ions in acidic soil solutions. Reaction rates involving the $H^+$ ion are not restricted by diffusion rates of $H_3O^+$. Therefore, proton transfer may be rapid to the wedge zone at the junction of 'open' to 'closed micas.' The rate of $H_3O^+$–$K^+$ exchange may be limited by the rate of $K^+$ diffusion.

The sources of $H_3O^+$ in soils include water through its dissociation, the dissolution of $CO_2$ from the air and soil atmosphere, strong mineral acids originated from weathering reactions, and low-molecular-weight organic acids. Hydronium ions are probably the cause of major release of structural $K^+$ by weathering reactions. This mechanism is especially important in the soil rhizosphere, which is the most important plant feeding zone. Therefore, the extraction of soils with boiling 1 mol $l^{-1}$ $HNO_3$, dilute HCl solutions, and other acid dissolution techniques have been proposed for measuring the K-supplying power of soils.

The activity of $K^+$ ions in the soil solution greatly influences the release of $K^+$ from micas by cation exchange. Therefore, the K release can be greatly driven by $K^+$ leaching and especially by cropping without $K^+$ fertilizer application. When the K level is less than the critical value, $K^+$ is released from the interlayer by other cations from the solution. On the other hand, when the K level is higher than the critical level, the mica-expansible 2:1 phyllosilicates adsorb $K^+$ from the solution. The critical K level is highly mineral-dependent, being much lower for muscovite than for trioctahedral micas. The nature and concentration of the replacing cations also influence the critical K level in the solution. For example, the critical K levels decrease in the order $Ba^{2+} > Mg^{2+} > Ca^{2+} \cong Sr^{2+}$ for the same concentration of these ions and with a constant mica particle.

Many organic and inorganic ligands in the soil solution also affect K release from minerals. The kinetics of K release from muscovite and biotite is greatly influenced by the nature of organic acids which are common in soil environments, especially in localized zones where biological activity is intense, such as in the rhizosphere and near decomposing plant residues. Besides organic ligands, some strong complexing inorganic ligands such as phosphate can influence the kinetics of K release from micas. In the immediate vicinity of the phosphate fertilizer zone, the concentration of phosphate can be as high as 2.9–4.5 mol $l^{-1}$. The rates of K release from muscovite, biotite, and illite are significantly enhanced by 1 mol $l^{-1}$ $NH_4H_2PO_4$ (pH 4.0).

Redox potential of soils influences the $K^+$ release from micas because the interlayer K is more tenaciously held after oxidation of the structural Fe(II). Other factors being equal, the extent of $K^+$ release from micas appears to be less in oxidized soil environments then in reduced soil environments.

**Potassium-bearing feldspars** The release of K from feldspars through weathering reactions is an important process in replenishing the labile pool of soil K. The weathering rate of K-bearing feldspars appears to be affected by their structural properties which include the chemical composition, principally the inclusion of Na in the structure, the order of Si and Al in the structure, crystal and particle size, crystallization of more than one mineral within one particle, and the nature of twinning. The rate of weathering reactions of K-bearing feldspars probably increased with Si/Al disorder and Na content. A perthitic feldspar would likely weather more rapidly than a homogeneous particle. Albite would be weathered more rapidly in perthitic structures than the K-feldspar. Weathering is also likely more rapid at the junction of twins because of disorder at these planes.

Hydronium ions are an important chemical weathering agent of K-bearing feldspars. The initial weathering of K-bearing feldspars proceeds by surface reactions in which $K^+$ is displaced by $H_3O^+$. The substitution of $K^+$ by $H_3O^+$ weakens the feldspar structure. The kinetics of dissolution of $K^+$ from K-bearing minerals by 1 mol $l^{-1}$ $HNO_3$ indicates that, for the temperature range studied, the rates of $K^+$ release from muscovite, phlogopite, and biotite are about 2, 9–12, and 118 to 190 times, respectively, greater than the rate of $K^+$ release from microcline (Table 2). The release of $K^+$ from a tectosilicate is more difficult than from the interlayers of micas. The Arrhenius heat of activation for the release of $K^+$ from microcline is greater than for the micas. The heat of activation is interpreted as the energy level the structural K must acquire to be able to react. As the heat of activation increases, the rate constant decreases, since the mineral with higher heat of activation for the release of $K^+$ would release less K per unit time. The Arrhenius heat of activation and apparent rate constant for $K^+$ release from the K-bearing minerals (Table 2) are fundamental to understanding the availability of $K^+$ in feldspars and

**Table 2** Apparent rate constants and Arrhenius heats of activation for the release of structural K from K minerals

| *Mineral* | *Rate constant*<br>*301 K* ($h^{-1}$) | *311 K* ($h^{-1}$) | *Arrhenius heat of activation* ($kJ\,mol^{-1}$) |
|---|---|---|---|
| Biotite | $1.46 \times 10^{-2}$ | $3.09 \times 10^{-2}$ | 58.66 |
| Phlogopite | $9.01 \times 10^{-4}$ | $2.44 \times 10^{-3}$ | 77.81 |
| Muscovite | $1.39 \times 10^{-4}$ | $4.15 \times 10^{-4}$ | 85.52 |
| Microcline | $7.67 \times 10^{-5}$ | $2.63 \times 10^{-4}$ | 96.24 |

Reproduced from Huang PM, Crosson LS and Rennie DA (1968) Chemical dynamics of K-release from potassium minerals common in soils. *Transactions of the 9th International Congress of Soil Science* 2: 705–712 micas to plants. Although the release of $K^+$ from K-feldspars is generally more hindered than from micas, the K-feldspars are the largest natural reserves of K in many soils. Therefore, the role of K-feldspars in supplying K to plants cannot be ignored.

Besides $H_3O^+$ ions in soil solutions, complexing organic and inorganic ligands have significant roles in the release of K from the feldspar structure. Low-molecular-weight organic acids are very effective in releasing $K^+$ from K-bearing feldspars. Root growth and root length along with organic acids in the soil rhizosphere, thus, merit attention. In the immediate vicinity of the phosphate fertilizer zone, high concentrations of phosphate in the soil solution may also induce K release from K-feldspars.

Climate, topography, and the degree of leaching also influence the stability of K-bearing feldspars in soils. All experimental evidence shows that chemical breakdown of K-feldspars accelerates with raising temperature. Warm climate and high precipitation in tropical soils may have depleted all of their K-feldspars. However, these minerals may persist in temperate soils, especially in sand and silt fractions.

### Natural Soils

Chemical methods have been used to extract nonexchangeable K of soils. These include boiling $HNO_3$, $H_2SO_4$, hot HCl, electroultrafiltration, Na tetraphenylboron with ethylenediaminetetraacetic acid (EDTA), and ion exchange resins such as H- and Ca-saturated resins.

Soil K released to Ca-resin can be divided into three simultaneous rates of K release from the surface of the soil complex, the weathered periphery, and the micaceous matrix. The diffusion coefficient ($3 \times 10^{-19}$ $cm^2\ s^{-1}$) calculated for the slowest process, which is suggested to be from the mineral matrix, is of the same magnitude as that for illitic clay minerals. Comparing the values of rate constants for K release from the soils with those for K release from minerals, the rate constants of K release from the soils are within the range expected for K release from the mixtures of micas and K-feldspars. The rate constants of K release from six Great Plain soils by successive extraction with Ca-saturated cation exchange resin are highly correlated with mica contents of the soils, cumulative K uptake, and cumulative alfalfa yield.

The release of nonexchangeable K from soils is not only from clay-size K-bearing minerals but also from micas and feldspars in silt and sand fractions. The release of $K^+$ by the $H^+$-saturated ion exchanger from these size fractions as a function of time is demonstrated in **Figure 5**, based on the Elovich equation. The variation in the slope of the Elovich equation correlates significantly with the concentration of micaceous minerals in the sand, silt, and clay fractions, and may have an important impact on K for plant uptake.

Few studies have investigated the effect of anions on the kinetics of $K^+$ release from soils. The amounts of $K^+$ released from the Oxisol, Alfisol, and Entisol in $1\ mol\ l^{-1}$ $NH_4H_2PO_4$ (pH 4.0) for 4 weeks are 4.77, 114.0, and $207.8\ mg\ kg^{-1}$, respectively (**Figure 6**). The results clearly indicate that the K release in $1\ mol\ l^{-1}$ $NH_4H_2PO_4$ at pH 4.0 from the three soils is 2.3–2.9 times higher than that in $1\ mol\ l^{-1}$ $NH_4Cl$ at the same initial pH. Compared with the $NH_4Cl$ systems, the excess K released from the soils in the $NH_4H_2PO_4$ systems is attributed to the phosphate-induced alteration of K-bearing minerals of the soils, since $1\ mol\ l^{-1}$ $NH_4H_2PO_4$ induces alteration of micas and feldspars and the subsequent K release. The amount of the $NH_4H_2PO_4$-induced $K^+$ release is related to the degree of weathering of the soils, namely, Oxisol < Alfisol < Entisol. The combined effect of phosphate and proton on the alteration of K-bearing minerals and the formation of $NH_4$-taranakite are the major mechanisms of the $NH_4H_2PO_4$-induced $K^+$ release from the soils. In the immediate vicinity of the fertilizer zone, the P concentration is very high and the pH is quite low. Therefore, application of $NH_4H_2PO_4$ fertilizer to soils would promote dissolution of K-bearing minerals from the soils, thus enhancing their K-supplying rate.

## Availability of Soil Potassium

### Quantity/Intensity Relationships

The ratio of the activity of cations such as $K^+$ and $Ca^{2+}$ is defined by the relation $a_K/(a_{Ca})^{1/2}$ where $a$ is the ion activity. The concepts of quantity ($Q$) and intensity ($I$) have been applied to the mineral nutrient status of soils. Based on the ratio law, the $I$ of K in a soil at equilibrium with its solution could best be defined by the ratio $a_K/(a_{Ca} + a_{Mg})^{1/2}$ of the soil solution. This equilibrium activity ratio for K is referred to as $AR^K$ and has often been used as a measure of $K^+$ availability. However, it has been proposed that the activity of K in the soil solution is of primary importance as it is relevant for $K^+$ diffusion and K translocation in soils by mass flow to roots.

Soils showing the same value of $AR^K$ may not possess the same capacity for maintaining $AR^K$ while K is removed by plant uptake. Therefore, one must include not only the current potential of K in the labile pool but also the way in which the potential depends on quantity of labile K present. These findings brought about the classic $Q/I$ relationships where the activity ratio $AR^K$ value is related to the change in

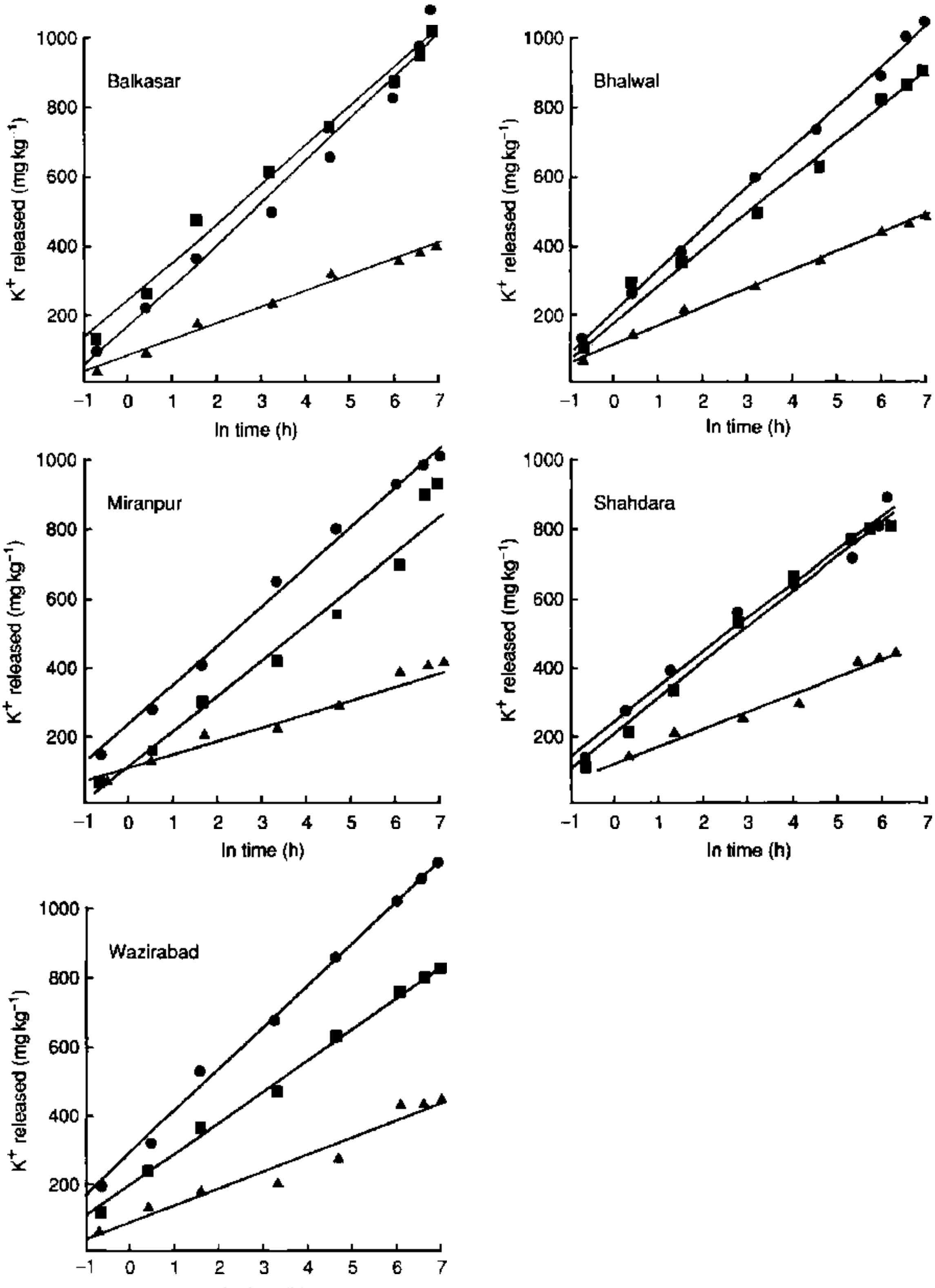


**Figure 5** Release of nonexchangeable K$^+$ from sand (triangles), silt (squares), and clay (circles) fractions. Reproduced with permission from Rahmatullah and Mengel K (2000). Potassium release from mineral structure by H$^+$ ion resin. *Geoderma* 96: 291–305.

exchangeable K to obtain the effect of quantity of labile K (exchangeable K) on intensity of labile K (AR$^K$). The *Q/I* concept has been widely disseminated in the soil literature in understanding soil K status.

The method used for constructing a typical *Q/I* curve involves equilibrating a soil with solutions containing a constant amount of $CaCl_2$ and increasing the amount of KCl. The soil gains or loses K to achieve the characteristic AR$^K$ of the soil or remains

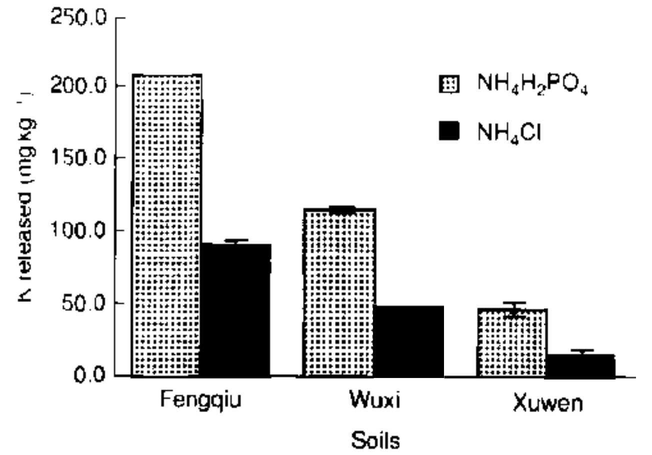


**Figure 6** Potassium released from the soil in 1 mol l$^{-1}$ $NH_4H_2PO_4$ and 1 mol l$^{-1}$ $NH_4Cl$ solutions (pH 4.0) at 25 C in 4 weeks. The Fenqiu, Wuxi, and Xuwen soils are Entisol, Alfisol, and Oxisol, respectively. The standard errors of the K release in $NH_4H_2PO_4$-Fengqiu and $NH_4Cl$-Wuxi soil systems were too small to be shown in the figure. Reproduced with permission from Zhou JM and Huang PM (1995) Kinetics of monoammonium phosphate-induced potassium release from selected soils. *Canadian Journal of Soil Science* 75: 197–203.

unchanged if its $AR^K$ is the same as the equilibrating solution. The $AR^K$ values are calculated from the measured concentrations of $K^+$, $Ca^{2+}$, and $Mg^{2+}$ converted to their ionic activities by application of the extended Debye–Hückel equation. The $AR^K$ values are then plotted versus the gain or loss of K to form the characteristic *Q*/*I* curve (**Figure 7**). Several parameters can be obtained from the *Q*/*I* plot to characterize soil K status. The value of $AR^K$ when the *Q* factor (ΔK) equals zero is a measure of the degree of K availability at equilibrium or $AR_e^K$. The value of ΔK when $AR^K$ equals zero is a measure of labile or exchangeable K in soils or ΔK°. The slope of the linear portion of the curve gives the potential buffering capacity of K ($PBC^K$) and is proportional to the cation exchange capacity of the soil. The number of specific sites for K ($K_x$) is the difference between the intercept of the curve and the linear portion of the *Q*/*I* plot at $AR^K$ equals 0. Advances in ion-selective electrode (ISE) technology have allowed for more rapid *Q*/*I* analysis.

Soils which have the same $AR_e^K$ may not have the same capacity for maintaining $AR_e^K$ while K is being uptaken by plant roots. Soils can have the same $AR_e^K$ but contain different amounts of labile K. K fertilization, liming, and cropping can change $AR_e^K$ values.

The ΔK° values are a better estimate of soil labile K than normal exchangeable K. Higher values of labile K (−ΔK°) indicate a greater $K^+$ release into the soil solution, resulting in a larger pool of labile K. The ΔK° becomes more negative or the labile K pool

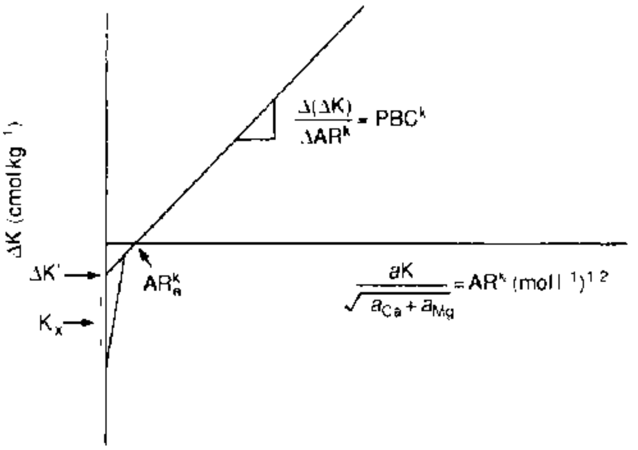


**Figure 7** Typical quantity/intensity (*Q*/*I*) plot. Reproduced with permission from Sparks DL and Huang PM (1985) Physical chemistry of soil potassium. In: Munson RD (ed.) *Potassium in Agriculture*. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

increases with K fertilization and liming on cropped soil.

The $PBC^K$ indicates the ability of a soil to maintain the intensity of $K^+$ in the soil solution. A high $PBC^K$ value is indicative of constant availability of $K^+$ in the soil over a long period. By contrast, a low $PBC^K$ value would indicate the need for frequent K fertilization. However, the lack of response of a Kalmia soil in the Delaware coastal plain in USA, which had a low $PCB^K$ value, to applied K was observed. This was attributed to the presence of the large amount of K-bearing minerals in the soil that with time could release $K^+$ to exchangeable and to solution form. Cropping and liming can change the $PBC^K$ values of soils.

### The Rhizosphere Chemistry and Labile Pool of Soil Potassium

The soil rhizosphere is the narrow zone of soil influenced by the root and exudates. The extent of the rhizosphere may vary with soil type, plant species, age, and many other factors, but it is usually considered to extend from the root surface out into the soil for a few millimeters. Microbial population in the rhizosphere can be 10–100 times larger than the populations in bulk soil. The rhizosphere is, thus, bathed in root exudates and microbial metabolites. Therefore, chemistry and bioavailability of soil K in the rhizosphere should differ from those in bulk soil.

A wide variety and considerable amount of organic compounds are released from plant roots in the rhizosphere. On average, between 30 and 60% of the net photosynthetic C is allocated to the roots in annual species; an appreciable portion (4–70%) of this C is released as organic compounds into the rhizosphere. The major components of rhizodeposition are sloughed-off cells and their lysates, mucilage, and low-molecular-weight organic compounds. Both the amount and proportion of the various organic compounds of root exudates vary substantially with plant species and cultivars. Moreover, the same plant cultivar grown in different soils varies in the kind and amount of low-molecular-weight organic acids present in the rhizosphere. Biological activity promotes $K^+$ release from micas. Besides the depletion of the K of the soil solution by plant roots, the overall action of biological activity is more complex when organic acids are produced. Organic acids substantially influence the dynamics of $K^+$ release from K-bearing minerals. The effect of organic acids in enhancing the release of K is attributed to the dissociated $H^+$ ions and complexing organic ligands. The weakening of the metal–O bonds by protonation of surface OH groups and the formation of inner-sphere surface complexes with organic ligands certainly enhance mineral dissolution and K release.

In view of the impact of plant species and cultivars and their interactions with soil properties on the formation of the kind and amount of organic acids and the influence of organic acids on the dynamics of $K^+$ release from the K-bearing minerals common in soils, the role of organic acids in the rhizosphere in affecting the labile pool of soil K merits close attention.

### Assessing Potassium Availability by Soil Tests

Most soils have a relatively large quantity of total K but relatively small amounts of plant-available K. Potassium soluble in water plus K adsorbed to exchange sites largely accounts for the labile pool of soil K. The exchangeable K is readily released to solution by the equilibrium process. During the course of a growing season, some initially nonexchangeable K may become water-soluble or exchangeable, and thus be plant-available.

The soil test methods for plant-available K are generally those that measure the soluble and exchangeable K or some parameter that relates to their equilibrium state. Soil test extractants for K are used to measure K easily and rapidly in soils and to estimate K availability. Based on the amounts of extractable K, recommendations based on field test calibrations can then be made on the amount of K required to obtain maximum yields.

Although soil tests for K usually measure the quantity of solution and exchangeable K, acids are usually used as extractants. Therefore, some nonexchangeable and structural K is also extracted. There are a number of soil test methods that are used to evaluate K availability of soils. The suitability of these soil test methods varies with soil types. The $1\ \mathrm{mol\,l^{-1}}$ $NH_4OAc$ at pH 7 procedure is probably the most widely used of all soil test methods for measuring so-called plant-available K on a global scale. Water-soluble $K^+$ or soil solution $K^+$ is such a small fraction of the total extracted by this method that it is not measured separately but is included along with that extracted (soluble and exchangeable). Both the $K^+$ in solution and that in the nonexchangeable form are in dynamic equilibrium with the exchangeable form. Therefore, the exchangeable K is sometimes referred to as a moving target.

Several other methods of determining exchangeable $K^+$ in soils have been developed for various soil conditions or for various methodological expediences. The Mehlich no. 1 (double acid) method ($0.05\ \mathrm{mol\,l^{-1}}$ HCl in $0.025\ \mathrm{mol\,l^{-1}}$ $H_2SO_4$) is considered best for sandy, acid, low CEC soils. The Mehlich no. 2 method ($0.2\ \mathrm{mol\,l^{-1}}$ HOAc in $0.015\ \mathrm{mol\,l^{-1}}$ $NH_4F$ in $0.2\ \mathrm{mol\,l^{-1}}$ $NH_4Cl$ in $0.012\ \mathrm{mol\,l^{-1}}$ HCl) is considered good for a wide range of soil conditions. The ammonium bicarbonate-diethylenetriamine pentaacetic acid (AB-DTPA) method is considered best for alkaline conditions. The Bray P-1 extractant is also used in some soil-testing laboratories to measure exchangeable $K^+$.

The soil test methods currently used in the USA are Mehlich no. 1 and Mehlich no. 3 procedures in the north-eastern and south-eastern regions, the Morgan and modified Morgan procedures in parts of the north-eastern region, and the $1\ \mathrm{mol\,l^{-1}}$ $NH_4OAc$ at pH 7 procedure in the western region.

For the interpretation of soil tests, it would be useful to know which $K^+$ quantity originates from the hydrated $K^+$ fraction and which from the nonhydrated $K^+$ fraction. The routine method which distinguishes between these two fractions is electroultrafiltration (EUF), which is used in Germany and Austria on a broad scale for $K^+$ fertilizer recommendation.

Soils which have long-term K-supplying power are attributed to the presence of sufficient K-bearing minerals (micas and K-feldspars). There are soils in the Punjab and Europe which hardly respond to K fertilization in long-term fertilizer trials. These soils release much $K^+$ from the silt fraction (Figure 5). Such soils differ much from sandy soils in their K fertilizer strategy.

### Plant Analysis for Potassium

Plant K analysis is an approach to assess soil K status. This method is based on the principle that if an element such as K is deficient in a plant, growth rates and yields will be decreased. If increasing amounts of K are added, the K concentration in the plant or plant part increases until an optimum level is

**Table 3** Critical K concentrations in agronomic crops

| *Crop* | *Time of sampling* | *Plant part* | *Critical concentration*[a] *(g kg$^{-1}$)* |
|---|---|---|---|
| Sugarbeet | | Blade | 10 |
| Cotton | | Leaves | 9–15 |
| Wheat | Jointing (GS6) | Total tops | 20–25 |
| | Early boot (GS9) | Total tops | 15–20 |
| Rice | Flag leaf to mid-tillering | | 10–14 |
| Corn | At tassel | Ear leaf | 19 |
| | At tassel | Leaves | 17–27 |
| | At silk | Sixth leaf from base | 13 |
| | At silk | Leaf opposite and just below ear shoot | 20 |
| Grain sorghum | Full heading | Second blade below apex | 18 |
| Alfalfa | | Whole top | 8–22 |
| Red clover | | Tops | 15–22.5 |
| Bermudagrass | | Tops | 13–15 |
| Orchardgrass | | Tops | 23–25 |
| Tall fescue | | Tops | 24–38 |
| Kentucky bluegrass | | Tops | 16–20 |

[a] Critical concentration is that nutrient concentration at which plant growth begins to decrease in comparison with plants above the critical concentration. After Westerman RL (1990). Reproduced from Sparks DL (2000) Bioavailability of soil potassium. In: Sumner ME (ed.) *Handbook of Soil Science*. Boca Raton, FL: CRC Press.

attained. This approach is used to express growth and yield as a percentage of the maximum. The zone between the deficient and optimum concentration is referred to as the transition zone. The transition zone is the zone between nutrient concentrations that cause a 20% reduction in growth or yield and those that produce optimum or 100% in the maximum yield. This latter zone is referred to as the critical range where researchers select the yield reduction and nutrient concentrations that are acceptable. The critical K concentrations for various agronomic crops are shown in **Table 3**. These critical K concentrations are determined by relating yield (percentage of maximum yield or growth rate) to K concentration (g kg$^{-1}$) for a specific plant part sampled at a given stage of development.

*See also*: **Fertigation**; **Fertility**; **Fertilizers and Fertilization**; **Macronutrients**

## Further Reading

Huang PM (1987) Aluminum and the fate of nutrients and toxic substances in terrestrial and aquatic environments. In: Singh M (ed.) *Systems and Control Encyclopedia*, pp. 262–268. Oxford: Pergamon Press.

Huang PM (1989) Feldspars, olvines, pyroxenes, and amphiboles. In: Dixon JB and Weed SB (eds) *Minerals in Soil Environments*, 2nd edn, pp. 975–1050. Madison, WI: Soil Science Society of America.

Huang PM, Crosson LS, and Rennie DA (1968) Chemical dynamics of K-release from potassium minerals common in soils. *Transactions of the 9th International Congress of Soil Science* 2: 705–712.

Jackson ML (1963) Interlayering of expansible layer silicates in soils by chemical weathering. *Clays and Clay Minerals* 11: 29–46.

Marschner H (1998) Soil–root interface: biological and biochemical processes. In: Huang PM, Adriano DC, Logan TJ, and Checkai RT (eds) *Soil Chemistry and Ecosystem Health*, pp. 191–231. Madison, WI: Soil Science Society of America.

McClean EO and Watson ME (1985) Soil measurements of plant-available potassium. In: Munson RD (ed.) *Potassium in Agriculture*, pp. 278–308. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

Mengel K and Kirkby EA (2001) *Principle of Plant Nutrition*, 5th edn. Dordrecht: Kluwer Academic.

Rahmatullah and Mengel K (2000) Potassium release from mineral structures by $H^+$ ion resin. *Geoderma* 96: 291–305.

Rich CI (1968) Mineralogy of soil potassium. In: Kilmer VJ, Younts SE, and Brady NC (eds) *The Role of Potassium in Agriculture*, pp. 79–108. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

Rich CI (1972) Potassium in soil minerals. *Proceedings of the 9th Colloquium of the International Potash Institute* 9: 15–31.

Scott AD (1968) Effect of particle size on interlayer potassium exchange in micas. *Transactions of the 9th International Congress of Soil Science* 2: 649–660.

Somasiri S, Lee SY, and Huang PM (1971) Influence of certain pedogenic factors on potassium reserves of selected Canadian Prairie soils. *Soil Science Society of America Proceedings* 35: 500–505.

Song SK and Huang PM (1988) Dynamics of potassium release from potassium-bearing minerals as influenced by oxalic and citric acids. *Soil Science Society of America Journal* 52: 383–390.

Sparks DL (2000) Bioavailability of soil potassium. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. D38–D53. Boca Raton, FL: CRC Press.

Sparks DL and Huang PM (1985) Physical chemistry of soil potassium. In: Munson RD (ed.) *Potassium in Agriculture*, pp. 201–276. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

Westerman RL (ed.) (1990) *Soil Testing and Plant Analysis*. Madison, WI: Soil Science Society of America.

Zhou JM and Huang PM (1995) Kinetics of monoammonium phosphate-induced potassium release from selected soils. *Canadian Journal of Soil Science* 75: 197–203.

# PRECIPITATION, WATERSHED ANALYSIS

**J V Bonta**, USDA Agricultural Research Service, Coshocton, OH, USA

Published by Elsevier Ltd.

## Introduction

Variability in current and historic weather and long-term climate, acting on spatially variable watershed characteristics, is the driving force that causes observed variability in stream flows, sediment transport, and water quality. Watershed characteristics such as spatial distribution of soil-water storage, hydraulic conductivity, tension characteristics, subsurface flow, and drainage affect flow, erosion, and water-quality sources and transport in a watershed. Weather varies significantly with time at a point on the landscape. Climate and weather also vary spatially over a watershed from one area to another and with elevation, and temporally from season to season.

The most obvious component of weather that affects watershed variability is precipitation, which has both deterministic and random components. Errors in measurement of precipitation, streamflow and other environmental variables, temporal and spatial variability and inadequate quantification of processes of water, sediment, and chemical movement in watersheds cause uncertainties in estimating the impacts of land-management changes on hydrology and water quality. This article summarizes current understanding of watershed processes, precipitation characterization and measurement, and watershed, weather, and precipitation modeling.

## Measurement of Precipitation

Precipitation measurements and characterization have inherent inaccuracies that must be recognized before they are used in computer models and in other analyses. Accurate precipitation measurement is complicated because of its significant spatial and temporal variability, and because of its different forms (liquid, solid, mixture). There are three components to precipitation measurement – sensing, representativeness, and resolution. Precipitation sensing involves using a sensor to measure a control volume of atmospheric particles presented to the sensor. Representativeness refers to how accurately the sensed particles represent the volume of precipitation actually reaching the ground, which is important to evaluating a watershed's response to precipitation. Resolution refers to how often (time resolution – minute, hourly, daily) precipitation is measured, and to the depth increment measured (depth resolution – 0.1 mm, 1.0 mm, etc.).

Precipitation measurements must be accurate and representative because watershed model studies have shown that small changes in precipitation can cause large changes in watershed response, highlighting the nonlinearity of watershed systems. Few multidecadal records of representative short-time increment precipitation data with good spatial coverage exist. Synthesized precipitation data are often unrepresentative because precipitation 'data' synthesized by precipitation models are parameterized using the nonrepresentative raw precipitation data. Much has been learned from precipitation-measurement studies that can provide guidance for future precipitation measurements.

### Precipitation Sensing and Resolution

A storage gauge is typically a cylindrical container with the orifice about 1 m off the ground. The depth of precipitation is often recorded every 24 h.

A recording weighing gauge (**Figure 1a**) weighs the precipitation that enters the orifice, and records the weight with a strip chart or data logger. The weighing gauge gives the best measurement of precipitation because the data can be collected at any time or depth resolution (e.g., 1 min and 0.25 mm); there is no limit on high intensities. Also, beginning and ending times of bursts of precipitation can be determined.

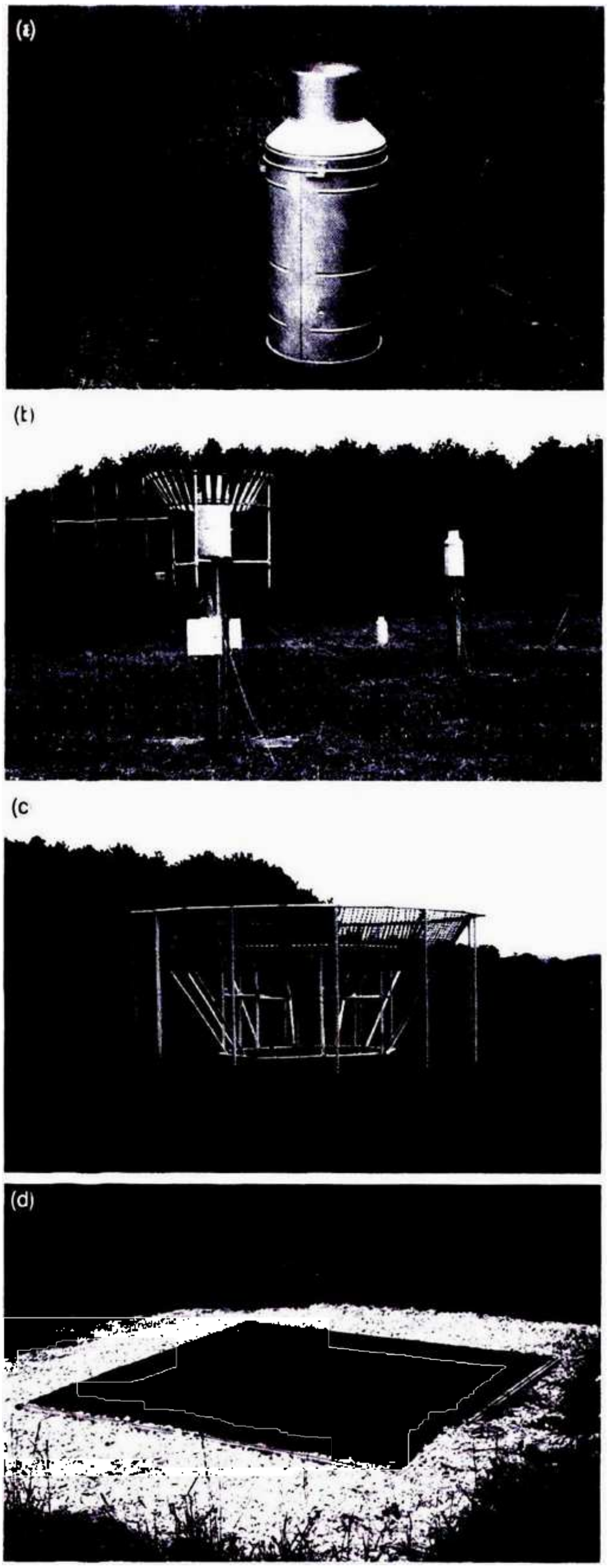


**Figure 1** Types of precipitation gauges and shield configurations at the North Appalachian Experimental Watershed at Coshocton, Ohio: (a) standard gauge; (b) dual-gauge system; (c) Wyoming shield; and (d) pit gauge

In tipping-bucket gauges, water from the orifice is funneled to a small hole where it drops into one side of a small 'bucket' that fills to a known volume before it empties its contents (e.g., 0.1 mm). This device is often used because tips are easily recorded as pulses with a data logger. However, the gauge must be calibrated, because at large rainfall intensities the system becomes overwhelmed and the volume of water from each tip of the bucket is less than the volume of a bucket. Therefore, long-term precipitation records can become truncated at larger intensities, often the ones of interest in watershed studies. Ponding above the buckets under high rainfall rates, and snow measurements can be problems. If heaters are used to melt frozen precipitation, they cause lower measurements because of evaporation and convection currents. Accurate beginning and ending times of storms are unknown with these gauges because water must accumulate in a bucket before it tips.

### Precipitation-Measurement Representativeness

Standard precipitation gauges (**Figure 1a**) typically underestimate true ground-level precipitation, the measurement of interest in watershed modeling and analysis. This is because wind diverts many smaller-diameter rain drops over the top of the gauge orifice. Light-density snow flakes are particularly vulnerable to transport by the wind, bypassing the orifice. Lysimeters (large isolated blocks of soil that act as a ground-level gauge) at the North Appalachian Experimental Watershed (NAEW) at Coshocton, OH, on average measured 9% more annual precipitation and 27% more snow and mixed precipitation than an adjacent recording gauge. Other studies of wind effects on precipitation-gauge catch found that a standard gauge can measure as much as 60% less snowfall than measured at ground level on windy days.

Many studies have shown that shielding gauge orifices increases the precipitation catch. A ring of freely hanging baffles surrounding a gauge orifice (Alter shield) has been shown to be about 75% effective in measuring ground-level precipitation (depending on wind speed). A Wyoming gauge shield (**Figure 1c**) composed of a standard gauge at the center of two concentric wind shields made of snow fence, and a dual-gauge system (**Figure 1b**) composed of two elevated standard gauges (one shielded and one unshielded) have been shown to give representative ground-level precipitation measurements.

A pit gauge (**Figure 1d**) is considered to be the standard measurement system for rainfall. A coniferous forest opening is considered the best place to measure snow. A Tretyakov gauge composed of dual concentric octagonal snow fences surrounding a gauge is an alternative standard for measuring snow because forest openings are often not available. A round inverted bell-shaped shield (Nipher shield) that surrounds a gauge gives good measurements of snowfall.

Representativeness is also affected by terrain slope. Horizontal orifice measurements can be much different compared with measurements from orifices parallel to the slope. Another factor affecting representativeness is obstacles in the vicinity of precipitation gauges.

Representativeness is also important over large areas as discussed later.

## Characterization of Precipitation

Even though precipitation is highly variable across both small and large areas, from season to season, and from storm to storm, hydrologists have developed methods to quantify precipitation for use in models for watershed analysis. Characterization of precipitation involves spatial and temporal components important for watershed evaluations. Recent research has focused on stochastic simulation methods to generate long records of precipitation with similar statistical characteristics for the location of interest. Practical characterization of spatial variability for modeling purposes is a current research topic.

Four elements of a precipitation record require characterization for different purposes: (1) dry time between storms (TBS); (2) storm duration; (3) storm depth; and (4) within-storm intensity variation and corresponding mass curves.

### Dry Times Between Storms

In any precipitation record (**Figure 2**), bursts ($B_i$) of varying intensities of precipitation are surrounded by dry periods ($D_i$). These TBS periods can range from a minute to months. A question facing an analyst is: what is the minimum dry-period duration (MDPD) for which $D_i >$ MDPD separates storms, and $D_i <$ MDPD is included in a storm. Intuitively, a 5-min period of no rain most likely will be included in the same storm, but a 5-month period would not. Quantification of MDPD is the first step in analyzing a record of precipitation data because storms can be identified, providing a data set for analysis of the other three elements. MDPD can be found by analyzing precipitation data independent of, or dependent on, watershed characteristics.

Arbitrary fixed-storm separation techniques are often used to identify storms. However, a fixed MDPD does not capture the spatial and monthly characterization of dry periods required for some analyses. Other specific storm-identification methods can be assigned, such as used in determining the 'R factor' (rainfall-runoff erosivity factor) of the Revised Universal Soil Loss Equation (RUSLE), a model used extensively for estimating soil loss on hill slopes (*See* Erosion: Water-Induced).

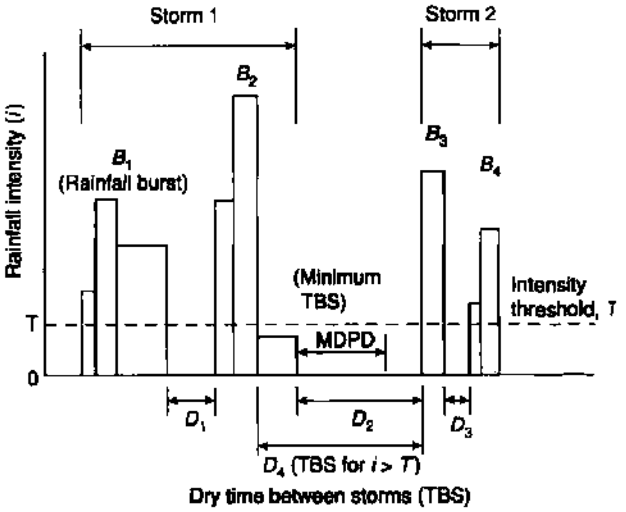


**Figure 2** Definition sketch for identifying storms in a precipitation record. See text for details

Rank correlation uses equally spaced time intervals of rainfall intensities that are lagged by increasing time intervals until time intervals are not correlated. The rank correlation method uses the rainfall intensity data to determine the dry time between storms instead of using the dry periods. Autocorrelation analysis has also been applied but the procedure does not always give well-defined estimates of MDPD.

Another method of determining MDPD is to assign a fixed time period for rainfall intensities greater than a threshold intensity, such as $T$ in **Figure 2**. In this case, $D_2$ would be replaced by $D_4$. $T$ represents an unknown parameter that is likely to be seasonally and spatially dependent, and interpretation must be qualified.

Another technique is called the 'exponential method.' In this method, MDPD is obtained by approximating groups of bursts of rainfall and dry periods (storms) as a Poisson process. Under this assumption, storms are considered statistically independent 'events' and the interarrival times of events are described by the exponential frequency distribution. In this method, dry periods greater than MDPD form an exponential distribution, and dry periods less than MDPD are included in storms. The exponential distribution is a good method to represent TBS, as it characterizes seasonal variation of TBS. Examples of best-fit exponential frequency distributions for several months at Bridgeport, NE, are shown in Figure 3.

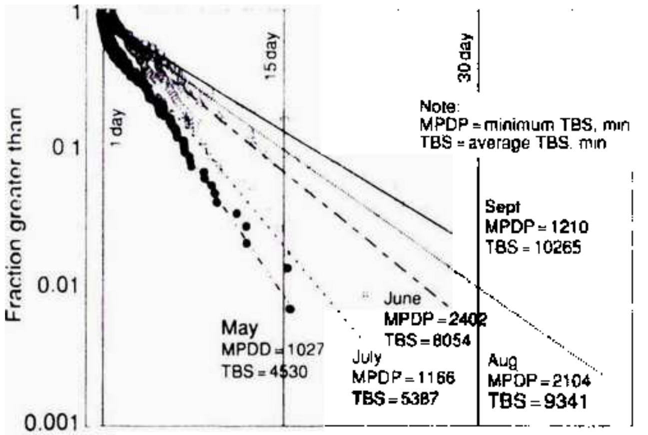


**Figure 3** Seasonal variation of exponential frequency distributions of times between storms (TBS) for Bridgeport, Nebraska, USA. -●- May; -○-June; -#- July; --△--August; -□-September.

The average TBS is the parameter of the exponential distribution.

Finally, some precipitation analyses have not used the concept of an MDPD value, but used only daily precipitation totals. For example, during the development of the watershed curve number (CN) technique, used worldwide for estimating watershed runoff potential, daily (and not storm) precipitation was used because these were the most common data available from a storage-gauge network. Recent investigations of the CN technique now match actual storms with observed runoff data for curve-number development.

Most methods of determining MDPD assume that storms occur independently from the watershed. However, as in the case of determining precipitation that causes observed runoff hydrographs used in determining a watershed CN, identifying storms that incorporate watershed characteristics will give different results. This is because identifying precipitation that causes runoff events includes watershed storage that attenuates watershed response to rainfall intensities. Increasing watershed size increases storage effects, and spatial variability of precipitation also becomes more important. One or more storms (or partial storms) identified using the exponential method may comprise the causal storm for a watershed hydrograph. The method used for identifying causal precipitation for CN analysis incorporates the threshold $T$ value in **Figure 2.**

Stochastic simulation of the occurrence of individual storms (alternative to using 24-h total precipitation) is possible by sampling the appropriate monthly exponential frequency distribution of TBS (**Figure 3**) using the Monte Carlo technique.

### Storm Depths and Durations

Storm depths and durations are usually analyzed together. Often daily total precipitation has been used, such as in the development of the CN method, and individual storm durations are ignored. However, times of observation of 24-h totals may be inconsistent, because precipitation may not necessarily have been recorded over the same 24-h period. Yet precipitation analyses often assume 24-h totals at midnight. This inconsistency has led to the approach of identifying individual storms using the exponential method.

One form of depth and duration data characterization is through intensity–duration–frequency curves (IDFs). IDFs summarize the return periods of different intensity rainfalls for a given duration (e.g., 5-year, 30-min intensity). IDFs are widely used in engineering design of culverts and sewerage systems through peak-runoff estimation methods. They are also used in engineering design with synthetic mass curves of rainfall. IDFs represent only the maximum intensities in a storm, with no information on time distribution of rainfall. Maps of isopleths of IDF curves are often available. An assumption made when using IDF curves is that the return periods of precipitation and peak-runoff rate are the same.

Newer precipitation-analysis approaches eliminate the 24-h constraint and use storms as described above using the exponential method. Storm durations can range from a few minutes to several days. Once storms have been identified by monthly MDPD values, storm depth and durations can be analyzed for particular purposes. Storm depths and durations are typically correlated (**Figure 4**); however, there is much scatter. Storm simulation involves sampling from graphs like **Figure 4,** using the Monte Carlo method.

### Within-Storm Intensities and Mass Curves

There is an infinite number of ways that a storm total can be distributed, as evidenced by the sample of 181 storm mass curves at Coshocton, OH (**Figure 5**). Historically, the use of only daily precipitation totals has led to simple methods for disaggregating the 24-h or IDF total to determine the time distribution of precipitation. Fixed patterns of rainfall intensities ('design storms') have been used to distribute the total in time (**Figure 6**). The USDA Natural Resources Conservation Service often uses design storms for determining peak flow rates in design (type I, II, etc. curves). However, fixed patterns do not describe the variability found in nature (**Figure 5**).

The most satisfactory mass-curve representation has been found to be Huff curves (**Figure 6d**) because the wide variability of curves is similar to that

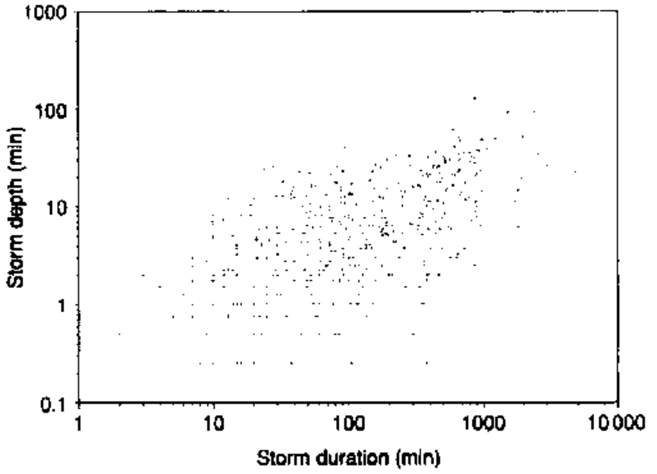


**Figure 4** Example of the correlation between storm depths and durations for July at the North Appalachian Experimental Watershed at Coshocton, Ohio, USA.

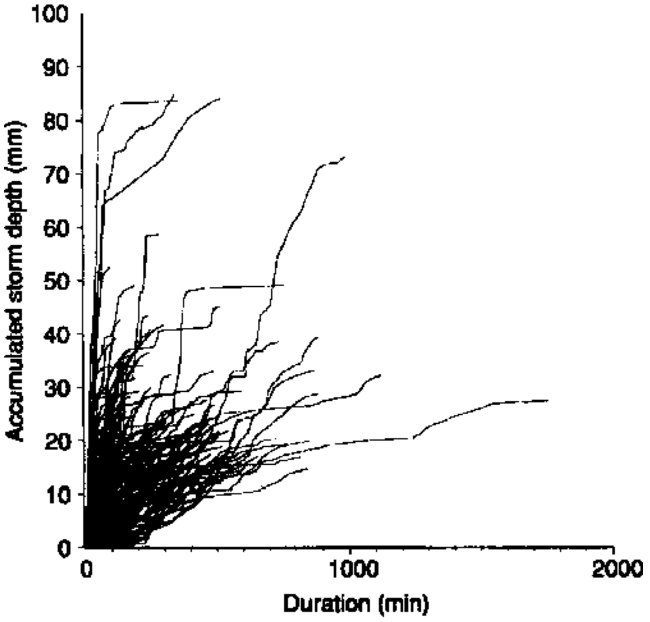


**Figure 5** Example of the wide variability in mass curves at the North Appalachian Experimental Watershed at Coshocton, Ohio, USA.

observed in natural precipitation. These are isopleths of the frequency of occurrence of dimensionless accumulated depths for given dimensionless elapsed times in a storm. There are several curves that can be used, compared with a single mass curve of precipitation in a design storm. Documented methods to use the information contained in Huff curves in watershed analysis are an unresolved issue.

### Spatial Variability

Spatial variability of storm depth and duration are often linked by snapshots of storm mass curves at the same time by using maps showing isopleths of

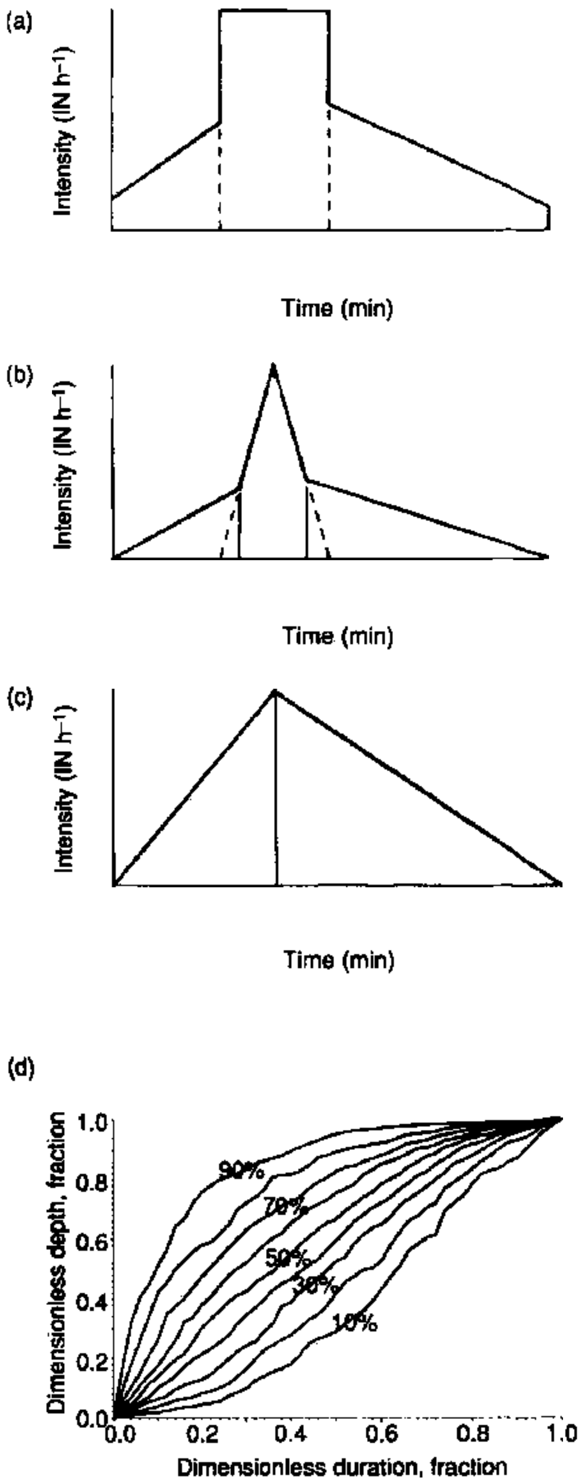


**Figure 6** Design storms used for distributing total precipitation (a) mixed trapezoids (Sifalda V (1973) Entwicklung eines Berechnungsregens für die Bemessung von Kanalnetzen. *Wasser Abwasser* 114: 435–440); (b) mixed triangles (Desbordes M and Raous P 1976) Un exemple de l'intérêt des études de sensibilité des modèles hydrologiques. *La Houille Blanche* 1: 37–43); (c) triangular (Yen BC and Chow VT (1980) Design hyetographs for small drainage-structures. *Journal of Hydraulics Division* 106: 1055–1076); and (d) Huff curves at a point (Bonta JV (1997) Proposed use of Huff curves for hyetograph characterization. In: Richardson CW *et al.* (eds). *Proceedings of Workshop on Climate and Weather Research*, pp. 111–124. USDA-Agricultural Research Service 1996–03

precipitation depth from several rain gauges during the storm (isohyetal maps). Isohyet interpolation can be by linear interpolation between known precipitation amounts, inverse-squared mapping, kriging, and by other methods. Average precipitation is computed by weighting isohyets by the area between isohyets. Thiessen polygons have also been used to capture the spatial variability of precipitation. These are polygons surrounding rain gauges such that the area enclosed by the polygons is closer to the enclosed rain gauge than to adjacent rain gauges. Polygon boundaries are the perpendicular bisectors between adjacent gauges. The polygons provide area-based weights for measured precipitation amounts.

Mapping of precipitation only grossly characterizes the spatial variability of precipitation. This is because precipitation distribution over an area can vary rapidly over time and with elevation in unknown ways over rolling hills and large mountains. Furthermore, precipitation depths can vary on the windward and leeward sides of mountains (i.e., 'rain shadow' effect). The effects of these factors can change monthly. Objective forms of mapping by interpolation do not consider orographic influences, and in mountainous areas, isohyets must often be adjusted manually.

One method of spatial characterization of precipitation that overcomes the problems of spatial variation with elevation is the use of conditional regression. This method of mapping utilizes regression relationships developed from precipitation data, and accounts for elevational, coastal, and other factors known to affect precipitation. Average monthly and annual precipitation have been mapped on a 4-km$^2$ grid across the USA. These maps are useful for general characterization, exploring relationships between average monthly weather elements and other variables (e.g., TBS), and mapping parameters for weather simulation models.

## Watershed Response to Precipitation due to Physical Characteristics

### Precipitation and Soil Variability

Watersheds that have similar appearance and land use respond differently to precipitation because of spatial and temporal variability of soil characteristics. Even watersheds having soils mapped in the same soil-map unit from an available soil survey and with the same land management can have significant differences in annual runoff. This is caused by natural spatial variability of soils and complex interactions between precipitation and surface–subsurface water. Studies in experimental watersheds attribute runoff differences to factors such as heavier-textured soil in different parts of the watershed, small seeps on the hillslope due to geologic clay layers that drain perched water to the seeps, and persistently high soil moisture in different areas. Heavier-textured soils and wetter watershed surfaces tend to run off sooner during the higher-intensity parts of storms than coarser-textured and drier areas, and runoff generation during a storm is nonuniform in the watershed. A local soil-survey map cannot necessarily be used to characterize small watersheds adequately because these factors are not typically mapped.

Watersheds exhibit complex interactions between surface and subsurface flow paths. Spatial and temporal variability of precipitation superimposed on watershed variability is an important factor to consider in understanding and modeling watershed hydrology. As watersheds become larger, more diversity in watershed characteristics can be expected.

### Watershed Processes

Infiltration at the land surface significantly affects the timing and amounts of high and low watershed flows. Infiltration during rainfall events is affected by soil type, porosity, texture, vegetative cover, land management, preferential flow in soil and bedrock due to roots, cracks, and soil fauna. Macropores in natural systems can contribute significant amounts of infiltrated water and pollutants quickly to shallow and deep waters under high-intensity rainfalls. Advanced infiltration models require short-time increment precipitation to simulate matrix and macropore flow.

Land management can significantly affect infiltration. For example, rainfall on a field plowed with a moldboard plow tends to cause development of a relatively impervious surface due to crusting, whereas surface residue protects the soil surface from crusting with no-till cropping.

Infiltration is also affected by the degree of saturation since the last rainfall event (TBS, described above) as influenced by evapotranspiration (ET) and drainage rates. ET and drainage between storm events vary with soil type, vegetative cover, position on the landscape, aspect, geology, land use, climate, and time and weather since the last rainfall (TBS). The rate of water loss in the soil affects the storage available for water at the beginning of the next rainfall event ('antecedent soil water conditions').

In some watersheds, water moves horizontally near the surface as 'interflow.' The interflow process stores storm water for slow release either during a storm or after a storm has ended. It can also contribute to wetter antecedent conditions. Its effect on runoff, water quality, and erosion depends on the time until rainfall starts again (TBS) and drainage and ET rates. In addition, interflow can re-emerge at the land surface, causing localized seeps (this is known as 'exfiltration') that immediately generates runoff during a storm.

Interflow, groundwater flow, and spatial variability of soil characteristics affect antecedent water conditions nonuniformly on a watershed surface. Typically, higher soil-water levels can be found in concave parts of the hillslope and along the periphery of stream channels. This is because soil water accumulates by gravity in these low areas during and between storm events. This leads to areas of watersheds that are wetter than others and that can run off sooner in a storm compared with drier areas. However, runoff can be randomly generated from any part of the landscape due to local factors. The nonuniformity of antecedent conditions leads to areas of the watershed that expand as areas become wetter during a storm, leading to a 'variable source area' of runoff generation. Runoff from these varying areas is sensitive to changing rainfall intensities during storms, affecting timing and amounts of runoff, chemicals, and sediment transported.

## Watershed and Weather Modeling

### Watershed–Precipitation Interactions

Watershed models are one tool for investigating the impacts of alternative land-management practices on hydrology, water quality, and sediment transport. Lack of data and simplifications of watershed processes are necessary. Often some processes are eliminated in modeling because of lack of understanding of these processes.

Few models adequately include the important processes of preferential movement of water as infiltration, subsurface flow, and interflow. The observed effects of scale of watershed on water yield must include a subsurface component to link upland and lowland areas quantitatively. Inadequate modeling of these processes and surface–subsurface precipitation interactions lead to errors in hydrology, erosion, and water-quality modeling.

Few models include spatial variability of infiltration and soil-water storage, except through land-surface discretization methods. These methods include representations of the land surface as square grids, triangles, and irregularly shaped areas. All these methods typically assume uniformity within the discretized element. Even within very small watersheds, however, there is significant variability in the distribution of soil characteristics. When small watersheds are subjected to precipitation, the spatial variability of characteristics can have significant effects on watershed hydrology and water quality. Sometimes a small area of a watershed may contribute a large fraction of sediment or chemicals downstream because it is wetter and tends to runoff sooner under rainfall, or it may receive anthropogenic inputs (e.g., fertilizer, vulnerable land-management practices). Another method for characterizing watershed elements is by using frequency distributions of infiltration characteristics within a 'uniform' element. Knowledge of the spatial variability of infiltration, runoff-producing areas, and antecedent conditions can be useful for developing management practices that minimize the effects of chemicals and eroded sediment.

### Watershed Models

Although understanding of natural watershed processes is deficient, there have been many attempts to model watershed hydrology and water quality. The CN model mentioned earlier is widely used for estimating runoff volumes from watersheds, and is often a component of more complex models. Models that measure responses to precipitation have been developed for field-level investigations of surface runoff, sediment, pesticides, subsurface water, leachate and water quality, and erosion on hillslopes.

### Weather and Precipitation Modeling

Weather and precipitation provide the deterministic and random driving forces in watersheds, and affect quantities and timing of runoff, sediment, and chemicals. Often measured weather and precipitation data are used as inputs to watershed modeling. Unfortunately, records are often not long, and short time-increment precipitation data exist only at points on a landscape. Furthermore, if several precipitation-gauge records are required, the periods of records may not overlap, the gauges may have had different sampling intervals (15-min vs 24-h), or the gauges inadequately characterize spatial variability of precipitation over watersheds. Consequently, weather and precipitation models have been developed. These models stochastically simulate weather for the long periods needed for analyzing modeled watershed outputs using frequency distributions. Stochastic simulation involves analyzing historic weather data to develop parameters that describe the frequency distributions of measured data. The models are designed so that the long records of weather and precipitation data have the same statistical characteristics as the measured data for a given location.

Two better-known weather models are CLIGEN and GEM (Generation of Weather Elements for Multiple Applications). Both models synthesize daily values of solar radiation, maximum and minimum air temperature, and precipitation, wind speed, wind direction, and dew point. Both models simulate daily precipitation that depends on whether there was precipitation during the previous day. Both models are 'point' models, meaning that simulated precipitation is modeled at a point and not over an area.

GEM additionally generates weather elements that are dependent on whether a day has precipitation ('wet' or 'dry' days). Intuitively, for example, on a day with rain, the sky is cloudy, and there is less likelihood of a large daily solar radiation value. Consequently, the frequency distribution describing solar radiation is different for dry and wet days. Furthermore, an advanced version of the GEM model incorporates spatially dependent weather elements. Extensive databases for CLIGEN have been created by developing parameters at individual gauges, requiring extrapolation to ungauged areas. GEM databases have been interpolated for many areas in the USA on a 4-km$^2$ grid.

### Spatial Variability of Weather and Precipitation

Under ideal conditions, weather inputs should be spatially available for watershed models at appropriate time steps, particularly for large areas. Often weather elements such as solar radiation and temperature do not vary greatly over large areas (except for variations in elevation), and a daily time resolution is adequate. However, precipitation intensities can vary greatly over small areas and characterization for both spatial and temporal variability for small areas and short time steps are required. Precipitation falling on areas of a watershed with potentially high runoff-generating potential (soil near saturation, feedlots, mined and urban areas, etc.), if not properly characterized in time and space, will result in inadequately modeled volume and timing of runoff. The spatial variability and uncertainties in precipitation, weather, and watershed modeling require watershed data and model outputs to be analyzed in terms of probabilities and frequency distributions (e.g., chemical loads, peak flows, runoff volumes).

## Summary

Precipitation characterization for analyzing watersheds depends on data that are representative of ground-level precipitation. Precipitation is characterized in terms of times between storms, storm durations and depths, within-storm intensities, and spatial variation of these storm elements. Spatial and temporal characterization of precipitation is important because watershed landscapes also have significant spatial variability in characteristics that affect runoff generation. Water-movement processes on watershed surfaces are inadequately understood and quantified. However, scientists and engineers have developed many methods and computer models for characterizing natural variability in storms and watersheds to solve practical runoff and water-quality problems.

Much remains to be learned regarding the interactions between precipitation and watershed characterizations. Factors that affect parameterization of weather and precipitation models are weather modification, changing climate, and persistence in weather caused by large-scale atmospheric forcings (e.g., El Niño Southern Oscillation). Atmospheric forcings cause persistent dry or wet periods lasting years in different parts of the world. Unless data are categorized to include these persistent periods, this effect will not be considered. Parameterization to reflect weather modification is lacking. Data sets from which parameters are developed may be changing because of potential changes in climate ('nonstationarity' data). Therefore, long periods of synthesized weather data for watershed modeling may not be representative of future conditions.

*See also:* **Climate Change Impacts**; **Climate Models, Role of Soil**; **Erosion:** Water-Induced; **Water Cycle**

## Further Reading

American Society of Civil Engineers (1996) *Hydrology Handbook*, (Manual No. 28), 2nd edn. Prepared by the Task Committee on Hydrology Handbook of Management Group D of the ASCE.

Amerman CR (1965) The use of unit-source watershed data for runoff prediction. *Water Resources Research* 1(4): 499–507.

Bonta JV (1997) Proposed use of Huff curves for hyetograph characterization. In: Richardson CW *et al.* (eds) *Proceedings of Workshop on Climate and Weather Research*, 17–19 July 1995, pp. 111–124. USDA-Agricultural Research Service 1996–03.

Bonta JV (1998) Spatial variability of runoff and soil properties on small watersheds in similar soil-map units. *Transactions of the ASAE* 41(3): 575–585.

Chow VT (1964) *Handbook of Applied Hydrology*. New York, NY: McGraw-Hill.

Daly GH, Gibson WP, Taylor GW, Johnson JL, and Pasteris PA (2002) A knowledge-based approach to the statistical mapping of climate. *Climate Research* 22: 99–113.

Desbordes M and Raous P (1976) Un exemple de l'intérêt des études de sensibilité des modeles hydrologiques. *La Houille Blanche* 1: 37–43.

Hershfield DM (1961) *Rainfall Frequency Atlas of the United States for Durations from 30 Minutes to 24 Hours and Return Periods from 1 to 100 Years*. Technical Paper No. 40. Washington, DC: US Government Printing Office.

Johnson GL, Hanson CL, Hardegree SP, and Ballard EB (1996) Stochastic weather simulation: overview and analysis of two commonly used models. *Journal of Applied Meteorology* 35: 1878–1896.

Larson LW and Peck EL (1974) Accuracy of precipitation measurements for hydrologic modeling. *Water Resources Research* 10(4): 857–863.

Maidment DR (1992) *Handbook of Hydrology*. New York, NY: McGraw-Hill.

McGuinness JL (1966) A comparison of lysimeter catch and rain gage catch. ARS-41-124.

Nicks AD and Gander GA (1994) CLIGEN: a weather generator for climate inputs to water resource and other models. *5th International Conference on Computers in Agriculture*. ASAE publication 3–94.

Restrepo PJ and Eagleson PS (1982) Identification of independent rainstorms. *Journal of Hydrology* 55: 303–319.

Sifalda V (1973) Entwicklung eines Berechnungsregens für die Bemessung von Kanalnetzen. *Wasser Abwasser* 114: 435–440.

Yen BC and Chow VT (1980) Design hyetographs for small drainage-structures. *Journal of Hydraulics Division* 106: 1055–1076.

# PRECIPITATION–DISSOLUTION PROCESSES

**W P Robarge**, North Carolina State University, Raleigh, NC, USA

## Introduction

Precipitation and dissolution reactions in soils and the underlying strata have long been of interest to soil scientists and soil chemists, because primary and secondary minerals are inherently part of many soil processes, e.g., chemical weathering, soil genesis, and nutrient availability (which can impact human and animal nutrition). More recently, the importance of precipitation and dissolution in soils has been emphasized because of environmental issues, e.g., mineral dissolution reactions as a buffering mechanism against inputs of acidic deposition, and formation of new solid phases as a means of remediation for trace metal contamination. Indeed, it is a mistake to attempt to limit the importance of precipitation and dissolution reactions in soils to mineral-weathering reactions or formation of secondary precipitates around fertilizer granules. Rather, it is necessary to acknowledge that precipitation–dissolution reactions span a range in physical scales that extends from the atomic and submicron through laboratory, watershed, and even continental processes.

Understanding precipitation and dissolution reactions in soils requires knowledge of the chemical and physical laws assumed to govern these processes. These laws are derived from consideration of either macroscale or microscale reactions. Macroscale processes are based on the science of thermodynamics and allow prediction of precipitation or dissolution under equilibrium conditions. The power of this approach is that it is possible to make predictions of precipitation or dissolution without any knowledge of the actual reactions that occur at mineral surfaces or during the formation of nuclei that lead to the formation of stable precipitates. The disadvantage of this approach is that it cannot easily predict the rate at which reactions will occur. Theories that focus on microscale reactions assume that specific chemical-reaction mechanisms occur at mineral surfaces and are more successful in predicting the rate of mineral dissolution due to changes in the composition of the surrounding soil solution. Unfortunately, equations predicting rates for microscale processes tend to be empirical in nature and require information specific to a given set of conditions that can usually be generated only within the laboratory. Most of the time, laboratory conditions only approximate the actual weathering environment in the soil. A knowledge of both macroscale and microscale processes is important in understanding precipitation and dissolution reactions in soils.

## Macroscale Processes

The dissolution of a mineral is a chemical reaction. Thermodynamics predicts that a chemical reaction will proceed when energy can be released. A simple chemical reaction can be written as follows:

$$A + B \Leftrightarrow C + D \qquad [1]$$

where chemical species A and B react to form chemical species C and D. The symbol ⇔ means that the chemical reaction has reached equilibrium and that there is essentially no net change in the concentrations of A, B, C, or D. Thermodynamics allows eqn [1] to be written in the following format:

$$K = \frac{(C)(D)}{(A)(B)} \qquad [2]$$

where $K$ is termed the equilibrium constant and parentheses refer to the activity of the chemical species in solution. The activity of a chemical species is a function of its concentration in solution. The equilibrium constant is related to the energy of the chemical reaction through the expression:

$$\Delta G_r = -RT \ln K \qquad [3]$$

where in thermodynamics $\Delta G_r$ is termed the standard Gibbs energy change, $R$ is the molar gas constant, $T$ is absolute temperature, and $K$ is the equilibrium constant when all the reactants and products are in their standard states. For sparingly soluble compounds, $K$ is written as $K_{so}$ and defined as the solubility product when the chemical reaction is written as a dissolution reaction. Most secondary minerals found in soils can be considered sparingly soluble in the presence of water: dissolution does not noticeably change the total mass of the solid phase present. Standard state assumptions concerning the solid phase are thus applicable, and the solubility of the mineral can be described by a solubility product. The solubility product of a mineral, therefore, is a relative index of its potential stability (whether it will dissolve, remain stable, or conditions are favorable for its formation) in the soil environment.

Representative values of $\log K_{so}$ calculated from standard Gibbs free energies are listed in **Table 1**. Negative values of $\log K_{so}$ indicate low solubility in the presence of water. Positive values of $\log K_{so}$ suggest that the corresponding mineral phases are not stable and would not attain equilibrium in the typical aqueous soil environment. This comparison is only qualitative, however, as the dissolution reaction for most silicates and phyllosilicates requires the presence of $H^+$ as well as water, thus the dissolution reaction (and therefore mineral stability) is also a function of pH. Calculated values of $K_{so}$ are also dependent on the degree of uncertainty associated with the standard Gibbs free energies for each of the constituents in the reaction and variations in calculated $\log K_{so}$ values of several whole numbers or more are not uncommon.

### Ion Activity Products

The dissolution of a mineral can be written as the transition from one solid phase to another:

$$\text{mineral A} + \text{water} \rightleftharpoons \text{mineral B} + \text{solution species C} \qquad [4]$$

or the destruction of the original mineral with the stoichiometric (in proportion to the atomic structure of the mineral) release of the constituent ions to solution. The dissolution reaction for the secondary mineral kaolinite can be written as:

$$\begin{aligned} &Al_2Si_2O_5(OH)_4(s) + 6H^+(aq) \\ &\leftrightarrow 2Al^{3+}(aq) + 2H_4SiO_4^0(aq) + H_2O(l) \end{aligned} \qquad [5]$$

where the subscript '(s)' refers to the solid phase, '(aq)' to the ion in solution, and '(l)' for the bulk aqueous phase. The calculated $\log K_{so}$ for this reaction is 5.4 (**Table 1**). Solving for $K_{so}$ yields the following equation:

$$K_{so} = \frac{(Al^{3+}(aq))(H_4SiO_4^0(aq))^2(H^+(aq))^{-6}(H_2O(l))}{(Al_2Si_2O_5(OH)_4(s))} \qquad [6]$$

Under standard conditions, the expressions for the solid-phase kaolinite and the bulk liquid-phase water are defined as being equal to a value of 1, leaving the numerator as an ion activity product (IAP) for the constituent ions in solution:

$$IAP = (Al^{3+}(aq))(H_4SiO_4^0(aq))^2(H^+(aq))^{-6} \qquad [7]$$

When the $IAP > K_{so}$, the solution is said to be supersaturated with respect to the solid phase (conditions favor the precipitation of kaolinite). When the $IAP < K_{so}$, the solution is undersaturated with respect to kaolinite (conditions favor the dissolution of kaolinite). When $IAP = K_{so}$, the solution is considered saturated and at equilibrium (kaolinite should neither dissolve nor precipitate). With tabulated values of solubility products (**Table 1**) and measurements of the dissolved concentrations of ions in soil solutions or streams, soil scientists can predict the possible absence or presence of minerals by calculating IAPs.

### Activity Diagrams

The IAP provides a numerical way to estimate the stability of a mineral in a given soil environment. A visual representation of the relative stability of a mineral can be obtained using an activity diagram. Activity diagrams vary in complexity, depending on the number of mineral phases involved. They are all, however, based on rearrangement of dissolution reactions (e.g., eqn [5]) into a linear form where the $y$- and $x$-axes are represented by constituent ions common to all the mineral phases being addressed. This rearrangement often involves the introduction of $H^+$ into the dissolution reaction by adding the hydrolysis reaction of water. The resulting $K$ values are sometimes designated as $K_{diss}$ (where 'diss' is 'dissolution') in contrast to $K_{so}$ to reflect the mathematical convenience of incorporating $H^+$ into the dissolution reaction.

**Table 1** Calculated solubility products for selected minerals using standard Gibbs free energies of formation[a]

| *Name* | *Chemical formula* | *Dissolution reaction* | *log* $K_{SO}$ |
|---|---|---|---|
| *Carbonates* | | | |
| Calcite | $CaCO_3$ | $Ca^{2+} + CO_3^{2-}$ | −8.3 |
| Dolomite | $MgCa(CO_3)_2$ | $Mg^{2+} + Ca^{2+} + 2CO_3^{2-}$ | −17.6 |
| Siderite | $FeCO_3$ | $Fe^{2+} + CO_3^{2-}$ | −10.1 |
| Rhodochrosite | $MnCO_3$ | $Mn^{2+} + CO_3^{2-}$ | −9.9 |
| *Sulfates* | | | |
| Gypsum | $CaSO_4 \cdot 2H_2O$ | $Ca^{2+} + SO_4^{2-} + 2H_2O$ | −4.6 |
| Jarosite | $KFe_3(SO_4)_2(OH)_6$ | $K^+ + 3Fe^{3+} + 2SO_4^{2-} + 6OH^-$ | −95.3 |
| Alunite | $KAl_3(SO_4)_2(OH)_6$ | $K^+ + 3Al^{3+} + 2SO_4^{2-} + 6OH$ | −79.9 |
| *Phosphates* | | | |
| Variscite | $AlPO_4 \cdot 2H_2O$ | $Al^{3+} + PO_4^{3-} + 2H_2O$ | −21.8 |
| Strengite | $FePO_4 \cdot 2H_2O$ | $Fe^{3+} + PO_4^{3-} + 2H_2O$ | −26.1 |
| Vivianite | $Fe_3(PO_4)_2 \cdot 8H_2O$ | $3Fe^{2+} + 2PO_4^{3-} + 8H_2O$ | 35.5 |
| K-Taranakite | $H_6K_3Al_5(PO_4)_8 \cdot 18H_2O$ | $6H^+ + 3K^+ + 5Al^{3+} + 8PO_4^{3-} + 18H_2O$ | −176.5 |
| Octacalcium phosphate | $Ca_8H_2(PO_4)_6 \cdot 5H_2O$ | $8Ca^{2+} + 2H^+ + 6PO_4^{3-} + 5H_2O$ | −92.6 |
| Hydroxyapatite | $Ca_{10}(OH)_2(PO_4)_6$ | $10Ca^{2+} + 2OH^- + 6PO_4^{3-}$ | −114.9 |
| Fluorapatite | $Ca_{10}F_2(PO_4)_6$ | $10Ca^{2+} + 2F^- + 6PO_4^{3-}$ | −314.3 |
| *Sulfides* | | | |
| Alabanite | $MnS$ | $Mn^{2+} + S^{2-}$ | −12.7 |
| Sphalerite | $\alpha$-$ZnS$ | $Zn^{2+} + S^{2-}$ | −24.4 |
| Pyrite | $FeS_2$ | $Fe^{2+} + S_2^{2-}$ | −26.6 |
| Covellite | $CuS$ | $Cu^{2+} + S^{2-}$ | −35.6 |
| Black-metacinnibar | $\beta$-$HgS$ | $Hg^{2+} + S^{2-}$ | −51.0 |
| *Oxides/hydrous oxides* | | | |
| Quartz | $\alpha$-$SiO_2$[a] | $H_4SiO_4^0$ | −4.0 |
| Quartz (amorphous phase) | $SiO_2$[a] | $H_4SiO_4^0$ | −2.7 |
| Gibbsite | $\gamma$-$Al(OH)_3$ | $Al^{3+} + 3OH^-$ | −33.5 |
| Gibbsite (amorphous phase) | $Al(OH)_3$ | $Al^{3+} + 3OH^-$ | −31.9 |
| Boehmite | $\gamma$-$AlOOH$[b] | $Al^{3+} + 3OH^-$ | −33.4 |
| Amorphous iron hydroxide | $Fe(OH)_3$ | $Fe^{3+} + 3OH^-$ | −37.9 |
| Hematite | $\alpha$-$Fe_2O_3$[b] | $Fe^{3+} + 3OH$ | −41.4 |
| Goethite | $\alpha$-$FeOOH$[b] | $Fe^{3+} + 3OH^-$ | −41.5 |
| *Silicates*[c] | | | |
| Olivine | $Mg_{1.6}Fe_{0.4}SiO_4$ | $1.6Mg^{2+} + 0.4Fe^{2+} + H_4SiO_4^0$ | 25.8 |
| Diopside | $CaMg(SiO_3)_2$ | $Mg^{2+} + Ca^{2+} + 2H_4SiO_4^0$ | 20.9 |
| Pyroxene | $CaAl_2SiO_6$ | $Ca^{2+} + 2Al^{3+} + H_4SiO_4^0 + 2H_2O$ | 34.8 |
| Albite | $NaAl_2Si_3O_8$ | $Na^+ + Al^{2+} + 3H_4SiO_4^0$ | 3.6 |
| Microcline | $KAlSi_3O_8$ | $K^+ + 2Al^{3+} + 3H_4SiO_4^0$ | 1.0 |
| Anorthite | $CaAl_2Si_2O_8$ | $Ca^{2+} + 2Al^{3+} + 2H_4SiO_4^0$ | 23.0 |
| *Phyllosilicates*[c] | | | |
| Muscovite | $KAl_2(AlSi_3O_{10})(OH)_2$ | $K^+ + 3Al^{3+} + 3H_4SiO_4^0$ | 13.3 |
| Vermiculite | (see below)[d] | $2.71Mg^{2+} + 0.02Fe^{2+} + 0.46Fe^{3+} + 0.06Ca^{2+} + 0.1K^+$<br>$+ 1.14Al^{3+} + 2.91H_4SiO_4 + 0.36H_2O$ | 38.1 |
| Montmorillonite | (see below)[e] | $0.49Mg^{2+} + 1.71Al^{3+} + 0.22Fe^{3+} + 3.81H_4SiO_4^0$ | 2.7 |
| Kaolinite | $Al_2Si_2O_5(OH)_4$ | $2Al^{3+} + 2H_4SiO_4^0 + H_2O$ | 5.4 |

[a] Standard Gibbs free energies of formation taken from Lindsay WL (1979) *Chemical Equilibria in Soils*. New York: John Wiley; Stumm W and Morgan JJ (1996) *Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters*, 3rd edn. New York: John Wiley.

[b] Water added as a reactant (e.g., $SiO_2 + 2H_2O = H_4SiO_4^0$).

[c] Water and $H^+$ added as reactants.

[d] $Mg_{2.71}Fe(II)_{0.02}Fe(III)_{0.46}Ca_{0.06}K_{0.1}(Si_{2.91}Al_{1.14})O_{10}(OH)_2$.

[e] $Mg_{0.2}(Si_{3.81}Al_{1.71}Fe(III)_{0.22}Mg_{0.29})O_{10}(OH)_2$.

Simple activity diagrams usually include a series of linear lines, reflecting differences in the solid phase under consideration (**Figure 1**). These differences are often associated with variations in solubility due to assumed crystallinity or size of the solid phase. Experimental data points plotted on such graphs which fall along the linear lines are taken to reflect equilibrium with the solid phase. Data points falling above or below the line indicate undersaturation or supersaturation of the soil solution with respect to the mineral phase (Figure 1). More detailed activity diagrams can also be used to delineate regions of stability among several mineral phases (**Figure 2**). Experimental data can be plotted on such plots, but the intersection of the

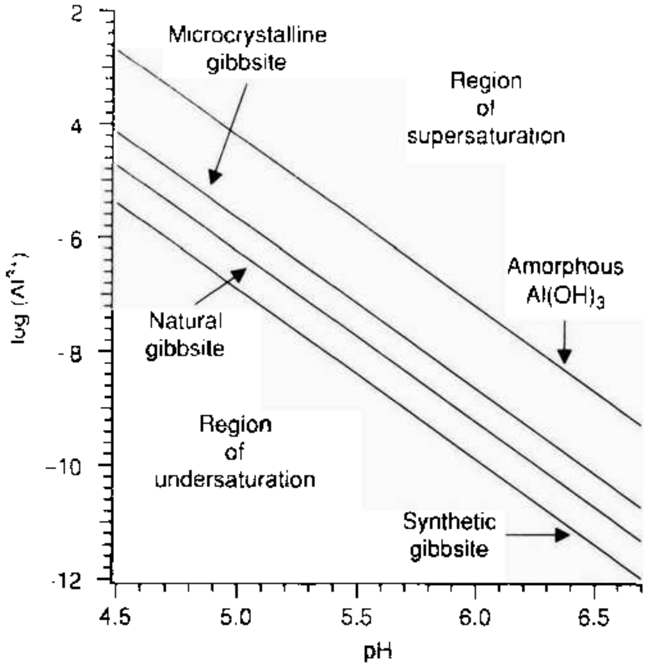


**Figure 1** Activity–ratio diagram for synthetic gibbsite, natural gibbsite, microcrystalline gibbsite, and amorphous aluminum hydroxide. Reproduced with permission from Johnson NM, Driscoll CT, Eaton JS, Likens GE, and McDowell WH (1981) 'Acid Rain', Dissolved Aluminum and Chemical Weathering at the Hubbard Brook Experimental Forest, New Hampshire. *Geochimica et Cosmochimica Acta* 45(9): 1421–1437.

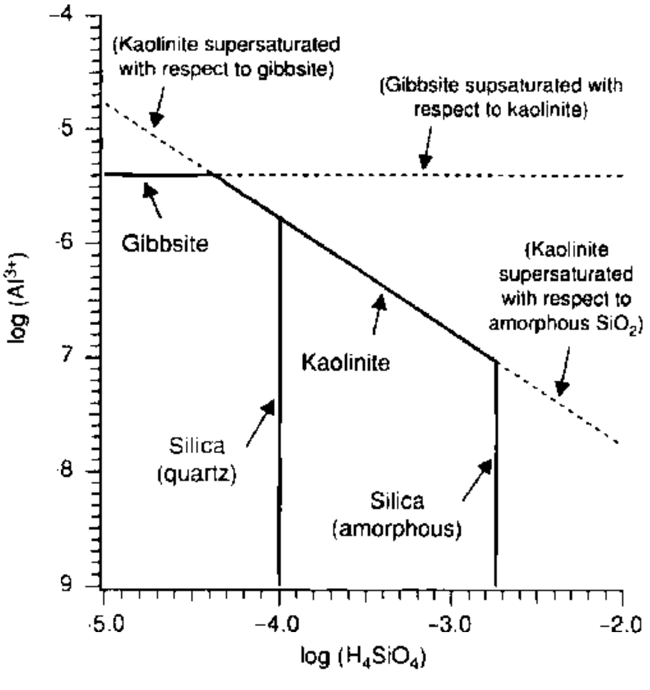


**Figure 2** Activity–ratio diagram for the gibbsite, kaolinite, silica mineral assemblage for pH 4.5. Dashed line representing supersaturation of kaolinite with respect to quartz is not shown. Adapted from Robarge WP (1999) Precipitation/dissolution reactions in soils. In: Sparks DL (ed.) *Soil Physical Chemistry*, 2nd edn. Boca Raton, FL: CRC Press, with permission.

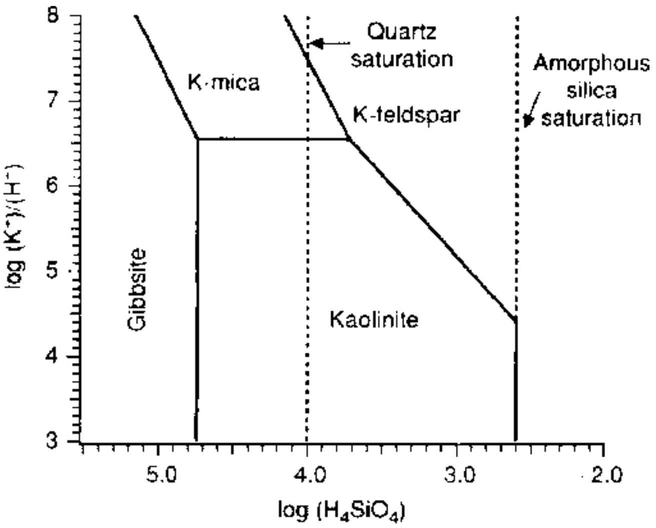


**Figure 3** Phase diagram for gibbsite, kaolinite, K-feldspar, K-mica, silica mineral assemblage. Adapted from Garrels RM and Christ CL (1965) *Solutions, Minerals and Equilibria*. New York: Harper and Row, with permission.

various linear representations of mineral solubility also provide predictive information concerning the stability of minerals in soil systems. For example, kaolinite is the most stable phase compared with gibbsite in soil environments where the log ($H_4SiO_4^0$) in soil solution is more than −4.5. The region of stability is increased if the activity of $H_4SiO_4^0$ in solution is controlled by amorphous silica in contrast to quartz (**Figure 2**).

Dissolution reactions as depicted in eqn [4] can also be displayed visually through a variation in activity plots known as a phase diagram (also known as an activity–activity diagram) (**Figure 3**). Linear lines in such plots represent equilibrium between two solid phases. Departure away from the linear lines indicates that only one of the mineral phases in question would be stable for the values of the constituent ions plotted along the axes.

Understanding the dissolution of minerals as a macroscale process relies on the assumptions inherent in thermodynamics and congruent dissolution of the solid phase. Congruent dissolution requires that the rates of release of the chemical species represented by the formula weight for a mineral are equal. In other words, it is assumed that the various chemical species that comprise a mineral enter the surrounding soil solution at the same time during the dissolution reaction. For many carbonate, sulfate, phosphate, sulfide, and even silicate-based minerals, this assumption is probably valid. With increasing complexity in mineral structure, however, the assumption of congruent dissolution becomes questionable. For example, the congruent dissolution of montmorillonite or muscovite (**Table 1**) involves the concurrent release to soil solution of up to five different ionic species. Given the differences in chemical bonding between these constituent species within the mineral structure, it is highly unlikely that dissolution of these minerals proceeds in a congruent fashion. It is more likely that the dissolution reaction is incongruent: the ratio of ions released to solution is not the same as that found in the bulk phase of the mineral. The extent to which incongruent dissolution invalidates the macroscale approach to understanding mineral solubility in soil systems is not easily determined. In general the macroscale approach is valid for predicting long-term, steady-state conditions or delineating boundaries for possible reactions.

There are also limits to using the condition of IAP > $K_{so}$ (supersaturation) to predict precipitation reactions in the soil environment. Minerals (including soil clay minerals) represent the selective accumulation of two or more constituent ions into an organized solid matrix. This can occur on the bulk soil solution (homogeneous precipitation) or on the surfaces of other soil solids that act as a catalyst to promote the growth of a new solid phase (heterogeneous precipitation). It is possible for many solids with relatively simple structures (carbonates, sulfates, sulfides, and some phosphates and oxides) to form in soils over relatively short time periods. For example, soil horizons that experience periodic wetting and drying due to a variable water table lose and gain color due to the dissolution and then subsequent precipitation of iron oxides. In more arid climates, it is not uncommon to find layers of carbonates and even sulfates that precipitate from the soil solution as the wetting front from a rain event slows and finally stops as it moves through the soil. Formation of more complex solids, especially silicates, phyllosilicates (which include most soil clay minerals) and apatites, however, should not be assumed to occur readily in the bulk soil environment. These solids generally require unique conditions within the soil that favor their formation, or occur as heterogeneous precipitation reactions on the surfaces of or near other minerals that are undergoing dissolution. Even when such solids form, they are often amorphous in nature, meaning that they have the expected chemical composition for the solid phase, but require more time to mature and develop the internal crystalline, well-ordered structure characteristic of a given mineral. Amorphous phases of a mineral are more soluble than its crystalline counterpart and may not survive long enough to develop into a stable separate phase within the soil. As with dissolution reactions, predicting the rates of precipitation in specific soil environments is best viewed as a microscale process.

## Microscale Processes

### Precipitation

Formation of a new solid phase when viewed as a microscale process requires a series of steps that must occur before a new solid can survive and be characterized as having a specific $K_{so}$ value. The first requirement is that supersaturation (IAP > $K_{so}$) conditions must exist whether the reaction results in homogeneous or heterogeneous precipitation. Supersaturation conditions are necessary because they promote the formation of nuclei as a result of the collisions in solution (or on a foreign surface) of the constituent ions that eventually could become the new solid phase. Once formed, the nuclei must resist the tendency to dissolve and must strive to grow in size by accumulating other constituent ions and transforming into crystallites. The formation of crystallites is a critical step, because they are large enough in size to begin to lower the supersaturation of the solution and lower the chances of other nuclei forming. This effectively promotes their own survival, because it lowers the competition for the remaining constituent ions in solution. The final step is the growth of the crystallites into crystals through the process called Ostwald ripening: an increase in the size of a selected number of crystallites to form crystals by the net transfer of material from smaller nuclei and crystallites via dissolution and precipitation. The larger crystals being formed in fact can be defined by an IAP and continue to grow until ideally IAP = $K_{so}$ (saturation).

Which crystallites survive the transformation into crystals and the eventual formation of a new solid phase can be predicted by consideration of two sources of energy which influence the growth of crystals in solution or on the surface of a foreign particle. The first source of energy is the formation of the chemical bond between the constituent ions that will compose the solid (here designated as $\Delta G_{bulk}$). The second energy term is related to the amount of work for the nuclei actually to be present within the solution (here designated as $\Delta G_s$). In other words, the act of forming nuclei must result in the displacement of the surrounding water molecules to create a volume in solution where the nuclei can physically exist. How well a nucleus or crystallite can do this is related to its size and its ability to interact with the surrounding water molecules. This latter ability is referred to as interfacial tension. The summation of the two energy terms equals to amount of energy available for the formation of a stable nuclei in solution (or on a foreign surface):

$$\Delta G^0_{r(nucleus)} = \Delta G_{bulk} + \Delta G_s \qquad [8]$$

For any nucleus, $\Delta G_{\text{bulk}}$ is related to the number of molecular units in the nuclei or crystallites and the degree of supersaturation in solution. As long as $IAP > K_{so}$, $\Delta G_{\text{bulk}}$ is negative, which promotes nuclei or crystallite growth. $\Delta G_s$ is a function of the interfacial tension between the surrounding water molecules and the surface area of the nuclei or crystallites. This term is often positive in value, meaning that it requires energy to make room for the particles to grow in solution.

Mathematical expressions for the two energy terms in eqn [8] can be derived but are a function of the geometry selected for the nuclei and will not be discussed here. Equation [8] does, however, provide a qualitative explanation for the results of precipitation and dissolution reactions often observed in the soil environment. For example, inclusion of $\Delta G_s$ illustrates why dissolution rates are often a function of particle size, i.e., smaller particles dissolve faster than larger particles. Small particles, typically less than 1 $\mu$m in size, have significant positive $\Delta G_s$ values because of the combination of their relatively large surface area to volume ratios and interfacial tension. Small particles therefore exhibit a different $K_{so}$ value than their larger counterparts, such that a solution may appear supersaturated with respect to a solid phase when in fact that solid may actually be present but as very small particles. The presence of the $\Delta G_s$ in eqn [8] also explains why amorphous phases may be present in soils even though they are more soluble than their more crystalline counterparts (e.g., compare $K_{so}$ values for amorphous silica with quartz in **Table 1**). For two solids having the same chemical composition, the solid that is more crystalline in nature should have a larger negative $\Delta G_{\text{bulk}}$ value. However, if the amorphous-phase nuclei have a lower interfacial tension than the crystalline phase, $\Delta G_s$ will have smaller positive value such that the sum of the two terms actually results in a $\Delta G^0_{r(\text{nucleus})}$ value that favors formation of the amorphous phase. Experimental evidence suggests that formation of amorphous phases is almost always preferred over crystalline phases under most conditions that favor precipitation. More crystalline phases usually develop later either through internal rearrangement of the amorphous-phase solid, or through the dissolution of the amorphous phase and subsequent growth of the crystalline phase. In either case, it should not always be assumed that the crystalline phases of minerals will dominate soil environments simply because of differences in $K_{so}$ values.

Lastly, the role of interfacial tension also offers an explanation of why heterogeneous precipitation occurs in the presence of foreign surfaces, often at supersaturated conditions below those necessary to promote homogeneous precipitation. If the interfacial tension between the constituent ions of a nuclei or crystallite is lower when interacting with a surface than with bulk water, the formation of a nucleus on the surface of a foreign particle would reduce the surface area exposed to solution, in effect lowering the value of $\Delta G_s$ in eqn [8]. This yields a $\Delta G^0_{r(\text{nucleus})}$ value that favors nuclei attempting to grow on the surface of a foreign particle in contrast to those attempting to survive in the bulk solution. Thus solid phases that nucleate and grow on a foreign surface perform less work in order to survive. Depending on the surface of the foreign particle (i.e., the interfacial tension between the foreign particle and the nuclei of the new solid phase), the growth of the nuclei into crystallites may actually proceed along the surface of the foreign particle before developing as a separate, three-dimensional phase extending away from the foreign surface. This affords one explanation for the development of coatings, such as iron oxide coatings on clay minerals, which are often observed in soils.

### Dissolution

The dissolution of a mineral can be defined as the detachment of constituent ions from the surface of a mineral and their subsequent transport to the bulk solution. It is now generally accepted that the rate of dissolution of minerals in soils is controlled by chemical reactions along the mineral surfaces at distinct surface features that can include dislocations, etch pits, and micropores. Dislocations represent areas in the mineral structure where the normal growth of the mineral has been disrupted or caused to alter in some fashion. Etch pits and micropores form in the surface of minerals and are visible as holes in the mineral surface when viewed using devices such as a scanning electron microscope. The presence of surface features such as dislocations, etch pits, and micropores is taken as an indication of a population of surface reactive sites where dissolution reactions are favored above those possible along the total surface area of a mineral.

The dissolution reaction at the surface reactive sites is considered a two-step process. The first step involves the relatively rapid formation of a surface species as the result of the reaction with entities in soil solution such $H^+$, $OH^-$, $H_2O$ molecules, and inorganic and organic ligands. The second step, and the one that is considered rate-limiting, involves the transformation of the surface species into an activated complex with the eventual release of constituent ions to the bulk solution. Dissolution at the microscale therefore can be considered a function of the reactive surface area of the mineral, the nature of the inherent chemical bonding structure that comprises the mineral, and the presence of various chemical species in the soil solution. Thus the dissolution rate of a mineral is not a fixed quantity but dependent upon a number of variables, which may change with time due to changes in the physical size of the mineral itself, as well as changes in the surrounding soil solution.

**Rate law** A general form of a rate law for chemical dissolution at heterogeneous mineral surfaces has been formulated to account for the various factors important in the solubility of minerals. A simplified version of this rate law can be expressed as:

$$\text{Rate} = k_o \cdot S_{\min} \cdot \exp^{(-E_a/RT)} \cdot \int\left(\text{pH} \cdot a_i^{N_i}\right) \cdot \int(\text{IAP}) \qquad [9]$$

where $k_o$ is the intensive rate constant (moles per meter squared per second), $S_{\min}$ is the reactive surface area of the mineral, and $E_a$ is the overall activation energy for the reaction. The possible effects of pH, ionic strength, and other ligands, reductants, oxidants, or sorbed ions ($\int(\text{pH} \cdot a_i^{N_i})$) are written as products to emphasize that these effects can be studied separately or in concert with one another, but that one cannot ignore the influence of any of these variables on the rate of dissolution. The final term, $\int(\text{IAP})$, is added to account for the fact that rate of dissolution is also a function of the degree of supersaturation or undersaturation of the soil solution. Technically eqn [9] can represent the rate of dissolution or precipitation, with rate = 0 at saturation ($\text{IAP} = K_{so}$).

The range in dissolution rates possible for various minerals, primarily from laboratory dissolution experiments, is shown in **Table 2**. Among the silicates and the phyllosilicates, those minerals referred to as framework silicates (quartz, albite, microcline, anorthite) and the sheet silicates (kaolinite, montmorillonite, muscovite) are more insoluble than the chain silicates (diopside) or the ring and orthosilicates (olivine). Assuming a mineral particle with an initial radius of 1 mm, minerals with log rate values of less than −13 could have lifetimes of more than 1 000 000 years, while log rate values of −8 to −5 would have lifetimes of less than one hundred to several thousands of years. Such predictions are generally valid as compared to actual field observations across many different types of geologic and hydrological systems. However, dissolution rates and mineral stability within specific local environments are dictated by the variables included in eqn [9]. Laboratory dissolution rates for calcite and dolomite (**Table 2**) suggest that these minerals are not stable across geologic time spans, yet under proper conditions these minerals are found naturally in soils.

**Table 2** Rates of dissolution for selected minerals

| *Mineral* | *Log rate ($mol\,m^{-2}\,s^{-1}$)* | *Conditions* |
|---|---|---|
| *Silicates* | | |
| Quartz | −13.2 | 40°C, $I = 0$[a] |
| | −11.2 | 25°C, pH 5 |
| | −10.7 | 70°C, pH 5.4, 0.25 mol l$^{-1}$ Na$^+$ |
| | −9.9 | 70°C, pH 6, $I = 0.35$ |
| Microcline | −12.5 | 25°C, pH 5 |
| | −11.8 | 25°C, pH 3, $I = 0$ |
| | −11.1 | 25°C, pH 3.1, 1 mmol l$^{-1}$ oxalic acid |
| Diopside | −11.3 | 25°C, pH 3.6 |
| | −10.2 | 25°C, pH 5 |
| Albite | −12.3 | 25°C, pH 5 |
| | −11.3 | 25°C, pH 5, oxalic acid |
| | −11.6 | 25°C, pH 3, oxalic acid |
| Anorthite | −12.0 | 25°C, pH 6.2, $I = 0$ |
| | −8.6 | 25°C, pH 5 |
| Olivine | −5.7 | 25°C, pH 3–5 |
| *Phyllosilicates* | | |
| Kaolinite | −13.8 | 25°C, pH 5 |
| | −13.0 | 25°C, pH 8 |
| Montmorillonite | −14.1 | 20°C, pH 8 |
| | −13.4 | 25°C, pH 5 |
| | −12.5 | 25°C, pH 3 |
| Muscovite | −13.0 | 25°C, pH 5 |
| | −12.5 | 25°C, pH 3 |
| *Other* | | |
| Gibbsite | −12.5 | 25°C, pH 5 |
| | −11.3 | 25°C, pH 3.1, $I = 0.0001$ |
| Calcite | −5.6 | pH 5, low $P_{CO_2}$ |
| | −4.5 | pH 3, low $P_{CO_2}$ |
| Dolomite | −6.5 | pH 5 |
| | −5.0 | pH 3 |

Data taken from Lasaga AC (1998) *Kinetic Theory in the Earth Sciences*. Princeton, NJ: Princeton University Press; Stumm W (1990) *Aquatic Chemical Kinetics: Reaction Rates of Processes in Natural Waters*. New York: John Wiley; and White AF and Brantley SL (eds) (1995) *Reviews in Mineralogy*, vol. 31. Washington, DC: Mineralogical Society of America.

[a] $I$, ionic strength.

## Summary

The dissolution of minerals is a vital geochemical process that has led to the development of life on Earth as we know it. The presence of minerals in soils and underlying soil parent material is critical to the sustainability of soils, especially forest soils, where fertilization to replace mineral elements such as calcium or magnesium lost by timber harvesting or enhanced leaching due to acidic deposition is currently cost-prohibitive. Intensive investigation into the solubility of minerals continues using both the macroscale and microscale approaches. New efforts are now being directed at incorporating *ab initio* quantum-mechanical calculations of mineral surface reactions to gain a better understanding of heterogeneous kinetics. This work is made possible by the rapid changes in computer technology, hardware, and software that have occurred since the 1980s. Just how molecular modeling will impact our understanding of the solubility of minerals and the precipitation of new solid phases, especially for larger and more complex chemical systems, is uncertain, but the merger of computational and experimental science has begun and will continue for the foreseeable future.

*See also:* **Acid Rain and Soil Acidification**; **Agroforestry**; **Chemical Equilibria**; **Clay Minerals**; **Factors of Soil Formation:** Biota; Climate; Human Impacts; Parent Material; Time; **Fertility**; **Metal Oxides**; **Micronutrients**; **Minerals, Primary**; **Productivity**

## Further Reading

Bragg L, Claringbull GF, and Taylor WH (1965) *Crystal Structures of Minerals: The Crystalline State*, vol. IV. Ithaca, NY: Cornell University Press.

Cygan RT (2001) Molecular modeling in mineralogy and geochemistry. In: Cygan RT and Kubicki JD (eds) *Molecular Modeling Theory: Applications in the Geosciences*. Reviews in Mineralogy and Geochemistry,vol. 42, pp. 1–35. Washington, DC: Mineralogical Society of America.

Dixon JB and Weed SB (eds) (1989) *Minerals in Soil Environments*, 2nd edn. Soil Science Society of America Book Series No. 1. Madison, WI: Soil Science Society of America.

Garrels RM and Christ CL (1965) *Solutions, Minerals and Equilibria*. New York: Harper and Row.

Lasaga AC (1995) Fundamental approaches in describing mineral dissolution and precipitation rates. In: White AF and Brantley SL (eds) *Chemical Weathering Rates of Silicate Minerals*. Reviews in Mineralogy, vol. 31, pp. 23–86. Washington, DC: The Mineralogical Society of America.

Lasaga AC (1998) *Kinetic Theory in the Earth Sciences*. Princeton, NJ: Princeton University Press.

Lasaga AC and Gibbs GV (1990) *Ab initio* quantum-mechanical calculations of surface reactions – a new era? In: Stumm W (ed.) *Aquatic Chemical Kinetics: Reaction Rates of Processes in Natural Waters*.New York: John Wiley.

Lindsay WL (1979) *Chemical Equilibria in Soils*. New York: John Wiley.

Robarge WP (1999) Precipitation dissolution reactions in soils. In: Sparks DL (ed.) *Soil Physical Chemistry*, 2nd edn. Boca Raton, FL: CRC Press.

Sposito G (1981) *The Thermodynamics of Soil Solutions*. London, UK: Oxford University Press.

Stumm W (ed.) (1990) *Aquatic Chemical Kinetics: Reaction Rates of Processes in Natural Waters*. New York: John Wiley.

Stumm W and Morgan JJ (1996) *Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters*, 3rd edn. New York: John Wiley.

White AF and Brantley SL (1995) Chemical weathering rates of silicate minerals: an overview. In: White AF and Brantley SL (eds) *Chemical Weathering Rates of Silicate Minerals*. Reviews in Mineralogy, vol. 31, pp. 1–22. Washington, DC: The Mineralogical Society of America.

## Precision Agriculture *See* Site-Specific Soil Management

## Preferential Flow *See* Unstable Flow; Macropores and Macropore Flow, Kinematic Wave Approach

# PRODUCTIVITY

**D L Karlen**, USDA Agricultural Research Service, Ames, IA, USA

## Introduction

Soil productivity is defined as the capacity of a soil to produce a certain yield of agricultural crops or other plants using a defined set of management practices. It encompasses soil fertility plus all of the other factors affecting plant growth and development, even though it is impossible to explain all of the potential interactions. Soil productivity is most accurately defined in terms of a specific kind of soil, for a specific crop or crop sequence, and with regard to a specific set of management practices. As a result of all these constraints, the concept is perhaps more accurately defined as an economic relationship than one of soil science.

Soil productivity is generally measured by comparing inputs with outputs. For agronomic situations, 'output' generally means crop yield, while 'inputs' usually refer to water, plant nutrients, genetic resources, pest control, or labor. Therefore, for a specific input (e.g., water or plant nutrients) productivity is sometimes calculated as efficiency (i.e., water-use efficiency or nutrient-use efficiency). With regard to sustainability, unintended off-site effects such as runoff, erosion, or leaching losses are also recognized as cropping system outputs that must be considered when evaluating productivity with regard to the social acceptability. The relationships between productivity and economic return are very evident at the farm level, while, from a global perspective, productivity is strongly associated with food sufficiency and sustainable food, feed, and fiber production systems. These multiple scales at which productivity can be evaluated mean that economically beneficial inputs at the field scale may or may not be environmentally and socially acceptable when evaluated at the whole-farm, watershed, regional, national, or international scale.

Productivity is also dependent upon the physical, chemical, and biological characteristics that for each soil reflect numerous interactions among the five basic soil-forming factors (parent material, climate – especially moisture and temperature effects – macro- and microorganism species and populations, topography, and time). Therefore, even though soil productivity *per se* is not an inherent soil characteristic, it will be different for every soil and is therefore one of the many characteristics used to classify more than 18 000 soil series in the USA alone.

Productivity is also affected by what some refer to as a sixth soil-forming factor – humankind. Others prefer to classify this response as 'dynamic soil properties,' since the changes are often controllable and reflect the soil-management decisions or practices that are imposed by humankind on the soil resource. Processes affecting the dynamic soil properties include erosion, nutrient mining (where removal is greater than the inputs), loss of organic matter, acidification, salinization, water logging, and compaction. Productive soils generally exhibit resistance or resilience, two properties that enable the resource to withstand degradation by natural (e.g., drought or flooding) or human-induced (e.g., tillage) stress without losing their ability to provide critical functions.

## Productivity Effects of the Inherent Soil-Forming Factors

The specific effect of each soil-forming factor on productivity will differ among soils, but for high agronomic productivity the soil must provide physical support, plant-available water, air for respiration, and the essential nutrients (C, H, O, N, P, K, Ca, Mg, S, B, Cu, Fe, Mn, and Zn). Parent material influences the availability of essential plant nutrients (those required to complete the life cycle) through its mineral composition and predominant particle size (i.e., sand, silt, or clay). Particle size is important because it influences soil texture as well as the type and rate of chemical reactions that occur within a specific soil. The process through which parent material was derived (e.g., volcanic activity, sedimentation, residual weathering of rocks) or transported to various sites (e.g., water, wind, or glaciers) affects productivity, primarily through soil texture and landscape topography. Climate, especially precipitation and temperature, influences inherent soil productivity through its effect on weathering of the primary minerals, erosion by wind or water, and the native vegetation (i.e., forest versus prairie). Chemical reaction rates for processes such as solution, hydration, and leaching are also regulated by water and temperature regimes. The primary type and quantity of vegetation found at various landscape positions, the type and amount of microbial growth, the type of enzymes produced, the number and type of soil animals (e.g., grazers versus burrowers), the predominant soil biota, and amount of organic material added to the soil from above- and belowground plant parts each year are also influenced by climate.

Topography influences soil productivity primarily through its effects on soil water, temperature, and erosion. Soils on sloping land, especially those with a medium or fine texture, will almost always capture less precipitation, irrigation, or runoff water and are more likely to be degraded through erosion than soils with similar vegetation on more level areas. Soils in low-lying or depression areas can be highly productive if rainfall or irrigation is evenly distributed, but they can be nonproductive if water accumulates, creating aeration problems or increasing leaching. Slope orientation and elevation affect productivity primarily through their effects on soil-forming processes as related to soil temperature and water content. Time of formation exerts its influence through the degree of soil development by processes of eluviation (loss of material) and illuviation (accumulation of material) at different positions within the soil profile.

Vegetation influences productivity through photosynthesis, which determines the quantity, quality, and spatial distribution (above- or belowground) of carbon input each year. This annual carbon input as soil organic matter affects nutrient cycling, soil aggregation, and soil structure. In this context, vegetation quality generally refers to carbon-to-nitrogen (C:N) or lignin-to-nitrogen ratios. These quality factors affect the food source for soil biota and hence nutrient cycling. Aggregation, the process through which primary soil particles (sand, silt, and clay) are bound together by natural forces and organic compounds derived from root exudates and microbial processes is important because, if soil aggregates are stable, rainfall or irrigation water are more likely to enter (infiltrate) the soil than to runoff. Aggregates also provide the building blocks for soil structure which influences productivity through its effect on root proliferation, aeration, and the volume of soil available for water retention. Vegetation also provides the fuel for naturally occurring and intentional fires, a natural ecosystem process that plays an exceptionally important role in nutrient cycling and reseeding of grasslands and coniferous forests. Periodic fires also prevent encroachment of trees into grassland areas, and also, in arid areas, they prevent shrubs or other woody species such as honey mesquite (*Prosopis glandulosa*) from becoming established.

## Humankind's Impact on Productivity

Humankind can have a tremendous impact on soil productivity, because with external inputs distinctly different soils can have very similar outputs. For example, two soils that respond similarly with low input (e.g., simply tilling, planting, and harvesting) may respond in very different ways if the inputs include fertilizer, lime, animal manure, or rotations with deeply rooted legumes. These differences illustrate the impact of humankind on dynamic soil properties through their land use and other decision-making processes.

For comparative purposes, the US Department of Agriculture (USDA)–Natural Resources Conservation Service (NRCS) soil survey program defines soil productivity assuming some specified level of management. A productive soil is generally defined as one that will produce good yields of adapted crops in relation to the level of labor, nutrient, and energy inputs. This interpretation of inherent soil productivity provides a basis for comparing crop yield per unit area among soil map units throughout the USA. The caveat for each comparison is that the reported yields are those that can be expected under a high level of management, but the specific management practices are not defined. Also, when a yield is not given for a specific soil map unit within a survey report, it implies that either the soil is not suited for that particular crop or the crop is generally not grown on that particular soil.

Long-term soil management through its effect on the dynamic soil properties has created some general expectations regarding soil productivity when evaluated at regional or major land resource area (MLRA) scales. Soils formed under grassland in humid regions often have higher organic matter content, an increased percentage of water-stable aggregates, and a structure that accommodates rapid plant growth, unlimited root proliferation, and thus increased productivity (i.e., crop yield). These soil characteristics generally occur because grasses in these regions replace most of their roots and top growth each year. As the aboveground plant dies, it falls on the soil surface, where it is decomposed or mixed into the upper part of the soil. Earthworms and other soil organisms incorporate a large portion of this material into the soil. The dead plant material serves as a food source for fungal and bacterial decomposition processes, leading to the formation of soil aggregates, and modifies effects of soil texture with regard to water and air relationships and root penetration. The increased water-stable aggregation, annual proliferation of plant roots, and associated wetting and drying cycles gradually develop a soil structure that is stable and, with good management, generally resists degradation (**Figure 1**).

Grassland soils are often less acid than those developed under forest vegetation. This occurs because grassland soils were generally formed with less annual precipitation which can cause greater leaching

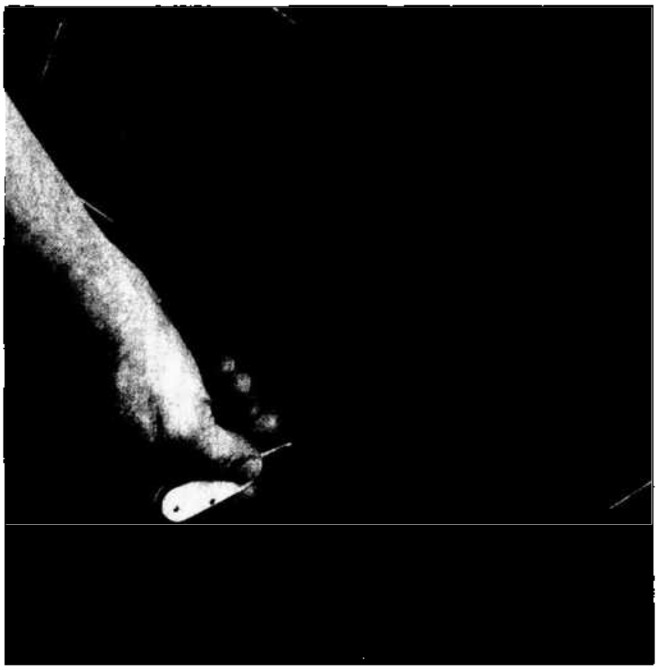


**Figure 1** A well-structured grassland soil showing characteristics associated with productivity.

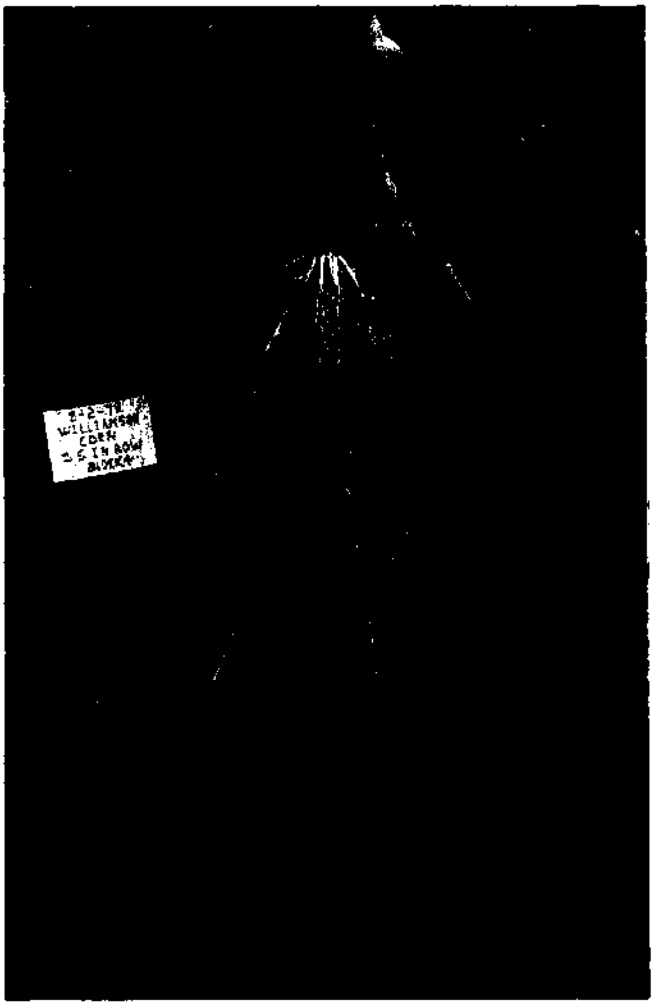


**Figure 2** Corn root proliferation restricted to the area where in-row subsoiling (deep tillage) disrupted the eluviated layer in a southeastern Coastal Plain soil.

of soluble elements such as K, Ca, Mg, and $NO_3$-N. Grassland soils also tend to have a darker color and organic matter accumulation to a greater depth than soils developed under forest. This occurs because trees have larger and longer-living roots than grasses, thus reducing the amount of new organic matter added to the soil each year. Furthermore, most organic matter under forest vegetation is concentrated in the litter layer and remains in the upper few inches of the surface or A horizon. Below that relatively shallow zone in forest-derived soils, the soil organic matter is often more eluviated (leached) than under grassland. The increased amount of organic matter and prolific plant root systems in grassland soils also enable them to be more effective at recycling leached nutrients than forest-derived soils. Nutrient recycling occurs when plant roots absorb ions from the soil solution and transport these materials to aboveground to leaves and other plant parts, which then die and return the nutrients to the soil through litter and root decomposition. As a result of these soil-formation processes, grassland soils (Mollisols) are often considered to have a higher inherent productivity for commonly grown crops than forest soils (Alfisols or Ultisols). The increased aggregation and desirable soil structure enhance air exchange, water entry and retention, and nutrient cycling for subsequent crops.

The eluviated soil layers typically found within forest-derived soils are often more dense and subject to compaction when tilled. This subsurface compaction reduces inherent productivity by decreasing the volume of soil that is available for water retention, air exchange, and plant root development. In the US southeastern Coastal Plain, eluviated horizons are dominated by sand-sized particles (0.05–2.0 mm) and generally have less than 10% by weight of highly weathered kaolinite or similar clay minerals. This results in low plant-available water content, nutrient retention, and organic matter, and thus reduces the absorption of water and nutrients by plants. The most common soil-management practice to enhance the productivity of these soils would be to use some type of deep tillage or mechanical disruption such as in-row subsoiling to loosen the very dense, root-restrictive eluviated soil layers (**Figure 2**). The short-term benefit of deep tillage on several of the dynamic soil properties is often positive. However, tillage also enhances oxidation of the already low levels of soil organic matter and, in the long term, results in even greater compaction due to the wheel traffic.

## Soil Productivity Comparisons

To illustrate soil productivity, selected chemical and physical characteristics, as well as the resultant corn and soybean productivity ratings, for a midwestern Mollisol representing the northern US Corn and Soybean Belt (Nicollet) and an Ultisol from the southeastern Coastal Plain (Norfolk) can be compared (**Table 1**). The soils differ in their productivity ratings for soybean and corn by 290 kg ha$^{-1}$ and 375 kg ha$^{-1}$ when given a high degree of management. Several factors contribute to crop yield differences, but a comparison of the soil-forming factors, dynamic soil properties resulting from management decisions, and the subtle interactions provide a better understanding of soil productivity.

The cation exchange capacity (CEC) for the two soils reflects inherent differences in clay mineralogy and the amount of organic matter. This impacts productivity by influencing the total amount of cations (primarily $Al^{3+}$, $Ca^{2+}$, $H^{+}$, $Mg^{2+}$, $K^{+}$, and $Na^{+}$) that the soil can retain. However, the specific effect of these differences on nutrient availability (*See* **Nutrient Availability**) and crop yield will depend on the total nutrient management program (*See* **Nutrient Management**). The lower pH within the Norfolk soil (i.e., more acid), coupled with the lower CEC, means that smaller but more frequent lime applications may be needed to optimize productivity. The lower amount of organic matter in the Norfolk soil will affect both soil structure and water-retention characteristics. The lower plant-available water capacity (AWC) means that plants growing on the Norfolk soil must receive rainfall or irrigation water more frequently during the growing season than those growing on Nicollet soil. Fortunately, annual rainfall in the southeastern Coastal Plain is greater, but so is the average temperature. Also, because of the restricted rooting associated with compaction, the volume of soil available for each plant to obtain water and nutrients effectively is generally less for the Norfolk soil than for the Nicollet. The effects of compaction and the lower nutrient and water retention can be partially compensated for by including deep tillage as part of the high-level management plan for Norfolk soils. However, use of a subsoiler/ripper requires approximately 16 l ha$^{-1}$ of diesel fuel,

**Table 1** The soil-forming factors, resultant chemical and physical properties, and productivity ratings for a US Midwestern Mollisol and a southeastern Coastal Plain Ultisol

| | *Nicollet* | *Norfolk* |
|---|---|---|
| *Soil-forming factor* | | |
| Parent material | Loamy glacial till, calcareous | Loamy marine sediments, acid |
| Time | Deposited during late Pleistocene | Deposited during Pliocene/Miocene |
| Climate | Mean annual temperature 8°C (mean winter −7°C, summer 22°C); annual precipitation 864 mm | Mean annual temperature 17°C (mean winter 7°C, summer 26°C); annual precipitation 1162 mm |
| Topography | Nearly level to gently sloping glacial till plains and moraines, 1–3% slope | Nearly level to sloping marine terrace uplands, 0–2% slope |
| Natural vegetation | Prairie vegetation characterized by Indian grass, little bluestem, switchgrass, needlegrass, blue grama, sideoats grama, and scattered trees and shrubs of bur oak, juniper, and sumac | Mixed pine and oak forest vegetation characterized by loblolly, longleaf, slash, and shortleaf pines; yellow poplar; sweetgum; and red and white oaks as overstory, and dogwood and farkleberry as understory |
| *Resultant properties* | | |
| Mineralogy | A mixture of minerals in the sand, silt, and clay fractions | Dominated by resistant sand grains and low-activity clays |
| Drainage | Moderately permeable but somewhat poorly drained, seasonal high-water-table 60–100 cm | Moderately permeable but well-drained, seasonal high-water-table 100–180 cm |
| Texture | Loam, clay loam, silty clay loam | Loamy sand, sandy loam |
| Organic matter (g kg$^{-1}$) | 40–80 | 5–20 |
| CEC (cmol kg$^{-1}$) | 25–40 | 1–5 |
| pH | 5.6–7.8 | 4.0–5.5 |
| AWC (cm cm$^{-1}$) | 0.17–0.22 | 0.06–0.11 |
| *Productivity ratings* | | |
| Corn | 7275 kg ha$^{-1}$ | 6900 kg ha$^{-1}$ |
| Soybean | 2960 kg ha$^{-1}$ | 2690 kg ha$^{-1}$ |

CEC, cation exchange capacity. A measure of the ability of the soil to retain positively charged ions such as potassium (K), calcium (Ca), or magnesium (Mg) that are essential for plant growth and development.

AWC, available water capacity. The amount of plant-available water that a soil can retain.

increasing the annual costs of production on that soil and further decreasing economic productivity when computed using output per unit input.

## Productivity Ratings

As illustrated by the comparison of Norfolk and Nicollet soils under natural or nonirrigated conditions, the critical functions that strongly influence the relative productivity rating of a soil for crop production are those needed to provide a nonlimiting rooting environment. For maximum productivity, this environment must be capable of meeting the plant's water and nutrient requirements throughout the growing season. The suitability of a soil for plant root proliferation thus provides a scientific basis for several soil productivity indices. One of the most common is:

$$\mathrm{PI} = \sum_{i-1}^{r} (A_i * C_i * D_i * \mathrm{WF})$$

where PI is the relative productivity index (0–1), $A_i$ is the sufficiency of AWC, $C_i$ is the sufficiency of bulk density, $D_i$ is the sufficiency of pH, WF is a weighting factor representing an idealized plant root distribution, and $r$ is the number of horizons in the rooting depth. Within this index, sufficiency is defined as the pH, $D_i$, and $C_i$ levels that do not restrict plant growth or development throughout the life cycle of the species for which the assessment is being made.

This approach for evaluating productivity uses readily available information, assumes a high level of management, assesses only those changes associated with nonreplaceable inputs, and assesses long-term productive potential without annual fluctuations in crop yield. It was used extensively during the 1980s to project the potential effects of soil erosion and demonstrate the usefulness of soils databases, especially the SOILS-5 and the National Resources Inventory (NRI). One conclusion drawn from a study designed to predict the effects of soil erosion in the US Corn and Soybean Belt was that the nonreplaceable productivity would decline by an average of less than 8% during the next 100 years. The rationale for such a small decline was that the soils in that region are typically deep and fertile, and the cropland tends to be on level-to-gently-rolling terrain. However, for soils in landscape positions where the slope exceeds 6%, the decline in long-term productivity could be much greater. Also, for soils with unfavorable subsoil material (e.g., bedrock or gravel), the loss of productivity could be much greater if erosion was allowed to continue.

## Processes Reducing Soil Productivity and Practices for Restoring It

Soil erosion, nutrient mining (removal or output greater than input), loss of organic matter, acidification, salinization, water logging, and compaction are among the most common causes for reduced soil productivity. As shown by the productivity index, reductions in plant-available water, nutrient retention, and the creation of a less-favorable environment for plant growth are the primary reasons for loss of soil productivity via erosion.

Erosion primarily removes surface soil, which often has the highest biological activity and greatest amount of soil organic matter. This results in an economic loss of essential plant nutrients and a potential environmental loss if those nutrients are transported to water resources, where they can create problems such as algal blooms and eutrophication. Visual indicators that suggest soil erosion may be reducing productivity include changes in soil color, decreasing soil horizon depth, deposition at field boundaries, or, in desert soils, the presence of moss and algal (crypotogam) crusts. The erosion processes occur not only through the effects of wind and water, but also through mechanical processes such as the removal of soil through the harvest of root crops or tillage and cultivation practices that move soil downslope. Management practices that can help minimize erosion and the subsequent productivity loss include modifying the landscape to control runoff amounts and rates (e.g., terracing, shortening slope length and steepness, strip-cropping, planting shelterbelts) or maintaining a protective cover on the soil (e.g., protecting the surface with crop residue with no-tillage, growing cover crops, improving aggregate stability, increasing water infiltration).

Loss of organic matter decreases productivity by reducing air and water movement within the soil, decreasing cation and anion exchange sites, increasing the potential for crusting and subsequent runoff, and reducing the availability of N, P, and S. Soil organic matter can be lost through erosion or decomposition processes. When soils are tilled, organic matter decomposes faster than under no-till conditions because of changes in water, aeration, and temperature. Although the actual amount of organic matter lost after clearing a wooded area or tilling native grassland will vary, a large amount (e.g., 30–50%) will often be lost within the first 10 years.

Practices that hasten the loss of soil organic matter and potentially decrease productivity include (1) decreasing the production of plant materials e.g., replacing perennial vegetation with short-season vegetation, replacing mixed vegetation with monoculture cropping, using cultivars with high harvest indices, increasing the use of bare fallow, or introducing more-aggressive but less-productive plant species); (2) decreasing the supply of organic materials by burning forest, range, or crop residue or through excessive grazing; and (3) increasing the rate of organic matter decomposition through tillage, drainage, or excessive N fertilization. Strategies to enhance soil productivity by increasing soil organic matter include: (1) increasing plant material production with irrigation, optimum fertilization, use of cover crops, reforestation, or restoration of grassland; (2) increasing organic matter supply by protecting from fire (provided the area is not degraded grassland, where fire actually improves productivity by preventing encroachment of shrubs and other woody plant species), applying animal manure or other carbon-rich materials, and controlling insects and rodents; or (3) decreasing rates of decomposition by reducing or eliminating tillage, preventing excessive drainage, or reducing soil temperature by preserving surface vegetative cover.

Acidification or decreases in soil pH often occur because of the excessive application of nitrogen fertilizers. Acidity reduces productivity by decreasing the availability of essential plant nutrients such as Ca, Mg, or P; decreasing CEC through accelerated mineral weathering; and increasing the solubility of Al, which can interfere with plant root development. Acidity also affects the activity of many microorganisms responsible for organic matter decomposition and chemical transformations within the soil. A soil pH between 6.6 and 7.3 is favorable for microbial activities that contribute to the availability of N, S, and P in soils.

Salinization decreases soil productivity by limiting the ability of plants to take up water. It occurs in response to any practice that affects the soil–water balance and allows water-soluble salts to accumulate in the soil. Irrigation, drainage, changes in plant community composition that alter soil cover or rooting characteristics, and summer fallow are some agricultural management practices that can result in salinization if the climate and hydrology conditions are favorable. The process can occur on the soil surface whenever the following conditions occur together: soluble salts, such as sulfates ($SO_4^{-2}$) of Na, Ca, or Mg are present in the soil, the water table is near the soil surface, evaporation rates are high, or annual rainfall is low. In semiarid areas, this process often results in saline seeps on the rims of depressions and edges of drainage-ways, at the base of hill-slopes, and in flat, low-lying areas surrounding sloughs or shallow bodies of water.

Early signs of salinization include: (1) increased soil wetness in semiarid and arid areas to the point that the soil does not support agricultural equipment, (2) growth of salt-tolerant weeds, and (3) irregular patterns of crop growth and lack of plant vigor. Advanced signs include a white crusting on the surface, a broken ring pattern of salts adjacent to a body of water, white spots and streaks in the soil (even where no surface crusting is visible), or the presence of weeds or other naturally growing, salt-tolerant vegetation.

Restoring soil productivity in saline areas is done primarily through water management. Sufficient irrigation water should be applied to leach salts below the root zone, but care should be taken not to overirrigate, especially in areas that do not have adequate natural drainage. It is very important to provide suitable drainage water outlets that do not degrade water bodies or other natural resources into which the drainage water flows. The use of cropping and tillage systems that promote adequate infiltration and permeability can also help restore productivity. This includes building soil organic matter for better soil aggregation and avoiding compaction. Planting salt-tolerant crops that use available soil moisture, removing excess water from recharge areas with actively growing, deep-rooted plant species such as alfalfa, and reducing summer fallow by continuous cropping are among the most effective cropping practices for controlling salinization, restoring productivity, and minimizing negative impacts on the environment.

Soil compaction reduces productivity by restricting rooting depth and the uptake of water and nutrients. Compaction occurs when soil particles are pressed together, increasing the weight of solids per unit volume of soil (bulk density). This reduces pore size and volume, increasing the proportion of water-filled to air-filled pore space. Although in some soils compaction can increase thermal conductivity, causing the soil temperature to be higher than under noncompacted conditions, soils in humid regions with a high water content often have a lower soil temperature, which suppresses the activity of many thermophilic and mesophilic microorganisms that drive soil organic matter decomposition and the subsequent release of nutrients. Compaction also decreases infiltration, increasing the potential for runoff and erosion.

Tilling, harvesting, or grazing when soils are too wet are some of the primary causes for soil compaction. The depth at which compaction occurs, the shrink–swell potential of soil (which is determined by the clay mineralogy), and climate are critical factors determining the persistence of problems and overall impact on productivity. Subsoiling or deep tillage may be necessary to break up compacted layers in some soils. Applying animal manure, compost, newspaper, or woodchips with an appropriate C:N ratio, or municipal sludge that does not contain toxic compounds can help restore productivity by improving soil structure and biological activity. The negative effects of compaction on soil productivity can also be minimized by reducing wheel traffic, the pressure or weight of equipment, and maintaining or increasing soil organic matter.

## Summary

Soil productivity encompasses soil fertility plus the inherent and management-related factors affecting plant growth and development. It is generally measured in terms of inputs versus outputs, which for agronomic situations generally refers to water and/or nutrient input versus crop yield. The critical soil functions influencing productivity within any soil are those that provide physical support, a rooting medium with plant-available water, air for respiration, and essential nutrients. Humankind can also have a tremendous impact on soil productivity through its effects on the dynamic soil properties. Agricultural management decisions regarding tillage, fertilization, crop rotation, irrigation, and drainage are among the practices that can significantly affect soil productivity.

*See also:* **Nutrient Availability; Nutrient Management**

## Further Reading

American Society of Agricultural Engineers (1985) *Eroson and Soil Productivity.* Publication 8-85. St. Joseph, MI: ASAE.

Foth HD and Ellis BG (1997) *Soil Fertility,* 2nd edn. Boca Raton, FL: CRC Press.

Francis CA, Flora CB, and King LD (eds) (1990) *Sustainable Agriculture in Temperate Zones.* New York: John Wiley.

Hatfield JL and Karlen DL (eds) (1994) *Sustainable Agriculture Systems.* Boca Raton, FL: Lewis Publishers/CRC Press.

Hatfield JL and Stewart BA (eds) (1992) *Limitations to Plant Root Growth, Advances in Soil Science,* vol. 19. NY: Springer-Verlag.

Hillel D (1991) *Out of the Earth: Civilization and the Life of the Soil.* Berkley, CA: University of California Press.

Lal R (ed.) (1999) *Soil Quality and Soil Erosion.* Ankeny, IA: Soil Water Conservation Society.

Larson WE, Foster GR, Allmaras RR, and Smith CM (eds) (1990) *Proceedings of Soil Erosion and Producitivity Workshop.* St. Paul, MN: University of Minnesota.

Magdoff F and van Es H (2000) *Building Soils for Better Crops.* Burlington, VT: Sustainable Agriculture Publications.

National Academy of Sciences (1994) *Soil and Water Quality: An Agenda for Agriculture.* Washington, DC: National Academy Press.

**Profile** *See* **Morphology**

# PROTOZOA

**W Foissner**, Universität Salzburg, Salzburg, Austria

## Introduction

Protozoa are unicellular, heterotrophic, eukaryotic organisms comprising four organization types: amebae, flagellates, ciliates, and parasitic sporozoans. About 1600 species, of which some are restricted to certain geographic regions, are known to live in terrestrial habitats; however, at least the same number is still undiscovered. Small body size and the ability to produce protective resting cysts are the main adaptations of protozoa to the peculiarities of the soil environment. Many soil protozoa feed, more or less selectively, on bacteria, while others are omnivorous or highly specialized fungal feeders. Protozoa (active and cystic) inhabit the soil in great numbers, that is, some 10 000–1 000 000 individuals per gram dry mass, and produce many generations annually. They significantly enhance the flow of nutrients and growth of plants and earthworms. Accordingly, they are important soil inhabitants, and studies on their dynamics and community structures thus provide a powerful means for assessing and monitoring changes in biotic and abiotic soil conditions. Unfortunately, methodological and taxonomical problems still limit the use of protozoa as bioindicators in terrestrial environments.

This article covers the diversity, ecology, and bioindicative value of soil protozoa. The knowledge on autotrophic soil protists, which mainly live on the soil surface because they depend on light, is still in its infancy. However, they play an important role, especially as a symbiotic partner of lichens in the crust soils of extreme regions, such as desert and high mountain areas.

## A Brief History of Soil Protozoology

Although there were predecessors, such as Ehrenberg, Goodey, and Russel and Hutchinson, soil protozoology as a science was established by the Austrian–Hungarian naturalist H.R. Francé and the English microbiologist D.W. Cutler in the 1920s. Their monographs showed that unicellular organisms inhabitate the soil in great number and diversity, and thus contribute significantly to soil fertility and sustainability. Francé's booklet saw many editions, the latest in the 1960s.

This first bloom of soil protozoology followed a long period where a few specialists added significantly to the knowledge available, mainly J.F. Darbyshire in the UK, L. Bonnet in France, J.D. Stout in New Zealand, S.S. Bamforth in the USA, and V.F. Nikolyuk in Russia. It was only in the last two decades of the twentieth century that soil protozoology was rejuvenated by the studies and reviews of W. Schönborn in Germany, F. Ekelund and R. Rønn in Denmark, W. Foissner in Austria, and J.F. Darbyshire, who edited the first book devoted entirely to the biology and ecology of soil protozoa. Today, the field is well established but still a domain of a limited number of specialists, while general soil ecologists have problems to incorporate the new knowledge appropriately, in contrast to marine and limnetic plankton researchers, where the 'microbial loop,' that is, the intimate connection of dissolved organic matter, bacteria, and protists, greatly changed views and research strategies.

## Diversity and Biology

### Diversity and Geographic Distribution

About 23 000 free-living, extant protozoa are known. Of these, roughly 1600 were originally described or later reported from terrestrial habitats (**Table 1**). For a long time, the soil protozoan community was considered as an ubiquitous part of limnetic species and some even suggested that nearly the same species of protozoa occur in soil, sewage, and activated sludge. However, more detailed research disproved these findings and showed that they were based on misidentification of species. Today it is widely accepted that the main part (>70%) of the soil protozoan species evolved in and are restricted to terrestrial environments. Refined morphological methods and gene sequence data indicate that a considerable portion of soil and freshwater protists have a restricted geographical distribution (**Figures 1 and 2**), disproving the old hypothesis that, in microorganisms, 'everything is everywhere, the environment selects.' This matter, however, is still under discussion.

**Table 1** Known and proposed diversity of soil protozoa

| *Taxa* | *Known species* | *Proposed species* |
|---|---|---|
| Naked amebae[a] | 60 | 600 |
| Testate amebae | 300 | 500 |
| Flagellates | 270 | 600 |
| Ciliates | 1000 | 2000 |
| Microsporidia and sporozoans | 3 | ? |

[a]Without several hundreds of slime molds.

**Table 1** shows the known and proposed soil protozoan diversity. While numbers of testate amebae and ciliates are supported by detailed data, those of the other groups are very speculative. Likewise, little is known of protozoan parasites. But even the known soil protozoan diversity is high and a major problem in basic and applied soil biology because comprehensive identification literature is lacking and few taxonomic specialists are available. The decreasing interest of the public and science in classical taxonomy, will soon dramatically increase our ignorance. Although molecular methods likely can solve some of the problems, nobody will be available to add names to the sequences. Thus, great efforts are necessary to educate young scientists in classical alpha-taxonomy.

### Morphological and Physiological Adaptations

The soil is a very special environment for protozoa. Three factors are most important for their existence and activity: (1) the structure of the porous space; (2) the astatic water supply; and (3) intraspecific and interspecific interactions. These factors caused two basic adaptations: small body size and the universal capability to produce protective resting cysts. These and some other important adaptations will be briefly described.

**Morphological adaptations** Mean body length, mean body width, and mean biomass are significantly smaller in soil ciliates and testate amebae than in ciliates and testaceans from fresh water. The small size, and volume, is achieved either by an overall reduction in size or by reduction of body width. Many soil ciliates are thigmotactic creepers with a reduced ciliature and a flattened and/or worm-like

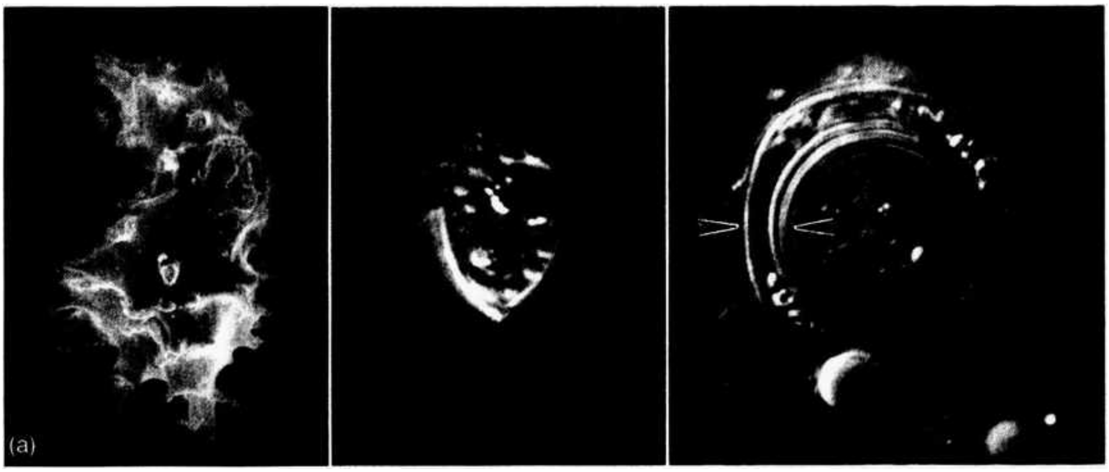


**Figure 1** Small-sized and cystic soil protozoa. (a) A naked ameba about 30 μm long with many tiny pseudopodia which can exploit even minute soil pores for food. (b) A 15 μm long soil flagellate, *Hemimastix amphikineta*, which only occurs in southern terrestrial environments, that is, in Gondwana. Note the two rows of flagella which make the organism look like a ciliate. (c) Resting cyst (diameter about 35 μm) of the ciliate *Exocolpoda augustini*. This ciliate occurs mainly in hot and dry soils, e.g., the Namib desert, and thus has a very thick cyst wall (arrowheads) composed of many membranes. The organism can remain viable in this condition for years.

body (**Figure 3b**). These features and several physiological peculiarities produced a soil ciliate community of unique structure, characterized by a high proportion of colpodids (**Figure 3a** and **3b**) and hypotrichs and a strong underrepresentation of sessile peritrichs and suctorians compared with fresh water ciliate communities. The soil testacean community is characterized by having a high proportion of species with reduced pseudostome size (opening of the shell), a flattened ventral side, and a globular shape (**Figure 2**). Soil naked amebae and flagellates have similar adaptations as ciliates and testate amebae (**Figure 1a** and **1b**): most are very small and the amebae, for instance, can exploit with their tiny and flexible pseudopodia micropores which have a diameter of only 1 μm.

**Physiological adaptations** Few physiological adaptations have been investigated in detail; it is probable that several are still to be discovered. Some physiological adaptations cause conspicuous morphological specializations.

*Food and feeding* Many soil protozoa, especially most naked amebae and flagellates, entirely or partially feed on bacteria, skimming off the dense bacterial standing crop provided by many soils. However, under natural conditions soil protozoa cannot significantly reduce bacterial abundance, and the absolute amount of organic matter consumed by the testacean community is small – 2.34 mg m$^{-2}$ during a 3-month period.

In soil ciliates, 39% feed mainly on bacteria, 34% are mainly predaceous, and 20% are omnivorous. Some are strictly mycophagous and the most characteristic species of the soil ciliate community. They have reduced the classical ciliate mouth almost completely and evolved a minute feeding tube producing discrete holes in the hyphae and spores of fungi and yeasts (**Figure 3b–e**). This type of feeding has been termed 'perforation lysis' and is also found in several naked amebae. There is evidence from pot experiments that such activities can reduce the inoculum level of plant pathogenic fungi in soils and can reduce the severity of take-all disease of wheat by the fungus *Gaeumannomyces graminis* var. *tritici*. Some flagellates are also strictly mycophagous, but take up the fungi by an ordinary cytostome. Many soil testaceans seem to feed on humus particles and/or fungal hyphae and spores, but other materials such as algae, protozoa, and bacteria are also ingested; some are polyphagous. A Gondwanan species, *Apodera vas*, can even feed on nematodes (**Figure 2b**). In some other species, *Difflugia lucida* and *Schoenbornia humicola*, a remarkable specialization has been described: they collect and store humus particles around the pseudostome during optimal periods and transport them into the cytoplasm during suboptimal environmental conditions (**Figure 2c**).

*r/K-selection* Field data and laboratory experiments show that colpodid ciliates are more *r*- than *K*-selected. Thus, they are particularly diverse and frequent in unpredictable and extreme habitats, such as leaf surfaces, lichens, and alpine and desert soils. In contrast, polyhymenophoran ciliates (mainly hypotrichs) are more *K*-selected and thus particularly diverse and frequent in predictable habitats (**Figure 4**). The ratio of these ciliate groups proved to be a sensitive indicator for biotope stability and can thus be used in studies on human-influenced soils.

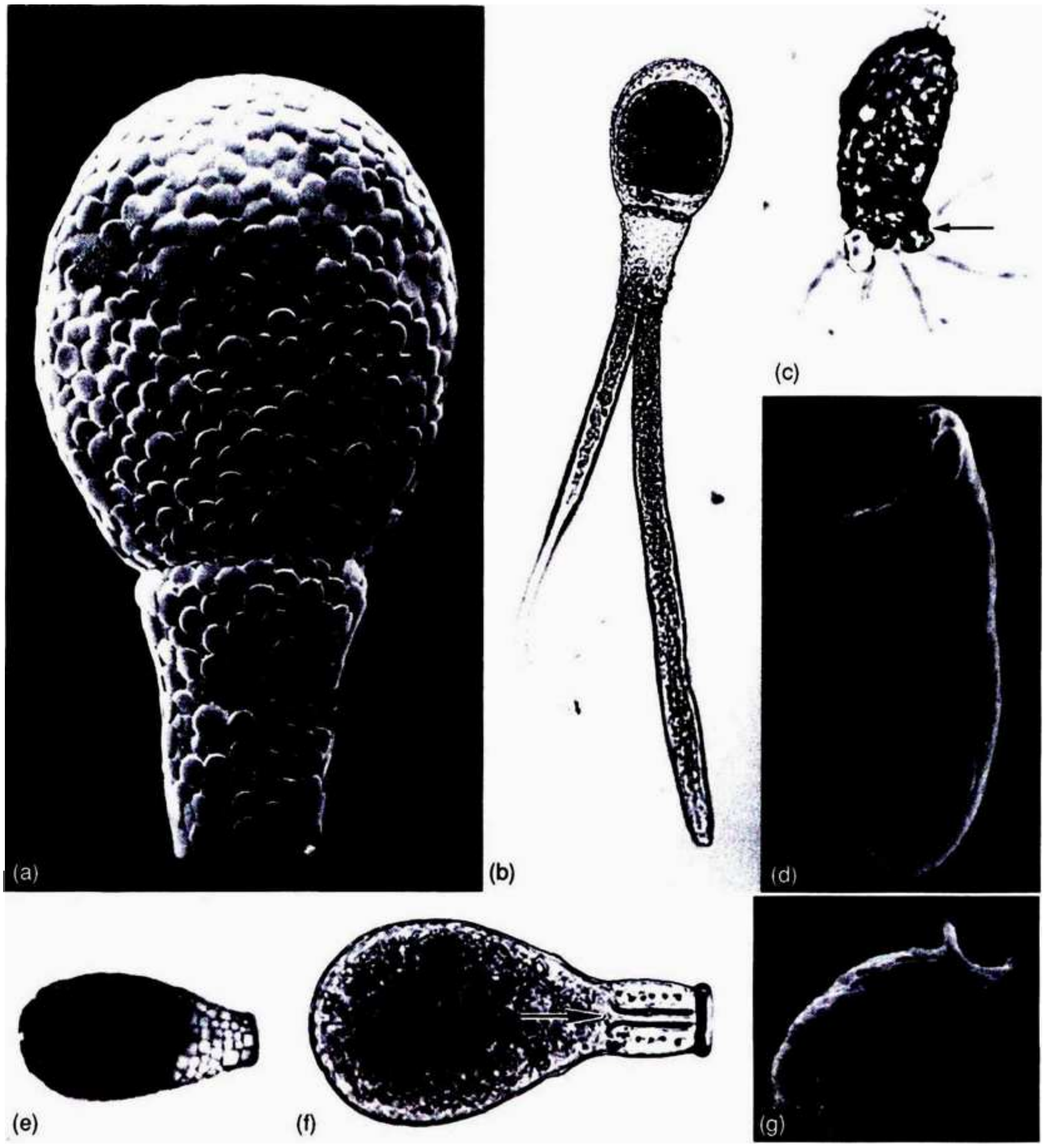


**Figure 2** Soil testate amebae. (a, b) *Apodera vas* is about 170 μm long and only occurs in the southern hemisphere, i.e., in Gondwana. It can feed on nematodes (b) and build its shell of platelets taken from the shell of another prey organism, the testate ameba *Trinema lineare* (d). (c) *Difflugia lucida* is about 60 μm long and produces food-bundles (arrow) around the shell opening under optimal environmental conditions. These food accumulations are ingested and digested when conditions are less favorable. (d) *Trinema lineare* is about 35 μm long and one of the most common soil testate amebae globally. Its shell is made of silicium platelets, which *A. vas* uses for building its own shell (a). (e) *Quadrulella* makes the test of quadrangular silicium platelets. (f) *Nebela certesi* is about 150 μm long and only occurs in the southern hemisphere (Gondwana). Note the oral tube (arrow) surrounded by minute teeth. (g) *Edaphonobiotus campascoides*, a soil testate ameba about 35 μm long has, as yet, been found only in central Europe (Laurasia). Its shell is highly similar to that of some marine sand testaceans.

Testate amebae are, in general, on the *r* side within the *r*/*K*-continuum, when compared with ciliates, naked amebae, and flagellates. However, a more fine-scaled rating suggests that autochthones are more *K*-selected than allochthones.

*Cysts* All soil protozoa survive detrimental environmental conditions in a dormant stage, termed 'protective resting cyst.' When encysting, the cell rounds up, loses water, and secretes a thick wall (**Figure 1c**). In encysted condition, protozoa can withstand, for instance, decades of drought and food depletion, liquid nitrogen, and strong acids. When conditions become favorable again, the cell excysts and commences feeding and reproduction. The receptors indicating favorable conditions to the encysted cell are still not known. Likewise, few reliable data are available on the factors stimulating excystment, but substances secreted by potential food organisms and $CO_2$ might be important. A slight increase in the $CO_2$ level to 0.5–0.8% enhanced excystment of soil amebae, but a further increase inhibited it.

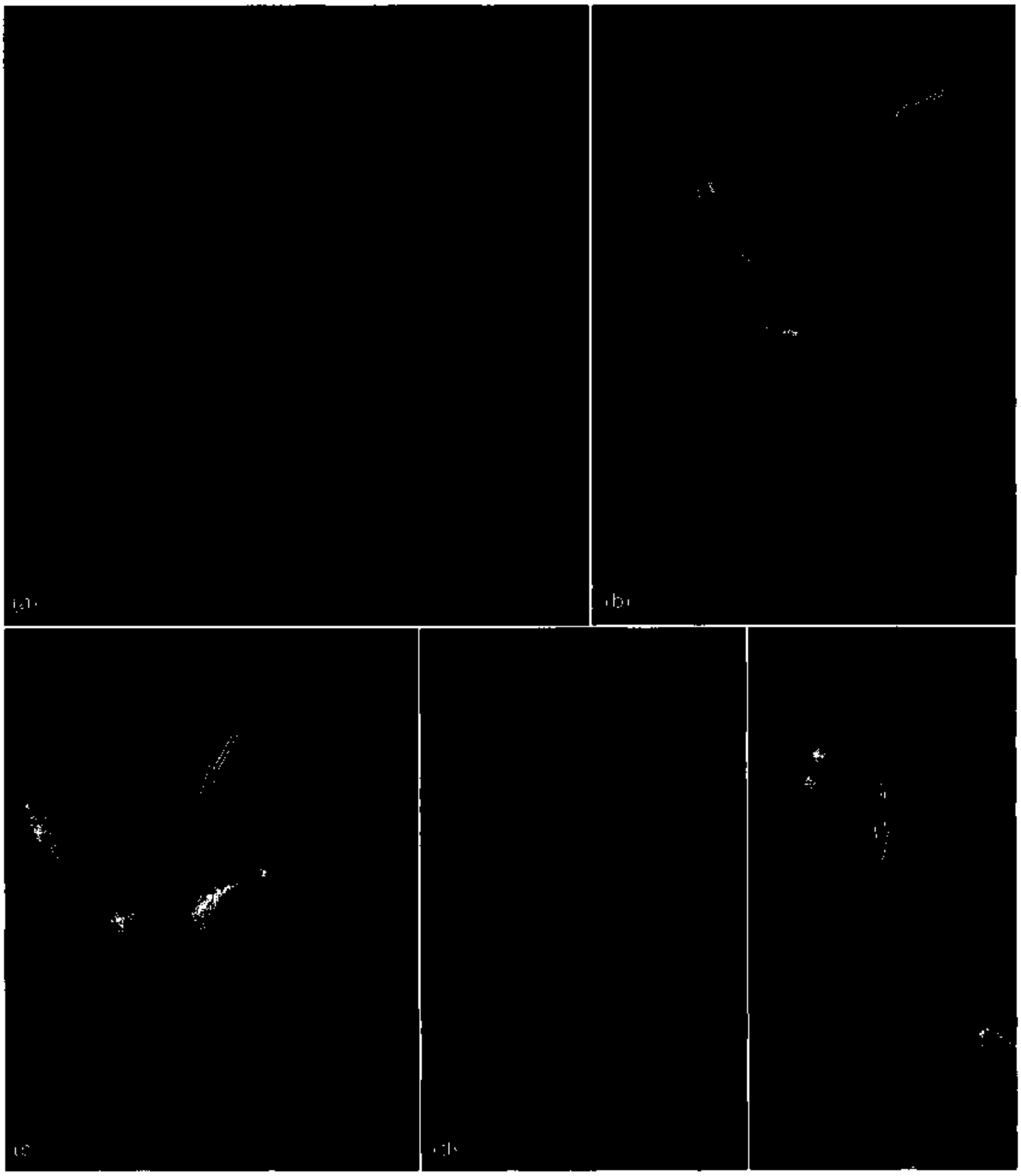


**Figure 3** (a) *Bresslauides discoideus* is a colpodid ciliate up to 600 μm large and likely restricted to the northern hemisphere. It is a predator with a large mouth and one of the largest ciliates occurring in soil. (b–e) *Pseudoplatyophrya nana* is a minute species about 25 μm long that feeds only on yeasts and fungi. It feeds by perforation lysis, that is, has a minute oral tube (arrows) that enzymatically produces small holes in the yeast's cell wall (c, d).

There is evidence that cysts are adapted to the overall environmental conditions in which the trophic cells live. Cysts formed, for instance, in arid environments can survive desiccation for decades, while those produced in rainforests survive drought for only a few weeks or months. The same is true for temperature. In experimental microcosms, diversity of soil ciliates and reproduction are highest at or near the mean annual temperature of the habitat, that is, around 20 °C for soils from the region of Bonn, Germany, and 25–30 °C for those from an African savannah. These data show the existence of many local physiological races adapted to the specific conditions of the habitat.

There are further cyst variations in soil protozoa. Testate amebae can produce a so-called 'precystic stage' for bridging short unfavorable periods, an adaptation to the often quickly changing soil environment. When forming this stage, the organism produces a thin membrane (epiphragm) closing the shell opening (pseudostome), and rounds up, but does not form a thick cyst wall. During excystment, the epiphragm is lysed or digested. Another spectacular cyst is produced by the epiphytic, colpodid ciliate genus *Sorogena*. Under favorable conditions, it multiplies rapidly, aggregates, encysts, and produces noncellular stalked aerial sori which look highly similar to those of cellular slime moulds. The sorogenesis requires alternate light and dark conditions, just as in the pink slime mold *Acrasia rosea*.

*Anabiosis* Anabiosis (cryptobiosis, anhydrobiosis) is the property of organisms to resist desiccation without marked morphological changes, such as the production of a special cyst or membrane. It is common in some groups of soil- and moss-inhabiting animals, such as rotifers, tardigrades, and nematodes, but is very restricted among soil protozoa, where it likely occurs in a few naked amebae and in a suctorian ciliate, as I myself have observed.

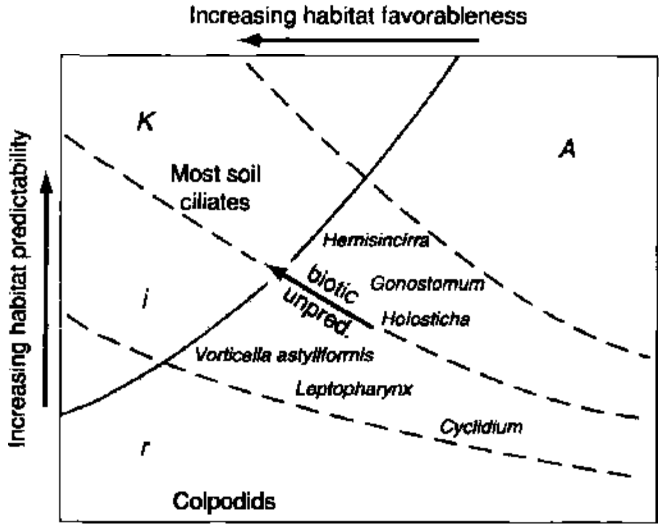


**Figure 4** Ordination of soil ciliate taxa on a two-dimensional template. Colonizing *r*-organisms, colpodids, characterize ephemeral habitats of low resources (lower left to lower right of diagram). A few intermediate *K*-selected species (e.g., *Cyclidium* and *Holosticha*) can adapt to harsh conditions. Species diversity increases towards the upper left corner of the diagram where the abiotic factors of the habitat improve, enabling more biotically unpredictable interactions due to competition and predation. *A*, adversity selection; *i*, intermediates. Reproduced with permission from Bamforth SS (2001) Proportions of active ciliate taxa in soils. *Biology and Fertility of Soils* 33: 197–203.

## Number and Biomass

In soil, protozoa are present as active (trophic) and inactive (encysted, dormant) cells. It is likely that many soil protozoan species are encysted most of their life. Thus, the encysted portion of the total number is usually much larger than the active. There is currently no reliable method available for separating active and dormant specimens. Furthermore, most counts rely on culture techniques followed by a most probable number (MPN) calculation of protozoan abundance. However, only few of the many protozoan species present excyst and grow to detectable numbers under the limited conditions of such cultures. All these methodological shortcomings mean that many of the quantitative data available in the literature are doubtful, especially as concerns the active and thus ecologically relevant portion. Only numbers obtained by direct microscopic counts in soil suspensions are reliable (**Table 2**). Unfortunately, naked amebae and flagellates can hardly be counted with direct methods because they strongly adhere to the soil particles and thus cannot be seen. Accordingly, all numbers on active naked amebae and flagellates are highly questionable.

In spite of these problems, it is clear that protozoa inhabit soil in great numbers, that is, some 10 000–1 000 000 individuals per gram dry mass (**Tables 2–4**, **Figure 5**). Biomass tends to correlate with abundances, and values between 0.1 g and 5 g m$^{-2}$ are frequent. Naked amebae and flagellates (active plus cystic) usually show numbers between 2000 and 1 000 000 g$^{-1}$ dry mass of soil. Standing crop numbers of active testate amebae are moderately high i.e. 100–1000 individuals per gram dry mass in arable land and mineral soils, 1000–10 000 in meadow and grassland topsoils, and 10 000–100 000 in forest leaf litter and upper humus horizons. However, as many testate amebae are rather voluminous, their standing crop and production biomass often surpasses that

**Table 2** Mean numbers of active ciliates per gram dry meadow soil (0–5 cm), spruce litter (0–2 cm), and beech litter (0–2 cm)

| | *Direct counts*[a] | | *MPN simplified*[b] | | *MPN complete*[c] | |
|---|---|---|---|---|---|---|
| *Sample* | *Mean*[d] | *CV (%)*[e] | *Mean*[d] | *CV (%)*[e] | *Mean*[d] | *CV (%)*[e] |
| Meadow (*n* = 7) | 60* | 80 | 102 000* | 177 | 88 000* | 207 |
| Spruce (*n* = 3) | 400 - | 57 | 79 000* | 70 | 71 000* | 93 |
| Beech (*n* = 3) | 8100 • • | 13 | 136 000* | 56 | 127 000* | 65 |

[a] Microscopic counts in soil suspensions

[b] Dilution culture method by Singh and most probable number (MPN) equation by Fischer

[c] Dilution culture method by Singh and refined MPN equation as available in computer programs.

[d] Values followed by the same symbol are not significantly different ($P < 0.05$)

[e] Coefficient of variation

Reproduced with permission from Foissner W (1999) Soil protozoa as bioindicators: pros and cons, methods, diversity, representative examples. *Agriculture, Ecosystems and Environment* 74: 95–112.

**Table 3** Effects of irrigation on the microfauna of a spruce forest soil

| *Parameters*[a] | *Soil depth (cm)* | *Irrigated*[b] | *Control* | n | *Statistics* |
|---|---|---|---|---|---|
| Soil moisture (% wet mass | 0–3 | $50.0^{*} \pm 9.9$ | $41.3^{\dagger} \pm 12.6$ | 15 | ANOVA ($P \leq 0.05$) |
| of air-dried soil) | 3–9 | $48.8^{*} \pm 4.3$ | $44.7^{*} \pm 5.9$ | 8 | ANOVA ($0.2 > P > 0.1$) |
| *Ciliates* | | | | | |
| Abundance | 0–3 | $311^{*} \pm 141$ | $489^{\dagger} \pm 258$ | 15 | U-test ($0.05 < P \leq 0.1$) |
| | 3–9 | $10^{*} \pm 10$ | $14^{*} \pm 24$ | 8 | U-test ($P \geq 0.2$) |
| Species number | 0–3 | $12.8^{*} \pm 4.5$ | $8.4^{\dagger} \pm 3.3$ | 15 | ANOVA ($P \leq 0.005$) |
| | 3–9 | $2.4^{*} \pm 2.1$ | $1.6^{*} \pm 1.9$ | 8 | U-test ($P \geq 0.2$) |
| *Testate amebae* | | | | | |
| Abundance | | | | | |
| Living | 0–3 | $22\,203^{*} \pm 7040$ | $17\,908^{*} \pm 3175$ | 4 | ANOVA ($P \geq 0.2$) |
| Empty tests | 0–3 | $361\,720^{*} \pm 17\,204$ | $319\,081^{*} \pm 58\,390$ | 4 | ANOVA ($P \geq 0.2$) |
| Species number | | | | | |
| Living | 0–3 | $11.5^{*} \pm 1.3$ | $9.8^{\dagger} \pm 1.0$ | 4 | ANOVA ($0.05 < P \leq 0.1$) |
| Total | 0–3 | $18.0^{*} \pm 0$ | $19.8^{*} \pm 2.2$ | 4 | U-test ($P \geq 0.2$) |
| *Nematodes* | | | | | |
| Abundance | 0–3 | $1197^{*} \pm 349$ | $824^{\dagger} \pm 349$ | 15 | ANOVA ($P \leq 0.01$) |
| | 3–9 | $308^{*} \pm 80$ | $391^{*} \pm 173$ | 8 | ANOVA ($P \geq 0.2$) |
| *Rotifers* | | | | | |
| Abundance | 0–3 | $227^{*} \pm 88$ | $181^{*} \pm 122$ | 15 | ANOVA ($P \geq 0.2$) |
| | 3–9 | $38^{*} \pm 25$ | $31^{*} \pm 18$ | 8 | U-test ($P \geq 0.2$) |

[a] Abundances (individuals $g^{-1}$ dry mass of soil; arithmetic mean ± standard deviation) were estimated with a direct counting method in soil suspensions. Values followed by the same symbol are not significantly different.

[b] Irrigated plot ($15\,m^2$) received $25\,l\,m^{-2}$ water every fourth day.

ANOVA, analysis of variance.

Reproduced with permission from Foissner W (1999) Soil protozoa as bioindicators: pros and cons, methods, diversity, representative examples. *Agriculture, Ecosystems and Environment* 74: 95–112.

**Table 4** Comparison of community parameters for soil testaceans from different sites of the world

| Parameters | *Moss under beech forest (moder)* | *Ash-maple forest (mull)* | *Alpine rendzina (moder)* | *Alpine mat (moder)* | *Aspen woodland (mor)* | *Beech forest (mull)* |
|---|---|---|---|---|---|---|
| Annual mean density ($\times 10^6\,m^{-2}$) | 1.7 | 33.6 | 39.6 | 20.1 | 261 | 84 |
| Annual mean biomass ($mg\,m^{-2}$) | 15.5 | 1033 | 2209 | 1165 | 723 | 1715 |
| Number of generations ($year^{-1}$) | 16.0 | 12.5 | ND | 8 | 90 | ND |
| Mortality rate (% $day^{-1}$) | 3.0 | 8.5 | ND | ND | 10.8 | ND |
| Production numbers ($\times 10^6\,m^{-2}$) | 145 | 940 | 29 | 118 | 90 930 | 358 000 |
| Production biomass ($g\,m^{-2}$) | 0.11 | 25.9 | 1.4 | 5.4 | 206 | 73 |
| Biomass turnover ($P_B/B$) | 8.1 | 43.9 | 0.7 | 5 | 285 | 43 |

ND, not determined.

Reproduced with permission from Foissner W (1987) Soil protozoa: fundamental problems, ecological significance, adaptations in ciliates and testaceans, bioindicators, and guide to the literature. *Progress in Protistology* 2: 69–212.

of all other protozoans (Table 4). Thus, their omission in many recent studies is an unfortunate mistake. Abundances of active ciliates are low in evolved soils (humus and mineral horizons), usually <100 cells $g^{-1}$ (Tables 2 and 3). However, up to 10 000 individuals per gram dry mass can be reached in developing soils and leaf litter (Table 2), showing that ciliates are important during the early stages of soil formation.

## Vertical and Horizontal Distribution

Like as many soil organisms, protozoa are usually most abundant and diverse in the upper 0–10 (30) cm soil layer, where food is concentrated in the form of leaf litter and the rhizosphere. However, there is a distinct, not yet fully understood, niche separation of ciliates and testate amebae. While active ciliates occur mainly in the uppermost (fresh) leaf litter layer, where abundances of up to 10 000 individuals per gram dry mass are reached, testate amebae peak underneath, that is, in the evolving humus horizon ('fermentation layer'; Table 3). Thus, habitats such as meadow topsoils and arable lands contain very few active ciliates, usually <100 $g^{-1}$ dry mass, although a lot of viable resting cysts are present. The inability to excyst (germ) in evolved soils has been termed ciliatostasis,

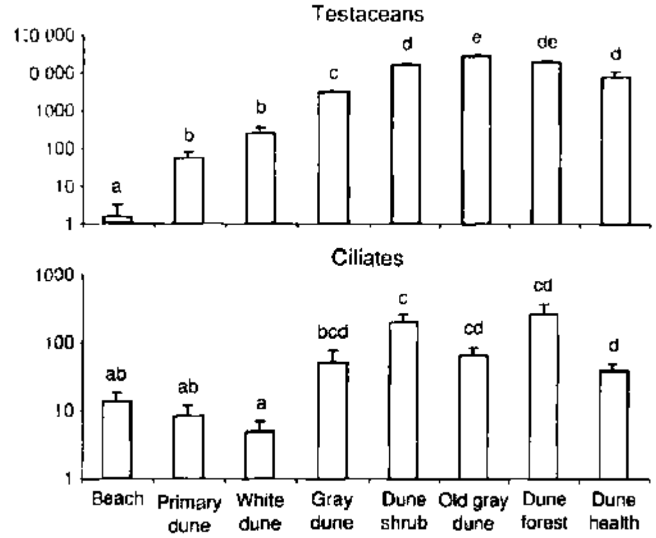


**Figure 5** Abundance (individuals $g^{-1}$ dry mass of soil) of active testate amebae and ciliates in dune soils along a series of soil (dune) development. Significant differences ($P < 0.05$) are indicated by different letters. Reproduced from Verhoeven R (2002) The structure of the microtrophic system in a development series of dune soils. *Pedobiologia* 46: 75–89.

in analogy to fungistasis. The reasons for ciliatostasis are not yet clear, but enrichment of certain detrimental substances during humification, density-dependent processes, and general food depletion are discussed.

Little is known about the horizontal distribution of soil protozoa. However, a very clumped distribution is evident from several faunistic and 'hot spot' investigations. Recent data suggest that spatial organization of protozoan communities occurs at scales below 10 cm.

### The 'Water Paradox'

Protozoa are aquatic organisms, and free water is thus indispensable for their life. Accordingly, it appears reasonable to assume that water is the key factor for soil protozoan abundance and diversity. Surprisingly, carefully designed field and laboratory studies do not support this assumption, but over the years have provided many conflicting results. Two examples from recent studies should demonstrate the 'water paradox.'

In one study, aspen woodland plots were irrigated during the seasonal dry period every 3 days with $3.78\,\mathrm{l\,m^{-2}}$ and testate amebae were sampled on three occasions. The general effect of adding water to the soil was to increase significantly the number of active testaceans and to decrease the number of encysted specimens. Numbers of active testaceans and of total living testaceans (active plus encysted) showed a significant, positive correlation with soil moisture content. At higher soil moisture levels, the larger ($>60\,\mu m$) species present tended to have a larger proportion of living individuals as cysts than did the smaller forms which numerically dominated the population. Generation times were shorter and secondary production higher in the watered plots.

In the other study, spruce forest plots were irrigated during the seasonal dry period every 4 days with $25\,\mathrm{l\,m^{-2}}$ (**Table 3**). This caused a marked decrease ($P \leq 0.05$) in the abundance of the active ciliates in the litter layer; species richness, in contrast, increased in both ciliates and testate amebae, while testacean abundance did not increase significantly. The nematode numbers increased dramatically by about 45%, while rotifer abundance did not change significantly.

As yet, there is no convincing explanation for these and many other contrasting results. However, most soil protozoologists agree on the basic assumption that water is a key factor for protozoan life, not only because they basically depend on water but also due to several morphological peculiarities most meaningfully explained as adaptations to the astatic water regime of the soil environment in general. I suggest, very briefly, two hypotheses for explaining the 'water paradox':

1. Water correlates with other, not yet recognized factors, especially the presence or growth of appropriate food organisms. This hypothesis is supported by the observation that laboratory microcosms, where food has been tested to be edible for the protozoans under investigation, show more consistent results, that is, increasing protozoan abundance and biomass with increasing soil moisture.
2. Protozoa can tolerate a wide range of soil moisture; only under extreme conditions do communities react. This hypothesis is supported by the occurrence of local, physiological races and the ability of many protozoan species to live in very thin films of water covering the soil pores and soil particles. In the spruce forest experiment described above, the litter of the control plot was hand-dry; most of the water present was a constituent of the needles (**Table 3**).

## Ecological Significance

### Standing Crop, Respiration, and Annual Production

Several ecosystem studies show consistently that the microflora (bacteria and fungi) accounts for about 90% of the heterotrophic soil respiration. Only about 10% of the total comes from the 'animals,' and about 70% of the 'animal' respiration is due to the protozoa, although their contribution to the standing crop is 'only' approximately 30% (**Figure 6**). This is due to their small size and fast reproduction. However, in earthworm-rich soils the proportion of

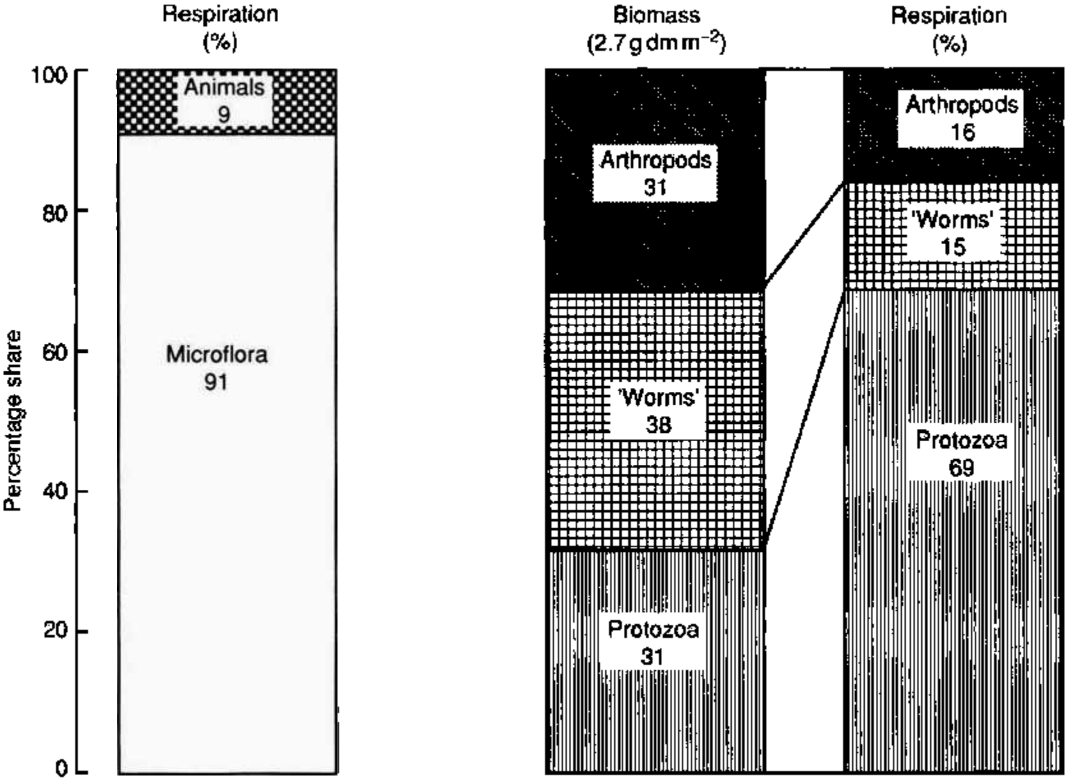


**Figure 6** Contribution of protozoa to dry biomass (dm) and respiration of soil organisms. The graphs show the mean of 14 ecosystem studies from various sites of the world. Reproduced with permission from Foissner W (1987) Soil protozoa: fundamental problems, ecological significance, adaptations in ciliates and testaceans, bioindicators, and guide to the literature. *Progress in Protistology* 2: 69–212.

protozoa to the total biomass is considerably lower than stated above, while it is higher in extreme regions, where earthworms are few or lacking, for instance, in coniferous forests, deserts, and high mountains.

Field studies of production of soil protozoa are mostly restricted to testate amebae, the shells of which can be used to estimate their mortality. The annual production is in the same order of magnitude as that of the earthworms, especially when productions of naked amebae and flagellates are added (Table 4). The high contribution of the protozoa to the total annual animal production is due to the many generations produced and shows that protozoa play an important role in the energy flux of the soil ecosystem.

### Mineralization and Plant Growth

Considering the fast growth of protozoa and their high annual production (Table 4), it is not surprising that they enhance flows of nutrients in soil to the benefit of plants and microorganisms. Experiments suggest that a complex 'microbial food-web' is operating rather than a simple 'microbial loop' during decomposition and growth. Convincing evidence is available that protozoa enhance plant growth significantly ($\geq$20%) either directly by increasing the availability of nitrogen in soil and leaching water and/or indirectly by nonnutritional effects, such as selective grazing on microbial populations, modification in the concentrations of plant hormones in the rhizosphere, or suppression of pathogenic bacteria. This matter is still under discussion. About 40% of the ingested microbial biomass N will be excreted by protozoa, leading to a constant release of nutrients available for plant uptake. Protozoa are not only responsible for respiring about 10% of the total carbon input but also for 20–40% of net N mineralization in the field. Furthermore, protozoa enhance growth of the plant root system and thus nutrient uptake (Figure 7).

Protozoa also cause a strong decrease in phosphate leaching. Phosphate is one of the most immobile nutrient ions in soil and continuous microdisturbance by protozoa and other microfaunal organisms (e.g., nematodes) is likely particularly important for microbial phosphate availability.

### Earthworm Growth

Early reports that earthworms feed on protozoa and are an essential constituent in the diet of the compost worm *Eisenia fetida* to reach sexual maturity were recently confirmed and extended by carefully designed experiments. These showed that *Aporrectodea caliginosa* actively searchs for places with high

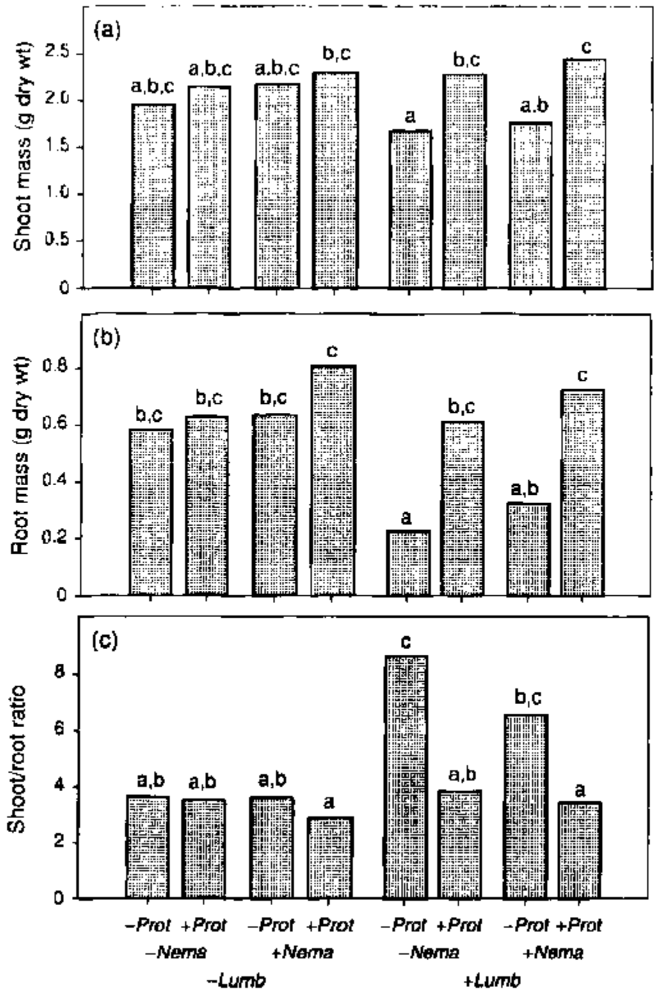


**Figure 7** Effects of protozoa (*–Prot* without protozoa, *+Prot* with protozoa), nematodes (*–Nema* without nematodes, *+Nema* with nematodes) and earthworms (*–Lumb* without earthworms, *+Lumb* with earthworms) on (a) shoot biomass; (b) root biomass; and (c) shoot/root ratio of *Hordelymus europaeus*, a grass, in experimental chambers at the end of the experiment (112 days). Bars sharing the same letter are not significantly different. $P > 0.05$ (three-way ANOVA). Reproduced with permission from Alphei J, Bonkowski M, and Scheu S (1996) Protozoa, Nematoda and Lumbricidae in the rhizosphere of *Hordelymus europaeus* (Poaceae): faunal interactions, response of microorganisms and effects on plant growth. *Oecologia* 106: 111–126.

protozoan densities and feeds on them. The weight gain of young *A. caliginosa* in soil with naked amebae was twice that in soil without amebae.

## Protozoa as Bioindicators in Terrestrial Environments

There is considerable literature on the use of protozoa as bioindicators in terrestrial environments, and recently appropriate soil protozoan bioassays became available. Previous reviews summarized data obtained until 1998 and discuss general problems and methodology of bioindication. It has been shown that the full power of bioindication needs identification of species and/or functional groups because the information content of individual numbers and biomasses is often rather limited. Here, some recent studies are reviewed to demonstrate the great potential of protozoa in the investigation of natural and human-influenced soil ecosystems.

Compared to the soil meso- and macrofauna, protozoa have several specific features making them especially useful for bioindication:

1. Protozoa are an essential component of terrestrial ecosystems, because of their large standing crop and production and their high food consumption and respiration per mass unit (**Figure 6**, **Table 4**). Thus, changes in their numbers and diversity influence the rate and kind of soil formation and soil fertility.
2. Protozoa, with their rapid growth, short lifetime, and delicate external membranes, can react more quickly to environmental changes than any other eukaryotic organism and can thus serve as an early warning system and excellent tool in bioassays. Results are obtained within a few days, faster than with any other eukaryotic test system.
3. The genome of the protozoa is similar to that of the metazoa. Their reactions to environmental changes can thus be related to higher organisms more convincingly than those of the prokaryotes.
4. Morphologic and genetic differentiation among globally distributed protozoan species is rather low, suggesting that many indicator species can be used worldwide.
5. Protozoa inhabit and are particularly abundant in those soil ecosystems that almost or entirely lack higher organisms, especially earthworms, due to extreme environmental conditions.
6. Protozoa are not readily dislodged in soil. The difficult problem of horizontal migration, especially with the epigaeon, does not affect the investigations.

Protozoa also pose problems when used as bioindicators; specifically, counting methods are time-consuming and species identification is troublesome due to the lack of comprehensive identification literature. There are indications that molecular methods and taxonomic reviews can partially solve these drawbacks.

### Soil Protozoa as Indicators in Natural Ecosystems

The main humus types, mull, moder, and mor, can be finely discriminated by certain protozoan species, especially testate amebae. Further, some ciliates are excellent indicators for periodic or sporadic oxygen depletion, which is often difficult to recognize and measure with physicochemical methods. Of course, protozoa can also be used to characterize certain habitats and soil development, for instance, in coastal dunes of Germany (**Figure** 5). This study shows, *inter alia*, that the content of organic matter is a key factor for the abundance and diversity of the microfauna, and testate amebae are the most important group, reaching up to 50 000 individuals per gram dry mass. Further, active ciliates are comparatively abundant. These data are in accordance with niche separation of testate amebae and ciliates and the concept of ciliatostasis reported in the section on vertical distribution.

### Protozoa as Indicators in Reclaimed Open-Cast Coal Mining Areas

Topsoil removal, as occurs in open-cast mining areas and during the preparation and use of high-mountain ski slopes, heavily impoverishes protozoa and soil life in general (**Figure** 8). However, colonization occurs within a few months. Abundances and biomasses in 2- up to 46-year-old afforested mine soils are in the same order as in undisturbed forest soils, but typical humus-inhabiting, large-sized testacean species are lacking or occur rarely, showing that reclamation was only partially successful. Six ubiquists out of 48 taxa contributed 61–87% to the overall abundance. Generally, all test sites had a distinct testacean community, whose structure depended on age, substrate, and stocking.

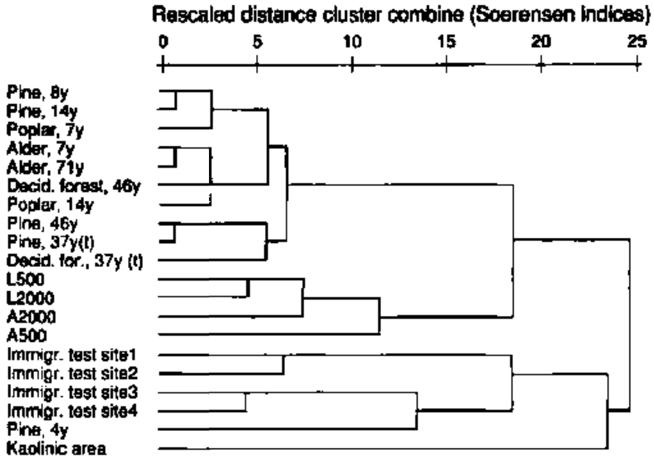


**Figure 8** Cluster based on similarity (Soerensen index, UPGMA-linkage) of numbers and species of testate amoebae in various immigration test sites, minicontainers (sites L, A), and recultivated forests. The distances are rescaled to fall in the range of 1–25. Quality and development of the substrates exposed for different time intervals were more important for immigration and colonization success than the adjacent source habitats. Reproduced with permission from Wanner M and Dunger W (2001) Biological activity of soils from reclaimed open-cast coal mining areas in Upper Lusatia using testate amoebae (protists) as indicators. *Ecological Engineering* 17: 323–330.

### Protozoa as Indicators of Soil Decontamination

Decontamination of polluted soils is a current challenge for scientists from various fields. Few data are available as concerns the soil fauna, and only one study included protozoa (**Figure 9**). Immediately after decontamination, active protozoa, nematodes, and collembola were not detected in treated soil. However, protozoan cysts survived treatment (5700 ± 4200 cysts per g dry soil). Following 1 week of exposure, no active protozoa were found. From the second sampling date onwards (after 3 weeks), protozoa were observed with significantly increasing biomass until the end of the experiment 12 weeks after exposition.

### Effects of Biocides

A considerable amount of literature is available on the effects of biocides on soil protozoa. Several reviews showed that: (1) the general reaction of soil protozoa to biocide stress matches that of other organisms; (2) many protozoan species seem to be just as sensitive to pesticides as other more commonly used test organisms; (3) insecticides are usually more toxic than herbicides; (4) insecticides disturb soil protozoa critically, i.e., populations often do not fully recover within 60 days; and (5) investigations should be performed at species and/or functional group level because the simple measure 'total protozoan abundance' is often too insensitive. This is also obvious from a recent study, where the effects of the fungicide fenpropimorph have been investigated in laboratory microcosms and field experiments. While the microcosms showed that bacterivorous protozoa were affected by fenpropimorph at concentrations

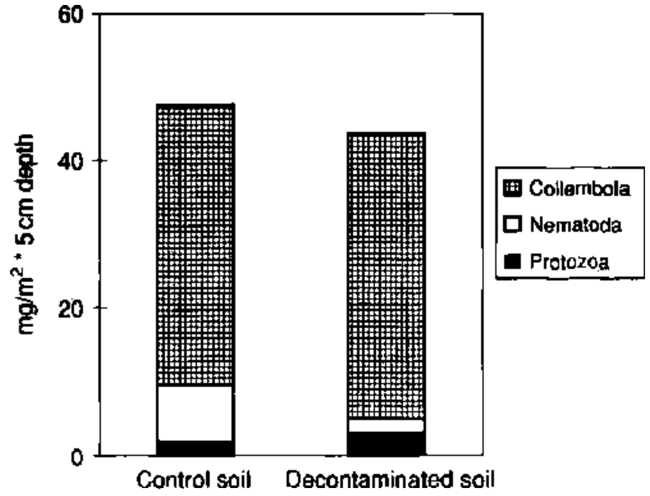


**Figure 9** Biomass of Collembola, Nematoda and Protozoa in control and decontaminated soil after 12 weeks of exposure in an arable field. Reproduced with permission from Böckl M, Blay K, Fischer K, Mommertz S, and Filser J (1998) Colonisation of a copper-decontaminated soil by micro- and mesofauna. *Applied Soil Ecology* 9: 489–494.

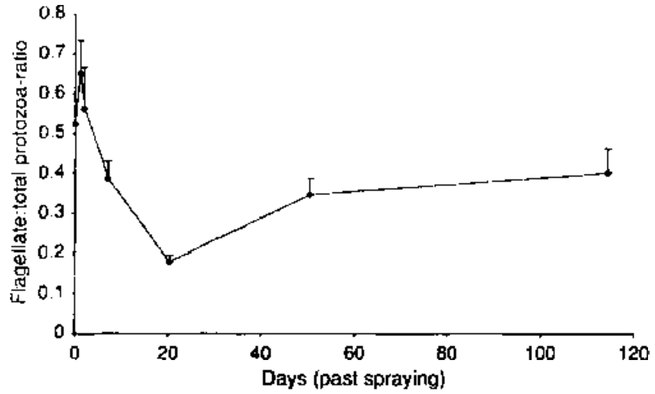


**Figure 10** A significant decrease in the flagellate:total protozoan ratio was recognizable when fenpropimorphic acid, a degradation product of fenpropimorph, peaked 20 days after pesticide application. Total protozoan abundance was, however, not significantly changed. Reproduced with permission from Bjornlund L, Ekelund F, Christensen S *et al.* (2000) Interactions between saprotrophic fungi, bacteria and protozoa on decomposing wheat roots in soil influenced by the fungicide fenpropimorph (Corbel®): a field study. *Soil Biology and Biochemistry* 32: 967–975.

lower than those expected in soil after normal field application, the field experiments showed only a slight decrease in the flagellate-to-total protozoan ratio after 20 days (**Figure 10**).

*See also*: **Bacteria:** Plant Growth-Promoting; Soil; **Fauna**; **Fungi**; **Microbial Processes:** Environmental Factors; **Mycorrhizal Fungi**; **Nematodes**; **Pollutants:** Biodegradation; **Root Exudates and Microorganisms**

## Further Reading

Alphei J, Bonkowski M, and Scheu S (1996) Protozoa, Nematoda and Lumbricidae in the rhizosphere of *Hordelymus europaeus* (Poaceae): faunal interactions, response of microorganisms and effects on plant growth. *Oecologia* 106: 111–126.

Bamforth SS (2001) Proportions of active ciliate taxa in soils. *Biology and Fertility of Soils* 33: 197–203.

Bjornlund L, Ekelund F, Christensen S *et al.* (2000) Interactions between saprotrophic fungi, bacteria and protozoa on decomposing wheat roots in soil influenced by the fungicide fenpropimorph (Corbel®): a field study. *Soil Biology and Biochemistry* 32: 967–975.

Böckl M, Blay K, Fischer K, Mommertz S, and Filser J (1998) Colonisation of a copper-decontaminated soil by micro- and mesofauna. *Applied Soil Ecology* 9: 489–494.

Bonkowski M and Schaefer M (1997) Interactions between earthworms and soil protozoa: a trophic component in the soil food web. *Soil Biology and Biochemistry* 29: 499–502.

Corliss JO (2000) Biodiversity, classification, and numbers of species of protists. In: Raven PH and Williams T (eds) *Nature and Human Society*, pp. 130–155. Washington, DC: National Academy Press.

Darbyshire JF (ed.) (1994) *Soil Protozoa*. Oxon: CAB International.

Foissner W (1987) Soil protozoa: fundamental problems, ecological significance, adaptations in ciliates and testaceans, bioindicators, and guide to the literature. *Progress in Protistology* 2: 69–212.

Foissner W (1994) Soil protozoa as bioindicators in ecosystems under human influence. In: Darbyshire JF (ed.) *Soil Protozoa*, pp. 147–193. Oxon: CAB International.

Foissner W (1999) Soil protozoa as bioindicators: pros and cons, methods, diversity, representative examples. *Agriculture, Ecosystems and Environment* 74: 95–112.

Verhoeven R (2002) The structure of the microtrophic system in a development series of dune soils. *Pedobiologia* 46: 75–89.

Wanner M and Dunger W (2001) Biological activity of soils from reclaimed open-cast coal mining areas in Upper Lusatia using testate amoebae (protists) as indicators. *Ecological Engineering* 17: 323–330.

# Q

## QUALITY OF SOIL

**B J Wienhold, G E Varvel, and J W Doran**, USDA Agricultural Research Service, Lincoln, NE, USA

Published by Elsevier Ltd.

### Introduction

Soil forms a thin layer over the surface of the earth that performs many functions essential to life. It serves as a substrate to support plant growth, a reservoir for many nutrients essential for plant growth, a filter to maintain air quality through interactions with the atmosphere, and a storage and purification medium for water. Soil also supports biological activity involved in the decomposition and recycling of animal and plant products; and it has been identified as having a major role in carbon sequestration and mitigation of climate change associated with anthropogenic emissions of carbon dioxide and other greenhouse gases. Maintaining these soil functions in natural and managed systems is essential for environmental quality, system productivity and biodiversity. Human activity has affected soil functions in many natural and managed systems. Management practices have been developed to manipulate the soil and its functions, to facilitate planting and harvesting of food, forage, and fiber, and to improve productivity. Commonly used management practices include grazing systems, prescribed fire, forest management, tillage, fertilization, crop rotation, water management, liming, and cover crops.

'Soil quality' refers to the capacity of a soil to sustain biological productivity, maintain environmental quality, and promote plant and animal health. Land-management practices influence soil quality to varying degrees, and this influence can be positive or negative. Any management practice misapplied can result in soil degradation, and this has been the case with many traditional practices. As our understanding of how management practices affect soil quality improves, new practices are developed that maintain or improve soil quality. Monitoring soil quality allows those managing the land to assess the impact their decisions are having on the soil and the functioning of that resource. Hence, soil quality serves as a tool for monitoring soil function and system sustainability. Soils are not managed for soil quality per se; rather, soils are managed to optimize inputs to maintain soil functions and generate an acceptable rate of return on investment. In managed systems, soils are managed to produce food, forage, and fiber needed by society. In addition, producers and other land managers may use soil-quality indicators to focus on the intended result of soil management, i.e., improvement of the fitness of a soil for a particular use, not merely the characterization of the soil material itself. Soil quality is a tool that managers can use to identify problems in the short term and to develop management practices that are sustainable over the long term.

### Management Systems

Soil-quality and land management have the dual goals of meeting the needs of society and conserving essential soil, water, and air resources for future generations. Rangelands are managed for forage production, water harvesting, recreation, and wildlife production. On rangelands, management is heavily focused on vegetation. In these systems, animal stocking rates are managed to maintain vegetation composition and productivity, while water-source and mineral-supplement locations are used to improve animal distribution and avoid overgrazing and compaction of the soil. Avoiding compaction and maintaining vegetative cover reduces the potential for erosion and maintains water quality in rangeland watersheds. Prescribed fire, fertilization, and seeding are other management practices that can be used to improve degraded rangelands and maintain critical soil functions.

In forests, soils are managed for wood and fiber production and other multiple uses such as water harvesting, recreation, and wildlife. During timber-harvesting operations, soil quality is maintained by using practices that minimize the potential for erosion and restore vegetative cover as quickly as possible. Prescribed fire, fertilization, and seeding (or seedling transplants) can be used to improve site productivity and maintain critical soil functions.

In agriculture, soils are managed primarily for the production of food and fiber. Historically, tillage has been used to prepare a seedbed, incorporate residue, control weeds, and distribute agricultural chemicals. Inorganic fertilizer, animal manure, or other organic by-products are applied to provide nutrients needed by the crop but not supplied in sufficient quantities by the soil or to replace nutrients removed by the harvested crop. Lime is applied to increase soil pH. Crop sequences and crop rotations are used to break disease cycles, diversify crop production, and provide additional marketing options. Soil quality monitoring allows agricultural managers to make assessments of the effects of various combinations of these management factors in the overall system.

## Soil Quality Assessment

Soil quality is an effective tool for monitoring soil function. Assessing soil quality involves measuring physical, chemical, and biological soil properties and using these measured values to identify properties of the soil that may be inhibiting soil function or to monitor how changes in management are affecting soil functions.

The initial steps in this assessment are to identify the management goals (e.g., food or fiber production, environmental protection) and to determine the soil functions that contribute toward accomplishing them. For a soil managed for food production, functions such as providing a suitable plant substrate and serving as a nutrient reservoir are important. For a soil managed for environmental protection, key functions are storage and purification of water, and interactions with the atmosphere.

Once soil functions important to a management goal are identified, soil properties influencing those functions can be selected for measurement (Table 1), depending on location and soil type. For example, soil properties important for a soil being managed for irrigated crop production in an arid environment (e.g., sodicity, depth to water table) will be different from those for a soil in a rain-fed environment (e.g., infiltration rate, potentially mineralizable N, extractable P). Selection of soil properties for soil quality assessment also needs to take into account the expertise and resources available to the person doing the assessment. Some soil properties are better indicators of soil quality than others, but special facilities or equipment may be required to quantify the property; in which case they are often of little practical use.

A one-time measure of these soil properties may identify properties that fall outside an acceptable range of values for that soil. Once soil properties exhibiting an unacceptable value are identified, management practices can be implemented to improve them and restore the associated soil function. If multiple management practices are being used on a single soil type, a one-time comparison of soil properties among the practices can be used to determine the effect the various practices have on soil quality. Such one-time assessments are termed 'comparative soil quality assessments.' A more effective approach is to sample the soil repeatedly (e.g., two or three times each year or at the same time of year for several years) and assess management effects on soil properties over time (i.e., during the year or from year to year). Repeated measurement of soil properties is termed

**Table 1** Potential soil properties for assessing soil functions

| *Soil function* | *Physical properties* | *Chemical properties* | *Biological properties* |
|---|---|---|---|
| Substrate for plant growth | Bulk density<br>Aggregate stability<br>Soil depth | Electrical conductivity<br>Sodium adsorption ratio | |
| Nutrient reservoir | | Inorganic N content<br>Extractable P<br>pH | Potential mineralization |
| Atmospheric interactions | Water-filled pore space | $CO_2$ flux<br>$NO_x$ flux<br>Methane flux<br>$NH_4$ flux | |
| Storage and purification of water | Infiltration rate<br>Water-holding capacity | $NO_3$-N concentration<br>Extractable P | |
| Biological activity | Water-filled pore space | | Microbial biomass<br>Soil respiration<br>Potential mineralization<br>Earthworm population |

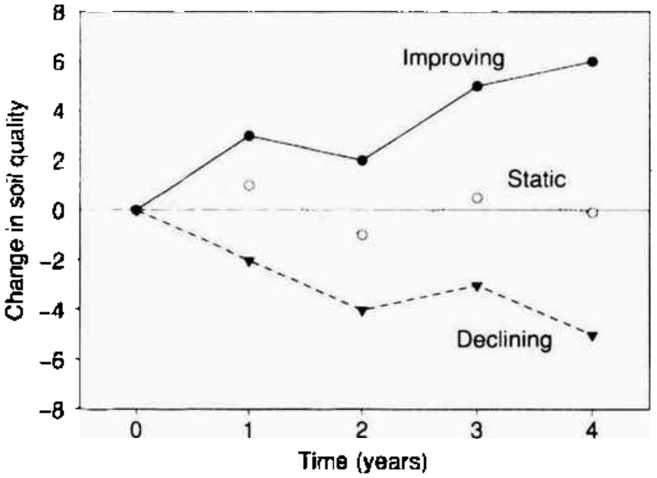


**Figure 1** Hypothetical outcomes for a dynamic soil quality assessment.

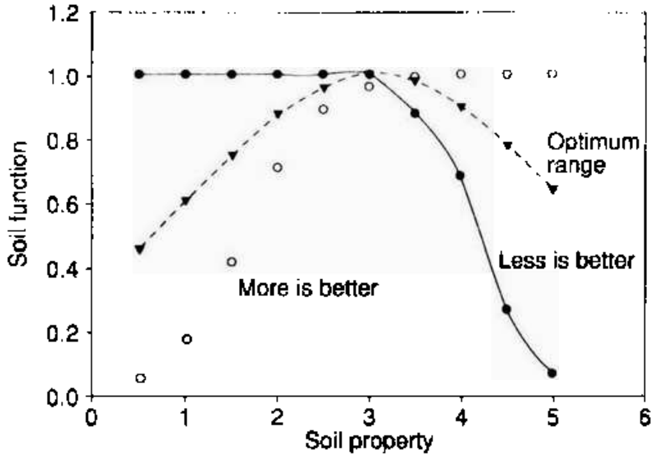


**Figure 2** Hypothetical indicator curves describing the relationship between a soil property and soil function. (Adapted from unpublished data from Andrews SS, Karlen DL, and Seybold C.)

'dynamic soil-quality assessment.' By repeatedly measuring a soil property, the direction and magnitude of change can be assessed and management practices can be adjusted (**Figure 1**).

## Soil Function–Soil Indicator Relationships

Relationships exist between soil properties and soil functions, which may take the form of: (1) lower or decreasing in value is better, (2) higher or increasing in value is better, or (3) an optimum range (**Figure 2**). As soil functions are identified, appropriate soil properties are selected to assess those functions. Identifying critical soil functions and selecting soil properties to be measured for assessing soil-quality management goals are decisions that are best made on a case-by-case basis.

Evaluation of soils functioning as a substrate for plant growth includes measurement of soil properties such as bulk density, soil depth, and electrical conductivity. These properties affect the rooting activity of plants and influence the ability of the plant to acquire water and nutrients. The relationship between bulk density and rooting activity takes the form of 'less is better', with rooting activity being depressed when the bulk density exceeds a certain value. The value at which bulk density restricts rooting activity varies with soil texture and ranges from 1.4 g cm$^{-3}$ for a clay soil to 1.7 g cm$^{-3}$ for a sandy soil. The relationship between soil depth and rooting activity takes the form of 'more is better', with the relationship leveling off when soil depths exceed the rooting depth of the crop. The relationship between rooting activity and electrical conductivity takes the form of 'less is better'. As electrical conductivity increases, the effect on rooting activity and subsequent plant growth is dependent on plant species. Crops such as bean (*Phaseolus vulgaris* L.), with a threshold of 1.0 dS m$^{-1}$, are much more sensitive to electrical conductivity (salinity) than others such as barley (*Hordeum vulgare* L.), with a threshold of 8.0 dS m$^{-1}$.

Evaluation of soils functioning as nutrient reservoirs includes measurement of soil properties such as the concentration of nutrients essential to plant growth and pH. These properties are measures of the soil's ability to provide nutrients needed by the plant. The relationship between nutrient concentration and ability to meet plant demand takes the form of 'more is better' for most nutrients. For some nutrients the relationship takes the form of 'more is better' at low concentrations and 'less is better' at high concentrations, due to negative effects on crop quality. The range of concentrations needed to meet crop needs varies among crops and soils. Values needed to develop these relationships are available in the soil nutrient-testing literature. The relationship between plant growth and pH is of the form of an optimum range. The availability of many nutrients essential to plant growth are highest at pH values between 5.5 and 7.5. At higher pH values, nutrient availability is reduced; at lower pH values, nutrient availability is reduced and availability of toxic elements such as aluminum increases.

Evaluation of soil interactions with the atmosphere includes measurement of soil properties such as respiration and flux of greenhouse gases. These properties provide a measure of whether the soil is serving as a source or sink for greenhouse gases – carbon dioxide, ($CO_2$), methane ($CH_4$), and nitrogen oxides ($NO_x$). The relationship between respiration and atmospheric interactions takes the form of an optimum range with lower values than optimum suggesting that beneficial microbial activity is too low (e.g., suboptimal release of nutrients from organic matter, low N-fixation rates, poor development of soil structure) and high values suggesting that microbial activity is too high (e.g., degradation of soil organic matter and release of nutrients in amounts exceeding what plants can utilize). Optimum value is determined by comparing organic matter inputs (crop residue and animal manure) with net respiration. The relationship between greenhouse gas fluxes and the atmosphere takes the form of 'less is better'. Low gas fluxes are a consequence of the activity of microorganisms naturally present in the soil. High gas fluxes represent the presence of labile C and nutrient substrates (e.g., nitrogenous fertilizers) under conditions (e.g., water-filled pore space of ~80%) that stimulate microorganisms that produce greenhouse gases as products of their metabolism (**Figure** 3).

Evaluation of soils as a medium for the storage and purification of water includes measurement of properties such as infiltration rate and nutrient concentrations. These properties provide a measure of the potential for contamination of surface-water and groundwater resources. The relationship between infiltration rate and storage and purification of water takes the form of 'more is better' and differs among soils as a function of soil texture. Low infiltration rates contribute to runoff and erosion losses of nutrients to surface water. The relationship between soil nutrient concentrations and storage and purification of water takes the form of 'less is better'. High nutrient concentrations in surface soils can contribute to contamination of surface waters when they are transported in runoff and erosion sediments. High nutrient concentrations in subsurface soils can contribute to groundwater contamination when they are leached from the root zone.

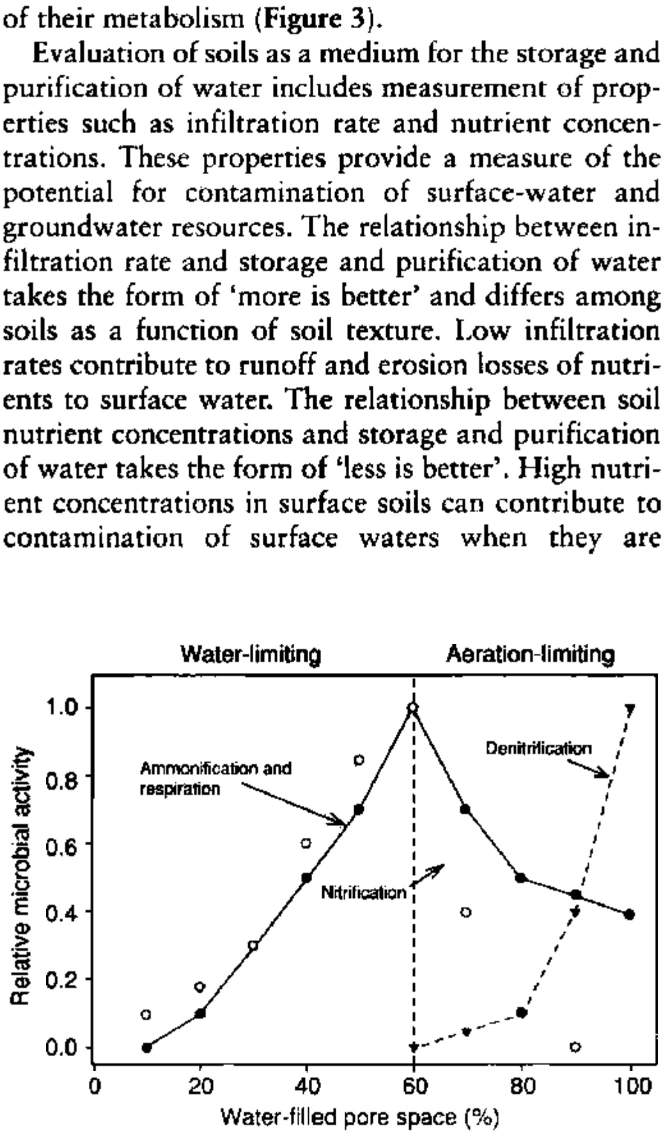


**Figure 3** Hypothetical relationship between soil water-filled pore space and microbial activity associated with respiration, nitrification, and denitrification. (Source: Linn DM and Doran JW (1984) Effects of water-filled pore space on carbon dioxide and nitrous oxide production in tilled and non-tilled soils. *Soil Science Society of America Journal* 48: 1267.)

Evaluation of the biological activity of a soil includes measurements of soil properties such as water-filled pore space and microbial biomass. These properties provide a measure of the soil's ability to decompose added material and to recycle nutrients contained in that material. Including physical soil properties in the evaluation of biological activity provides information about whether the products resulting from biological activity are from aerobic metabolic pathways (e.g., carbon dioxide, nitrate, sulfate) or from anaerobic metabolic pathways (e.g., methane, oxides of nitrogen, hydrogen sulfide). The relationship between water-filled pore space and biological activity takes the form of an optimum range. The optimum range corresponds to a moist, well-aerated soil. Microbial activity is greatest at 60% of water-filled pore space and declines linearly as the soil dries (**Figure** 3). When water-filled pore space reaches 80%, anaerobic processes such as denitrification increase. Microbial biomass responds more quickly to management than does soil organic carbon. Increases in microbial biomass are usually associated with improved management and an increase in soil quality.

For each soil function, soil quality is the integrated total of the selected set of properties i.e., the integration of soil physical, chemical, and biological processes specific to a particular soil and environment. Because it is location-specific, care must be taken when making soil-quality comparisons between soils. A soil that has been poorly managed for a long period of time may have a relatively low soil quality as compared to a similar, yet well-managed soil. Improved management is likely to have a much greater effect on the soil quality of the poorly managed soil.

## Summary

Since soil quality is assessed in terms of the fitness of a soil for its intended use, conflicts can arise when there are multiple uses and management goals for a soil. The quality of an agricultural soil may be quite different if the management goal being assessed is crop production than if the management goal being assessed is environmental services. A manager may incorporate residue using tillage to produce a seedbed and promote mineralization of nutrients. Fertilizers may also be used to provide adequate amounts of nutrients needed by the crop. Assessing the crop-production management goal for a given field may reveal that soil properties such as nutrient concentration, potentially mineralizable

N, and soil respiration are within the optimum range. Assessing the environmental-services goal for the same field may reveal that nutrient concentrations may be near levels with potential for leaching losses to groundwater, that denitrification rates are high following precipitation or irrigation, or that decomposition of the incorporated residue results in high carbon dioxide fluxes. Conflicting assessment outcomes are not a failure of the soil-quality concept. Knowledge of the soil-quality status for multiple management goals provides the potential for making adjustments in management practices to minimize negative aspects while maintaining positive aspects. The goal is that such assessment of management goals and adjustment of management practices will improve the sustainability of the system.

Concerns regarding the soil quality concept include the need to test underlying concepts in more soils and production systems and to develop procedures for addressing the potential for conflicting outcomes. Soil quality assessment is a management tool that is being improved to include more management goals, soil functions, and soil properties. The concept is being used in a wider range of soils and production systems. The utility of the soil quality concept improves as our understanding of the soil resource improves. Proponents and sceptics of the soil quality concept share a common goal: to improve management of the soil resources to maintain critical soil functions.

## List of Technical Nomenclature

| | |
|---|---|
| **Ammonification** | Biological process leading to ammonium formation from nitrogen-containing organic compounds |
| **Denitrification** | Biological reduction of nitrate and nitrite to nitrogen and nitrogen oxides |
| **Labile carbon** | Carbon-containing compounds that can be readily metabolized by microorganisms |
| **Nitrification** | Biological oxidation of ammonium to nitrate and nitrate |
| **Respiration** | Oxidative reactions providing energy to an organism |

*See also:* **Acidity; Carbon Cycle in Soils:** Dynamics and Management; **Compaction**; **Conservation Tillage**; **Crop Rotations**; **Crop-Residue Management**; **Cultivation and Tillage**; **Degradation**; **Erosion:** Water-Induced; Wind-Induced; **Fertility**; **Manure Management**; **Nutrient Management**; **Organic Matter:** Principles and Processes

## Further Reading

Acton DF and Gregorich LJ (1995) *The Health of Our Soils: Toward Sustainable Agriculture in Canada.* Ottawa, ON, Canada: Agriculture and Agri-Food Canada.

Andrews SS, Karlen DL, and Mitchell JP (2002) A comparison of soil quality indexing methods for vegetable production systems in Northern California. *Agriculture, Ecosystems and Environment* 90: 25–45.

Doran JW and Jones AJ (1996) *Methods for Assessing Soil Quality.* Soil Science Society of America Special Publication No. 49. Madison, WI: Soil Science Society of America.

Karlen DL, Mausbach MJ, Doran JW, Cline RG, Harris RF, and Schuman GE (1997) Soil quality: a concept, definition, and framework for evaluation. *Soil Science Society of America Journal* 61: 4–10.

Schjonning P, Emholt S, and Christensen BT (eds) (2004) *Managing Soil Quality – Challenges in Modern Agriculture.* Wallingford, UK: CABI Publishing.

Sojka RE and Upchurch DR (1999) Reservations regarding the soil quality concept. *Soil Science Society of America Journal* 63: 1039–1054.

# R

## RADIATION BALANCE

**J L Hatfield, T J Sauer, and J H Prueger**,
National Soil Tilth Laboratory, Ames, IA, USA

Life on Earth depends upon the flow of energy. Living objects balance their energy inputs with outputs to create a comfortable environment. Humans have an advantage because they create clothing and shelter to help modify their environment. Plants and soil also have inputs of energy, but have limited controls on the dissipation of energy and even less control on the energy inputs into the system. When we begin to think about the balance of energy in the soil or plant system we can begin to understand some of the remarkable balances that exist within the natural environment. Understanding the surface radiation and energy balances provides an opportunity to explore the role of different components, e.g., crop residue, tillage, cover crops, on soils and the environment.

The balance of radiation is similar to a checkbook or savings account. How much money is in it at any one time depends upon how much has been put into the account and how much has been taken out. This balance can be expressed very simply in word form. Net radiation is the difference between the gains (incoming solar radiation and absorbed longwave radiation) and losses (reflected and transmitted solar radiation and emitted longwave radiation). This 'budget' can be expressed mathematically as:

$$R_n = (1 - \alpha)S_t + L_d - \epsilon\sigma T_s^4 \qquad [1]$$

where $R_n$ is the net radiation at the surface expressed as the amount of energy (joules) per unit area (meters squared) per unit time (seconds) or watts per meter squared. All of the radiation terms ($S_t$, $L_d$, and $\epsilon\sigma T_s^4$) in eqn 1 must have the same units to avoid problems. As we expand these terms, $S_t$ is the solar radiation input, $\alpha$ the albedo or reflectance of the surface, $L_d$ the input of longwave radiation, $\epsilon\sigma T_s^4$ is the energy lost as longwave radiation from the surface. Shortwave radiation represents the energy in the 0.3- to 3-$\mu$m range, while longwave radiation is derived from wavelengths of more than 3 $\mu$m. Another way to consider this difference is that shortwave originates from the sun, with a surface temperature near 6000 K, while longwave is derived from objects with temperatures near 300 K. A representation of the radiation balance for two different days is shown in **Figure 1**. Radiation balance values throughout the day do not always produce a smooth curve because of the presence of clouds. Nighttime values are zero or below because the Earth's surface tends to lose more energy than it receives. Energy available as net radiation is further divided into other components. We express the energy balance as:

$$R_n = \mathrm{LE} + H + G \qquad [2]$$

where LE represents the energy used to evaporate water, $H$ the energy to heat the air, and $G$ the amount stored in the soil volume. The more energy input to the surface the greater the potential magnitude of these components. (*See* **Energy Balance.**)

### Incoming Solar Radiation

Solar radiation originates at the Sun and travels through space and the atmosphere before it reaches the Earth's surface. Solar radiation consists of two

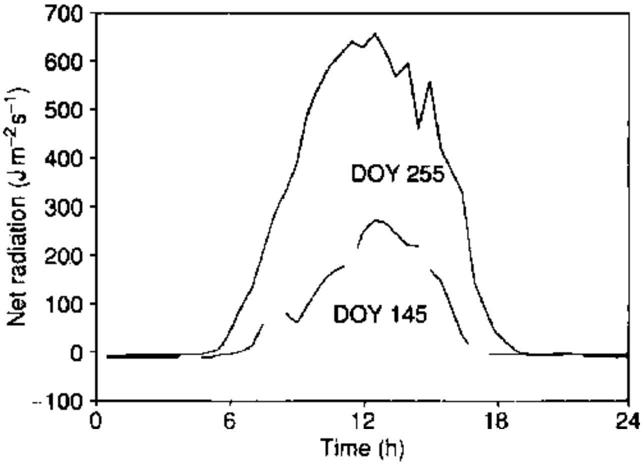


**Figure 1** Net radiation balance for 2 days (days 145 and 255) during the midsummer growing season in central Iowa.

components, the direct component ($S_b$) and the diffuse ($S_d$) component. Diffuse radiation comes from all angles, while the direct beam follows the path of the Sun throughout the day. Diffuse radiation is formed by the scattering of the direct beam by particles in the atmosphere. Physically, we can observe the effect of direct and diffuse radiation by sharpness of shadows. On a clear day, the outlines of the shadows are clearly defined, while on a cloudy day there are no shadows yet we still can see. Mathematically, we represent incoming solar radiation as:

$$S_t = S_b \sin\beta + S_d \qquad [3]$$

The angle of the Sun's beam above a horizontal surface, $\beta$, distributes the energy in the beam over a different area than if the surface were perpendicular to the beam. The magnitude of direct and diffuse radiation depends upon the solar elevation, the angle of the Sun's beam above the horizon, the amount of scattering and adsorption by atmospheric gases and aerosols, and the amount of scattering and adsorption by water droplets and ice particles in the clouds. The effect of clouds is easy to understand; however, particles in the atmosphere have a large impact, because they scatter radiation back to space and some toward the Earth's surface. The portion of diffuse radiation increases as the number of particles in the atmosphere increases.

There is large variation in the incident solar radiation at many locations throughout the year because of the changing angle of inclination of the Earth with respect to the Sun. At the equator there is little difference; however, at the midlatitudes there are large differences between summer and winter. Clear-sky solar radiation incident on the Earth at midday is approximately $26\,\mathrm{J\,m^{-2}}$ per day during summer, but approximately $9\,\mathrm{J\,m^{-2}}$ per day during winter at 42° N latitude. Clouds affect these totals and need to be considered. There are tables of clear-sky solar radiation incident on the Earth's surface for any location: these tables have been assembled through measurements and empirical calculations.

The slope of the surface is a critical factor in the radiation balance. If there is a south-facing slope in the northern hemisphere then more energy is going to be absorbed than if there is a north-facing slope. More subtle differences exist if the slope of the soil surface is changed with tillage implements. The creation of a furrow with a tillage tool that runs east to west across the field creates a surface in which one part will receive more direct solar radiation than the other. This will cause the soil surface and soil temperatures to be warmer on the south side and cooler on the north side in the northern hemisphere. Temperatures for bare soil at the surface under these conditions may be 10–20°C different at midday. This same furrow shape placed in a north–south direction will have differences in surface temperatures from morning to afternoon. The geometry of the surface plays a critical role in the radiation balance through the effect of direct-beam solar radiation. Manipulation of the soil surface through tillage implements can create a specific microclimate at the soil surface that influences temperature-dependent processes within the soil.

## Albedo

Albedo of a surface is the ratio of amount of solar radiation reflected back to the atmosphere to the amount of incoming solar radiation ($S_r$:$S_t$). We see a highly reflective surface as white, and a low-reflectance surface as black. A dark surface will absorb more radiation than a light-colored surface. The range of albedo values for different surfaces is shown in Table 1.

Albedo is more complex than assuming a single constant value for a surface. Albedo values for a soil surface change with soil water content, type of soil, tillage, amount and type of crop-residue cover, and age of residue. Starting with a smooth soil surface with an albedo of 0.15, the addition of water to wet the soil surface decreases the albedo to 0.10, and then there is a slow return to the original value as the surface dries. If the same surface is roughened, there is a further decrease in the albedo of a dry surface to near 0.10. This decrease is due to the change in the soil surface that traps more direct- and diffuse-beam solar radiation. Wetting this surface further decreases the albedo. Adding a fresh crop residue increases the albedo because of the highly reflective nature of the crop residue; however, as the residue ages, the albedo decreases and may be only slightly different from the soil surface in the spring. Soil surfaces are continually undergoing changes that affect their albedo. This will impact the radiation balance and ultimately the evaporation of water from the surface, heating the air, and warming the soil.

**Table 1** Albedo values for different natural surfaces

| *Surface* | *Albedo* |
|---|---|
| Bare soil, dry | 0.20 |
| Bare soil, wet | 0.15 |
| Bare soil, fresh corn residue | 0.35 |
| Bare soil, weathered corn residue | 0.22 |
| Corn canopy | 0.22 |
| Wheat canopy | 0.26 |
| Rice canopy | 0.24 |
| Deciduous forest with leaves | 0.18 |
| Coniferous forest | 0.16 |

The albedo values shown in **Table 1** represent an integrated value over the shortwave wavelengths and are determined by measured the incoming solar radiation with a pyranometer and reflected radiation with an inverted pyranometer of the same type. Within the solar radiation spectrum, there are differences in reflectance, as shown in **Figure 2** for a corn leaf. There is a dramatic change in the reflectance between the visible and near-infrared wavelengths from 0.6 to 0.7 μm. The difference in reflectance spectra across the wavelengths in plant material shows low reflectance values in the blue and red portions of the spectrum, an increase in the green wavelengths, and then a sharp increase in the near-infrared wavelengths. Reflectance values for a bare soil across the solar radiation wavelengths show a different pattern from a leaf and tend to increase linearly with wavelength, not showing the large differences among portions of the spectrum (**Figure 3**). Tilling the soil to change the roughness and increase the water content changes the reflectance of the soil. There is a decrease in the reflectance values for all wavelengths (**Figure 3**). Adding crop residue to the soil surface also changes the reflectance of the surface and changes the energy-balance components.

When solar radiation impinges on a leaf or any other object, the radiation may be reflected, absorbed, or transmitted. If we imagine ourselves underneath a thin layer of white cloth, from the upper side we would see the white color owing to reflection of all wavelengths; however, underneath the cloth we would still see owing to light being passed through (transmitted). Some of the light would be absorbed by the cloth. Under a 10-cm layer of soil, there would be no transmitted light and there would be no shortwave radiation, therefore no visibility. This same principle applies when layers of material are added to the soil surface. A layer with a high light transmittance allows solar radiation to reach the soil surface and warm the soil, evaporate water, or warm the air next to the soil surface. When a highly reflective and low-transmittance material is placed on the soil surface, there is little change in the temperature of the soil surface with varying solar radiation; hence snow affects the radiation balance. Transmittance values for different materials vary widely, as shown in **Table 2** for both the shortwave and longwave regions of the spectrum. Transmittance of solar radiation through materials, especially plant material, has a large effect on the radiation balance at the soil surface. Very dense canopies transmit some light through all of the layers of leaves to the soil surface. Humans are able to see in this region; however, a leaf receives little energy. Plants grown under other canopy layers often have very large leaves with less chlorophyll (more yellow) than the same plant grown in full sunlight. The radiation balance of these plants and the soil under dense canopies is much different from the environment without this canopy.

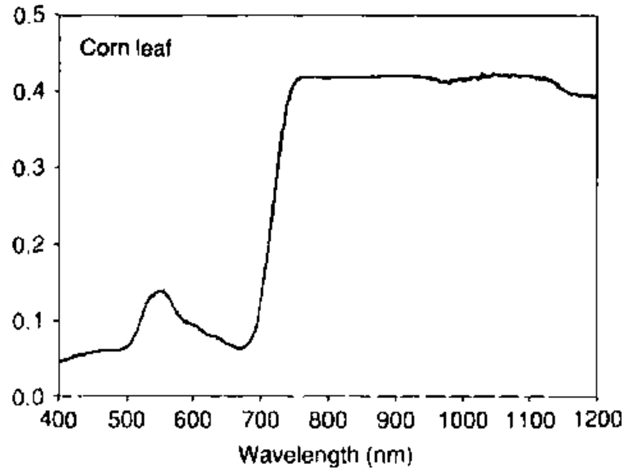


**Figure 2** Reflectance spectra for a corn leaf in the visible and near-infrared wavelengths.

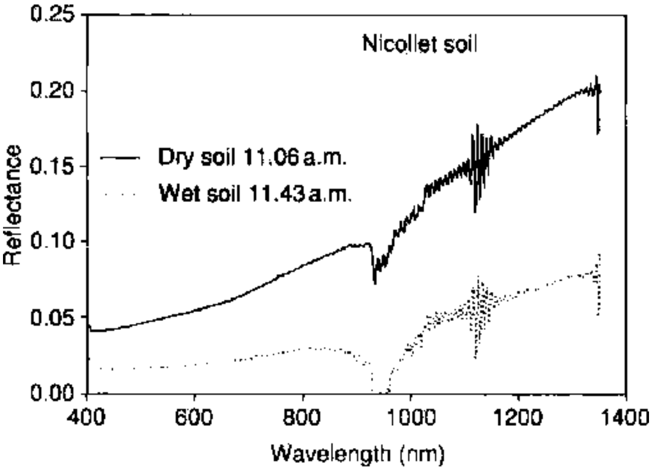


**Figure 3** Reflectance of solar radiation across visible and near-infrared wavelengths for a smooth, tilled, Nicollet soil when wet and dry.

**Table 2** Transmittance values for different natural surfaces

| *Surface* | *Transmittance* |
|---|---|
| Corn leaves, visible | 0.08 |
| Corn canopy, visible | 0.05 |
| Wheat leaves, visible | 0.07 |
| Wheat canopy, visible | 0.06 |
| Fresh corn residue, visible | 0.05 |
| Weathered corn residue, visible | 0.04 |
| White plastic, visible | 0.95 |
| Black plastic, visible | 0.0 |

## Incoming Longwave Radiation

All objects with an absolute temperature of more than 0 K emit energy. Objects in our atmosphere such as gases, particles, ice crystals, or water droplets have temperatures above absolute zero and emit energy. The energy emitted is equivalent to the 4th power of their temperature as described by the Stefan law, $E = es\ T^4$. The Stefan constant, $s$, is $5.67 \times 10^{-8}\ \mathrm{W\ m^{-2}\ K^{-1}}$. The amount of energy emitted by an object greatly increases as the temperature increases. The emissivity term can be considered an efficiency factor. The higher the emissivity of an object the more effective the emission of longwave radiation away from the object. Longwave radiation is measured with radiometers designed to capture only these wavelengths; however, there are relatively few of those instruments located around the world. Incoming longwave radiation is generally estimated from mathematical equations in which the emissivity of the atmosphere, $e_a$, is estimated through a combination of water vapor or temperature measurements to modify the energy relationship to $e_a s T_a^4$. In one of the original relationships for this estimation, it was proposed that $e_a = 0.51 + 0.66 e_0^{1/2}$, where $e_0$ is the water vapor pressure of the atmosphere. There have been several attempts to refine this relationship, for example, where $e_a$ is estimated by $0.575 e_0^{1/7}$ and $e_a$ is best represented by $0.70 + 5.95 \times 10^{-5} e_0 \exp(1500/T_0)$, with $T_0$ the air temperature in degrees Kelvin. Both $e_0$ and $T_0$ are obtained from meteorological station data at a height of 2 m above the surface and are available every hour throughout the day. A comparison of all of these different methods with data collected across the USA has found that there is little difference among the methods and all predicted longwave radiation to within 5% of the measured values. Emissivity values for clear skies can range from 0.6 to 0.75, and these equations apply to the clear-sky condition. When there are clouds, the $e_a$ term can be considered to be near 1. Consider what happens to air temperature on a clear versus a cloudy night. On the clear night there is a rapid drop in air temperature because the atmosphere is contributing little energy to the Earth's surface. On a cloudy night or an extremely humid night there is little change in air temperature because the longwave radiation lost from the Earth is captured by the clouds or water vapor and reemitted back to Earth's surface.

## Terrestrial Longwave Radiation

Objects on the Earth's surface have a temperature and emit energy in relationship to their temperature as discussed for longwave energy emitted from the atmosphere. The same principle applies to the soil surface or objects on the Earth's surface. The last term in eqn 1, $e_s s T_s^4$, represents the loss of energy away from the surface into the atmosphere. Emissivity values for objects on Earth are close to 1; however, there is variation, as shown in **Table 3**. The variation is small and often the assumption is that 0.95 or 0.97 are adequate mean values that can be used. The $T_s$ term in the relationship represents the surface temperature of the object. This would be the temperature at the outer surface of an object or at the very top layer of the soil surface. There are large gradients of temperature in the soil, and for a dry soil the surface temperature at midday may exceed 50°C while at 10 mm below the surface the temperature is less than 40°C. In the estimation of the longwave energy lost from the surface, a longwave radiometer can be inverted over the surface to measure the amount of energy escaping or estimated using measurements of surface temperature obtained from infrared thermometers combined with an estimate of emissivity for the surface. The latter approach works quite well for smooth or uniform surfaces; however, for rough or undulating surfaces or surfaces comprised of a combination of plants and soil, there is a large variation in surface temperature and measurements would have to be made that would represent the average value for the entire surface area. The magnitude of the energy lost from surfaces with different temperatures is shown by **Figure 4**. As the surface temperature increases, the energy emitted increases rapidly, because emission is a function of the 4th power of surface temperature.

**Table 3** Emissivity values for different natural surfaces

| *Surface* | *Emissivity* |
|---|---|
| Bare soil, dry | 0.97 |
| Bare soil, wet | 0.99 |
| Bare soil, fresh corn residue | 0.97 |
| Bare soil, weathered corn residue | 0.98 |
| Corn canopy | 0.94 |
| Wheat canopy | 0.96 |
| Deciduous forest, foliage present | 0.99 |
| Coniferous forest | 0.99 |

## Radiation Balance

Each of the terms in eqn [1] has been discussed separately in order to develop an appreciation for the variation within each component and the factors that affect the magnitude of the energy within each term. As these components are reassembled into the radiation balance for a surface it is possible to see how soil-management factors might affect the

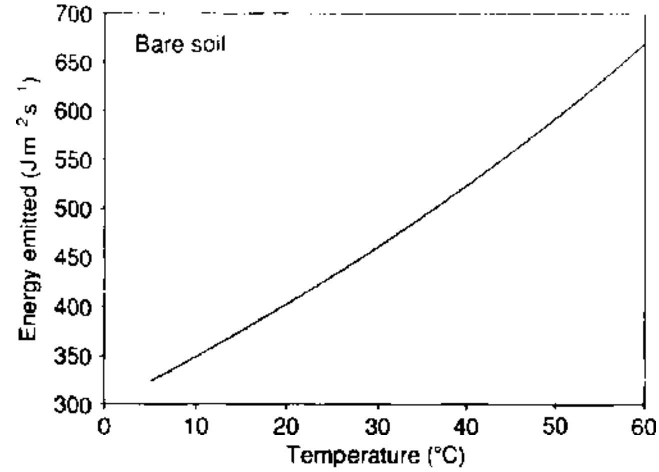


**Figure 4** Energy emitted from a soil surface under different surface temperatures.

radiation balance. There are changes throughout the day as shown in Figure 1. If the radiation balance is expanded into the components for each hour throughout the day, the differences that each component adds to the balance can be seen. The largest factor is the solar radiation term, and the difference between the two days is a direct result of the incoming solar radiation. The least amount of variation among days is introduced by the incoming longwave radiation, $L_d$, since this factor is primarily affected by air temperature. Longwave loss is actually highest during the day when the surface temperatures are highest. As these relationships are expanded into annual radiation balances, surface temperature and air temperature variations among days across the seasons have a large impact. If the annual radiation balance over the year is computed, the sum of the inputs and losses is nearly zero.

On a daily interval, changing the soil surface through tillage or crop-residue management has a large impact on the radiation balance. If a smooth, dry soil is tilled, there are several changes in the radiation balance components. Incoming solar and longwave radiation do not change. The albedo decreases so there is a greater adsorption of solar radiation leading to a greater amount of energy available at the soil surface. However, evaporation of soil water from the wet, rough surface decreases the surface temperature, creating a situation in which the surface emits less energy. The end result is a surface that has an increased amount of energy available than did the surface before tillage. This effect lasts only as long as there is soil water to evaporate, but there is some effect of the rough surface on reducing the albedo of a dry soil; however, this effect is small compared with the initial period with a wet soil surface.

The final example is the soil surface to which crop residue has been added. In the case of fresh crop residue there is an increase in the albedo of the surface and a reduction in net radiation. Weathered residue decreases the albedo and allows the surface to have an increased net radiation loading. Wetting and drying of the crop residue has some effect on the radiation balance because of the effect of water on both the albedo and emissivity of the surface. Soil-management practices have a large impact on the radiation balance, so understanding which factor is affected by management will help provide an improved understanding of the consequences of management decisions.

## Implications

The radiation balance of a surface represents the balance of shortwave gains and losses and longwave gains and losses. The radiation balance changes throughout the day and year owing to incoming solar radiation, which has the largest effect on the radiation balance, followed in importance by the cloud cover. The next major factor is the slope of the surface, which affects the amount of solar radiation absorbed. The slope factor can be both a large-scale (slope of hills and mountains) and microscale factor (creation of furrows and ridges with tillage implements). Humans have a large impact on the radiation balance and can manipulate the energy balance through all of the components in the radiation balance.

*See also:* **Climate Change Impacts**; **Crop Water Requirements**; **Crusts:** Structural; **Cultivation and Tillage**; **Energy Balance**; **Evaporation of Water from Bare Soil**; **Evapotranspiration**; **Remote Sensing:** Organic Matter; **Water-Use Efficiency**

## Further Reading

Hatfield JL, Reginato RJ, and Idso SB (1983) Comparison of long-wave radiation calculation methods over the United States. *Water Resources Research* 19: 285–288.

Monteith JL (1973) *Principles of Environmental Physics*. New York: Elsevier.

Rosenberg NJ (1974) *Microclimate: The Biological Environment*. New York: John Wiley.

## Radionuclides *See* Isotopes in Soil and Plant Investigations

## Rainfed Farming *See* Dryland Farming

# RANGE MANAGEMENT

**G L Anderson**, USDA Agricultural Research Service, Sidney, MT, USA

Published by Elsevier Ltd.

## Introduction

Proper rangeland management includes good stewardship of the rangeland biological components to preserve soil health and soil conservation. Certain management practices can also contribute to a host of soil-related problems (e.g., excessive wind and water erosion, nutrient loading, and increased sedimentation in reservoirs), ultimately leading to a general reduction in the land's ability to sustain healthy rangeland systems. Many of these problems can be avoided by the proper management of the aboveground or biological component of the range. While the concept of managing a healthy rangeland may seem simple and straightforward, different definitions of what constitutes a healthy rangeland result in divergent management approaches.

## What are Rangelands?

The natural evolution of different types of rangeland occurs under varying conditions and disturbances. Environmental factors such as drought, wind, and temperature affect the different types of plants that make up the rangeland plant communities found around the world. Other disturbances that influence the development of specific range types include: grazing (native animals and livestock), fire, weeds, and diseases. These disturbances have always been an important part of rangeland development and health.

In the USA a general example of these varying rangeland types includes: the presence of tall-grass prairies in the eastern Central Plains, mid-grass prairies in the Central Plains and Midwest, and short-grass prairies in the Midwest and West. Environmental conditions also result in varying amounts of shrubs, forbs, and trees. Rangeland in south Texas is dominated by trees, shrubs, and forbs; sagebrush dominates prairies in the west; and grasses dominate the northern and middle portions of the country.

Internationally, rangelands are divided into six broad categories: nonwoody grasslands, open shrublands, closed shrublands, savannas, woody savannas, and tundra. Africa is dominated by savannas with some open shrublands and closed shrublands, Australia is largely savannas along the coast and open shrublands in the interior, and Eurasia is dominated by nonwoody grasslands and open shrublands. All categories of rangeland are found in South America, except tundra, with the most diverse mixture of the different classes occurring along the lengthy western coast. Tundra is primarily limited to the extreme northern latitudes.

## How Much Area is Covered by Rangelands?

Worldwide, approximately 40.5% of terrestrial area is grasslands. This means that rangelands comprise more than 52.5 million km$^2$. This area can be further broken down into the different types of rangeland. Approximately 13.8% of terrestrial rangelands are woody savanna, 12.7% are open and closed shrublands, 8.3% are nonweedy grasslands, and 5.7% are tundra.

## What Determines How Rangelands are Managed?

Widely different environmental conditions and impacts of multiple disturbances make rangeland health difficult to define. One of the most influential factors behind range-management decisions is the manager's understanding of what comprises a healthy rangeland. There are essentially three major rangeland models or concepts that influence a rangeland manager's understanding and actions.

The climax model evaluates range condition and trend based on vegetation composition, plant production, ground cover, and soil erosion. By measuring and monitoring existing conditions, proponents of this approach hope to develop a meaningful index of rangeland health. The underlying assumption of this model is that a definable mixture of plant species composes the 'climax' or 'healthiest' vegetation type for a given ecosystem. Plant communities, over long periods of time, come to an equilibrium of predictable species based on the impact of human use, soil condition, geographic position, plant competition, animal use, microbial activity, burning, and climatic factors. Areas with minimal or no discernible impact by humans are often referred to as 'pristine' rangelands or rangelands that exemplify the purest, uncorrupted mixture of soils, microbes, plants, and naturally occurring animals, diseases, and pests.

While the climax model considers human use as a component associated with rangeland health, many people now view human activities (e.g., livestock grazing) as the primary cause of rangeland degradation. Therefore, any departure from the climax state is often attributed to misuse of the land by humans. An example of this type of thought underlies the beliefs of those proposing the development of 'buffalo commons,' which involve returning vast areas of US rangelands to a more 'natural' state, devoid of domestic grazing and other human uses. A synopsis of the climax model is: (1) We know what the composition of native plant and animal species should be, and (2) by limiting human influence the climax state can be maintained, and, given enough time, degraded ecosystems will ultimately return to a predetermined state of equilibrium.

A similar ecologic model of a healthy rangeland ecosystem also presumes a climax state exists, but that this state is rarely, if ever, achieved, because of natural and human impacts. The basic difference between this model and the previous one is the belief that nature often has as great, or greater, an impact on rangelands as do humans. In this model, plant community equilibrium is prevented by discrete occurrences of drought, fire, disease, pests, weeds, or human influences. But the basic idea that an ultimate climax state exists, remains. A synopsis of the ecologic model disturbance is: We know what the composition of native plant and animal species should be, but due to natural and human disturbance, this ideal state is rarely, if ever, achieved. However, the closer the system is to the expected equilibrium the healthier it is perceived to be.

The above two models presuppose that we know the basic combination of biotic and abiotic conditions that should exist in a given area. In reality we rarely, if ever, can know with any degree of certainty what a climax community should look like or even if climax in a grassland ever existed. This uncertainty has led to the development of the state and transition model, which presupposes that an ecologic system can transition (change) into several states (multiple types of vegetation communities), dependent on changing conditions that occur at critical junctures in a rangelands development. According to this model, the biological composition of rangelands can transition over time into a variety of different plant communities. Each transition can happen relatively quickly (e.g., from a grassland to a pine–juniper savanna) if the right environmental conditions occur, but once the change occurs it is very difficult to reverse. This means that some changes are ultimately permanent. Thus, this model recognizes that rangeland ecosystems are dynamic and changing, and that there is no 'one' combination of plants that is ecologically the 'right or ultimate' state.

An illustration can be found in an article written by Dr. Donald Jameson, who found that the combination of overgrazing followed by at least three wet years could result in the establishment of pines on rangelands previously devoid of the tree and that this change could ultimately be permanent and progressive without further human intervention. A synopsis of the state and disturbance model is: Many types of rangeland plant communities can develop in an area over time, and each can be 'healthy' and stable for long periods of time. However, the model takes into account that rangeland ecosystems can, and do, change when conditions are right.

In our brief discussion of what constitutes a healthy rangeland ecosystem, we have traversed from an absolute model (climax) to a more relative model of multiple steady states. How range managers approach their job is directly influenced by which concepts they use. As one progresses from one end of the ecologic spectrum to the other, a divergence of management approaches occurs. In the first instance (climax), managers view an ecologically defined set of specific plants as a healthy rangeland ecosystem, while at the other end (the transition model), managers expect to see any number of ecologic states where rangeland 'health' is not dependent on the existence of a specific type of plant community. In the former case, climax theory leaves the manager striving to achieve something that probably never existed on rangelands, while in the latter case the transition model incorporates change as a component of ecologic health. Under this conceptual model a range manager strives to achieve or maintain rangeland health; however, the ecologic definition of health depends on relatively 'recent' environmental conditions and the current mix of plants in the community.

## Biological Diversity

While divergent opinions on how best to manage rangelands often result in different impacts to the rangeland ecosystem, agreement generally exists on one key component: a healthy rangeland must maintain a mix of desirable (preferably native) plants that are self-sustaining. Biological diversity (or the number of different types of plants and organisms in the system) is an important consideration, because it provides food and habitat needs for different native and domestic grazers, birds, and the living microbial component of the soil.

The health of the belowground component of the plant and soil interface also depends on diversity (e.g., different plants using and influencing different portions of the ecosystem), a sustained rate of returning dead plant and animal material back into usable nutrients, maintenance of soil porosity (the rate at which water enters the soil), and water-holding capacity (the amount of water that the soil will retain for future use). For example, in the Northern Great Plains of the USA, most native rangeland plant species have fairly shallow roots, capable of capturing the limited amount of precipitation that occurs during the growing season. This shallow root mass is one of the primary zones of competition between plants, and it is the part of the biological component that minimizes soil loss. To a large degree, it is the soil–root interface that ultimately determines the stability and sustainability of the rangeland ecosystem. Therefore, it is here that changes in the aboveground plant community can have the greatest impact.

A common idiom in ecology is that 'nature abhors a vacuum.' When applied to this discussion, it simply means there are many environmental and soil conditions that exist across a rangeland, with different plants living and utilizing the resources in each of these areas or niches. If a specific plant group is missing (the vacuum), something will move in to take its place, and it is the range manager's job to ensure the niche is filled with the proper plant type. Unfortunately, more often than not, invading undesirable plants (weeds) exploit these areas and establish before management steps can successfully close the empty niches. Thus, the introduction of invasive weeds into the system can have a devastating impact on the soil–root interface, plant biodiversity, soil health, and finally the economic worth of rangelands.

## Introduction of Undesirable Plant Species

The impact of weeds on rangelands and range ecology is dramatic: First, native plants disappear, wind and water erosion are accelerated, and eventually the seedbank is altered to the point where there is no natural way for the system to recuperate. Second, weeds can also alter the soil chemistry so that little else is capable of growing in that location, and, finally, they often use excessive amounts of 'deep' water that give them a jumpstart in the early spring before native species begin their annual growth cycle.

The preponderance of weeds (undesirable plant species) that are proliferating across all types of rangelands is perhaps the single greatest biological threat to the maintenance of rangeland health today. Many of these undesirable plants have been introduced through several pathways: as contaminated seeds or plants used in landscaping, uncertified hay and landfill, and as unwelcome residents of improperly disposed waste material. While there are natural checks and balances for many of the native weeds present on rangelands, newly introduced species often propagate and disperse like a highly virulent disease, because existing natural control mechanisms are lacking. Consequently, many of these highly invasive species can establish in even the healthiest rangelands and quickly dominate the landscape.

One such plant is leafy spurge. Leafy spurge was introduced into the USA in 1827. The weed out-competes native vegetation for water in early spring and readily establishes in a wide range of climatic types. **Figure 1** shows leafy spurge stands dominating the riparian habitat in and around Devils Tower, Wyoming. The inset shows leafy spurge's ability to survive even in the harshest of areas (in this case growing out of a crack in a sandstone rock). Most invasive weeds are not used by domestic or native grazers. Limited grazing pressure increases a weed's ability to out-compete and native plants can result in monocultures that exist for decades.

Weeds impact most ecosystems; however, their impact on rangelands is usually much more pronounced for several reasons. First, the large area covered by rangelands, and their remoteness, may mean that invasive weeds are not discovered until they are well-established. Second, the individual cost to landowners, for combating weed infestations, is usually several times the economic return of the land. Finally, a failure to respond rapidly and aggressively to a weed infestation (either using herbicides or other control methods) leads to extensive problems that even most state and federal governments are financially unable to combat.

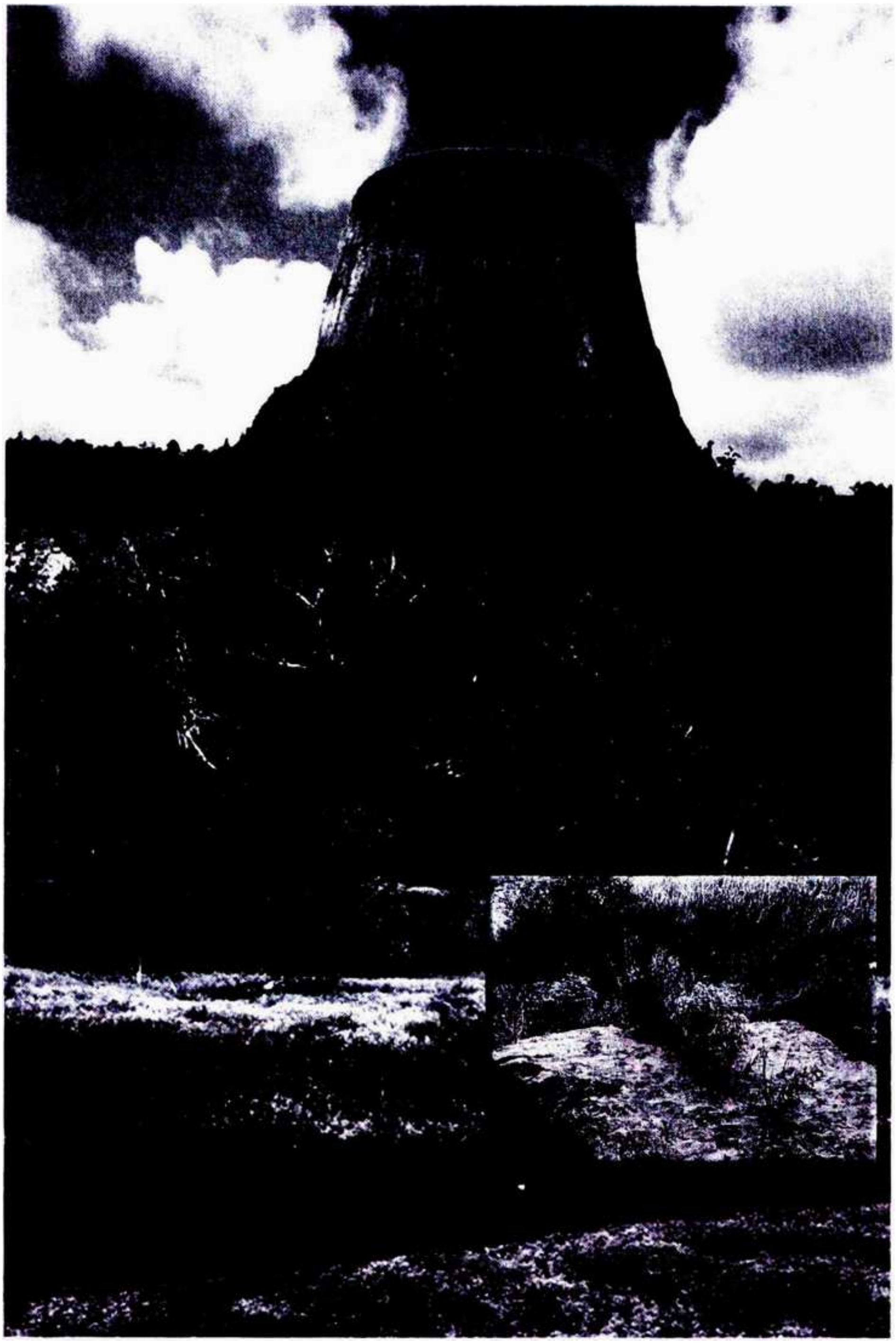


**Figure 1** (see color plate 53) Leafy spurge invasion near Devils Tower, Wyoming, USA. The yellow-green weed dominates the riparian flood zone and can even establish in the harshest of areas (inset, leafy spurge growing out of a sandstone rock).

Invasive and noxious weeds are expected to infest 57 million hectares of rangeland in the USA by the year 2010. When combined with the impact of other invasive species (not restricted to rangeland), the cost to the USA is estimated to be $137 billion per year. Given the rangeland's economic disadvantages and remoteness, range managers must prioritize the set of weeds they will attempt to manage and develop new approaches that are economically and ecologically sustainable. One innovative control method, refined since the mid-1990s, is to find natural enemies that occur where the weed originated (e.g., Europe, Asia, or the Middle East), ensure the natural enemy affects only the target plant, and establish and spread these populations across areas in North America infested by the weed. Secondly, managers need to look at additional control approaches (e.g., cultural, mechanical, and chemical) to aid the natural enemy and diminish further weed establishment. The scientific term for this type of control is 'biologically based integrated pest management,' because it combines natural enemies with other control options to provide affordable, ecologically sound, and sustainable weed control.

## The Role of Disturbance in Rangeland and Soil Health

Rangelands develop under conditions of periodic disturbance (e.g., drought, fire, grazing by domesticated animals and native wildlife, weed incursions, outbreaks of insect pests, disease, and human activities). These disturbances often help maintain rangeland ecosystem biological diversity and soil health over long periods of time. If the magnitude of the disturbance(s) is too great, the results can be devastating to the resource and result in permanent ecologic changes that are undesirable and frequently destabilizing to the point that the system may not recover.

Given that excessive disturbance can have an adverse effect on the rangeland ecosystem, managers generally accept that healthy rangelands depend on a certain amount of disturbance. However, the beneficial effects of certain disturbances are subject to debate, with livestock grazing being one of the most contentious. Many in the general public and some range managers believe rangeland health has been substantially degraded by livestock grazing. The idea is that healthier rangeland ecosystems will result from the elimination of livestock grazing. This was the rationale used to eliminate livestock grazing on many hectares of rangeland in the 1990s. But, the question remains: Is this policy effective in improving rangeland and soil health?

## Livestock Grazing

### No Domestic Grazing

Since some individuals attribute changes in rangelands to human influences, such as grazing, one solution often proposed is to do away with domestic grazing. While this may seem reasonable to those unfamiliar with the complexities and history of the rangeland ecosystem, the results are typically more detrimental to the resource than livestock grazing.

Grazing, even overgrazing for limited periods, has always been a part of rangeland ecosystems. The vast herds of buffalo and other native grazers migrating across the landscape no longer exist and the reestablishment of the dynamic interactions of native species that existed 100–200 years ago is difficult, if not impossible. While different native grazers may be introduced, historical movement of the herds throughout the landscape (typically memorized by the oldest member of the herd) has been lost. Physical boundaries, such as roads and fences, also preclude movement of larger native grazers and would concentrate native grazing much the same as they do for domestic livestock. Therefore, the reestablishment of native grazers is seldom effective over large areas and the cost–benefit ratio often precludes introduction altogether. The net result is the elimination or reduction of rangeland grazing pressure that was typical 200 years ago.

Given the limitations of reestablishing a natural symbiotic plant–animal interaction, several things happen to the rangeland resource if livestock grazing is entirely removed. One is the reduction in the turnover rate of nutrients from aboveground plants to the soil. This often stresses existing plants that grow in the nutrient-poor soil. As plants become stressed, they begin to lose root biomass. Soil compaction typically follows (reducing the ability of the soil to take up water) and dead, aboveground plant tissue can block incoming sunlight from living plants. There also tends to be an overall loss of biodiversity. The ultimate result is usually the dieback of a portion of the root system, an increased susceptibility to disease, and an increased amount of bare soil that results in unprotected areas (gaps) where weeds can readily establish and insect populations (e.g., grasshoppers that feed on the remaining plants) can reach high levels. While this situation may occur when the range is utilized for domestic grazing, it is nearly inevitable when rangelands are not managed at all.

Therefore, the removal of grazing or human management of native rangeland does not always result in a more natural or healthier rangeland. In fact, reduced human management often results in declining range health and a rapidly changing plant component that requires an inordinate amount of resources to maintain.

### Domestic Grazing

Grazing livestock on rangelands can have a negative impact when not managed properly. The existing body of scientific literature indicates that proper grazing is an important part of maintaining healthy rangelands. When rangelands are grazed using a proper grazing system (individual pastures are grazed at different times of the year to prevent overutilization of plants that prosper at different times of the year) and by only removing aboveground plant material to acceptable levels (leaving approximately 25% of the produced plant material), there is a positive impact on the soil–root interface. Roots continue to receive the photosynthetic materials needed to flourish and maintain a healthy mixture, many of the nutrients needed for plant growth are recycled quickly by the cattle and soil microbes back to the soil, and the stress of aboveground plant removal does not favor any single plant component (warm-season or cool-season plants). Rangelands flourish under proper grazing management. Rooting systems and species composition are maintained, soil structure and composition are protected from wind and water erosion, and the health of the system is more resilient to weed, insect, and disease impacts.

### Overgrazing

In this instance, the overutilization of the range due to excessive use or during periods of extended drought can worsen and ultimately cause increased soil erosion, soil compaction, and the permanent alteration of the soil seedbank and the rangeland plant community.

In any management application, there is always the possibility that the rangeland resource can be overutilized. However, the overwhelming majority of ranchers are good stewards of the range, more so because they choose to ranch as a lifestyle rather than for the earning potential of the job. To do that on an often limited income, they must maintain the quality of the rangeland needed for their livestock.

There are, however, times when even the best management plans can be taxed. Balancing grazing pressure during periods of drought, disease, or pest infestation is difficult. It is hard to predict how long adverse situations will last or how extensive problems will become. Therefore, there is often a lag-time between management changes and negative environmental impact. The result is, during periods of extended adverse conditions, management responses may not come quickly enough to prevent some damage to the rangeland resource.

What happens to rangelands when they are overgrazed for long periods or when improper management strategies are applied? Perhaps the most predominant impact of long-term overgrazing (or drought coupled with grazing) is the removal of more plant material than is recommended. For example, in the US Great Plains, the removal of more than 75% of aboveground production is considered detrimental to rangeland health. At this point there is not enough photosynthetically active plant material to provide the roots with the nutrients needed to maintain a healthy balance. The result is host dieback, and associated susceptibility to disease, repeated weeds and insect infestations.

## Summary

Maintaining or improving rangeland health is the primary concern of most range managers regardless of the ecologic concept they espouse. But the approaches they use to achieve that goal often diverge, particularly when a single rangeland concept or model is overemphasized. This is particularly true when the impact of periodic disturbances is considered.

Rangelands are dynamic and complex. Therefore our understanding and management strategies must also be dynamic to account for new knowledge. This is important for ensuring that rangeland health and soil health can be maintained for generations to come.

*See also:* **Grassland Soils**

## Further Reading

Devine R (1998) *Alien Invasion: America's Battle with Non-native Animals and Plants.* Washington, DC: National Geographic Society.

Halbach DW, Runge CF, and Larson WE (eds) (1987) *Making Soil and Water Conservation Work: Scientific and Policy Perspective.* Ankeny, IA: Soil Conservation Society of America.

Satterlund DR (1972) *Wildland Watershed Management.* New York: John Wiley.

Sheley RL and Petroff JK (eds) (1999) *Biology and Management of Noxious Rangeland Weeds.* Corvallis, OR: Oregon State University Press.

Stoddart LA, Smith AD, and Box TW (1955) *Range Management.* New York: McGraw-Hill.

Vallentine JF (1980) *Range Development and Improvement.* Provo, UT: Brigham Young University Press.

---

## Recycling of Organic Wastes *See* **Pollutants:** Biodegradation

---

## Redistribution *See* **Water Cycle**

# REDOX POTENTIAL

**R D DeLaune**, Louisiana State University, Baton Rouge, LA, USA
**K R Reddy**, University of Florida, Gainesville, FL, USA

## Introduction

Geochemists, soil scientists, and limnologists have used redox potential ($E_h$) measurements to characterize oxidation–reduction status of surface environments. The redox potential of soil, water, and marine systems is a measure of electrochemical potential or electron availability within these systems. Electrons are essential to all inorganic and organic chemical reactions. Redox potential measurements allow for rapid characterization of the degree of reduction and for predicting stability of various compounds that regulate nutrients and metal availability in soil and sediment. Redox potential is diagnostic for determining whether an area is functioning as wetland or nonwetland.

Oxidation and reduction reactions regulate many of the biogeochemical reactions in surface environments. Redox potential ($E_h$) is determined from the concentration of oxidants and reductants in the environment. The inorganic oxidants include oxygen, nitrate, nitrite, manganese, iron, sulfate, and $CO_2$, while the reductants include various organic substrates and reduced inorganic compounds.

## Theoretical Relationships

Oxidation and reduction reactions involve transfer of electrons from one compound to another and play a major role in regulating many reactions in biological systems. Oxidation–reduction reactions are coupled reactions.

Oxidation is defined as removal of electrons from a compound. Such a compound is usually referred to as 'electron donor' or 'reductant.' During this process it is oxidized and increases its oxidation number:

$$\text{Reductant} = \text{Oxidant} + \text{e}$$

Reduction is defined as the addition of electrons to a compound.

A compound that accepts electrons is usually referred to as 'electron acceptor' or 'oxidant.' During this process the compound is reduced and its oxidation number is decreased:

$$\text{Oxidant} = \text{Reductant} + \text{e}$$

The tendency of compounds to accept or donate electrons is expressed as 'reduction potential' or 'redox potential.' The redox potential of a substance depends upon:

- Affinity of molecules for electrons;
- Concentration of reductants and oxidants (referred to as redox pair);

Reductants (electron donors) in wetland soils are:

- Organic matter and various organic compounds;
- Reduced inorganic compounds such as $NH_4^+$, $Fe^{2+}$, $Mn^{2+}$, $S^{2-}$, $CH_4$, and $H_2$.

Oxidants (electron acceptors) are inorganic compounds such as $O_2$, $NO_3^-$, $MnO_2$, FeOOH, $SO_4^{2-}$, and $HCO_3^-$.

Anaerobic environments such as wetland soils are usually limited by electron acceptors and have an abundant supply of electron donors. Aerated soils are usually limited by electron donors and have an abundant supply of electron acceptors (primarily $O_2$).

## Redox–pH Relationships

Redox reaction in surface environments can be represented by the following half-cell reduction equation:

$$\text{Ox} + m\text{H}^+ + n\text{e}^- \rightarrow \text{Rd} \qquad [1]$$

where Ox is the oxidized component or electron acceptor, Rd is the reduced component or electron donor, $m$ is the number of hydrogen ions involved in the reaction, and $n$ is the number of electrons involved in the reaction. The electrons used in eqn [1] must be supplied by an accompanying half-cell oxidation reaction. In soils, organic matter ($CH_2O$) is the primary source of electrons. The oxidation reaction balances the reduction reaction. This is illustrated below with the reduction of $Fe(OH)_3$:

$$4Fe(OH)_3 + 12H^+ + 4e \rightarrow 4Fe^{2+} + 12H_2O\text{(reduction)}$$
$$CH_2O + H_2O \rightarrow CO_2 + 4H^+ + 4e \text{ (oxidation)}$$
$$4Fe(OH)_3 + CH_2O + 8H^+ \rightarrow 4Fe^{2+} + CO_2 + 11H_2O$$

The reduction reaction can be defined quantitatively through the change in Gibbs free energy ($\Delta G$):

$$\Delta G = \Delta G^0 + RT\ln\left(\frac{\text{Rd}}{(\text{Ox})(\text{H}^+)^m}\right) \qquad [2]$$

where $\Delta G^0$ is the standard free energy change, $R$ is the gas constant, and $T$ is absolute temperature. The Nernst equation expresses the reduction reaction in terms of electrochemical energy (millivolts) using the relationship $\Delta G = -nEF$

$$E_h = E^0 - \frac{RT}{NF}\ln\left(\frac{(\mathrm{Rd})}{(\mathrm{Ox})}\right) - \frac{mRT}{nF}\ln(\mathrm{H}^+) \qquad [3]$$

where $E_h$ is the electrode potential, $E^0$ is the standard half-cell potential, $F$ is the Faraday constant, $n$ is the number of electrons exchanged in the half-cell reaction, $m$ is the number of protons exchanged, and the activities of the various oxidized and reduced components are shown in parentheses. Substituting values of 8.31 J K$^{-1}$ mol$^{-1}$ for $R$, $9.65 \times 10^4$ C mol$^{-1}$ for $F$, and 298 K for $T$, and using the relationship $\ln(x) = 2.303 \log(x)$, eqn [3] simplifies to:

$$E_h(\mathrm{mV}) = E^0 - \frac{59}{n}\log\left(\frac{(\mathrm{Rd})}{(\mathrm{Ox})}\right) + 59\frac{m}{n}\mathrm{pH} \qquad [4]$$

Inspection of eqns [3] and [4] reveals that redox potential ($E_h$) increases with increasing activity of the oxidized component, decreases with increasing activity of the reduced component, and increases with an increase in $\mathrm{H}^+$ activity (or a decrease in pH).

Equation 4 shows the importance of pH in redox reactions. When the ratio of protons to electrons ($m/n$) is equal, there is a predicted 59 mV change in $E_h$ per pH unit. The $E_h$/pH slope can vary depending on the oxidants and reductants involved.

The $E_h$/pH slope predicted from the Nernst equation assumes that the redox couple controls the pH of the system. This is only true for pure chemical solutions. Measured slopes in natural systems due to mixed potential or various redox couples deviate from the predicted values. Applying the 59 mV correction factor cannot be used under those conditions.

The reduction reaction in eqn [1] can also be defined by an equilibrium constant:

$$K = \frac{(\mathrm{Rd})}{(\mathrm{Ox})(\mathrm{e}^-)^n(\mathrm{H}^+)^m} \qquad [5]$$

which can be rewritten in log form:

$$\log K = \log(\mathrm{Rd}) - \log(\mathrm{Ox}) - n\log(\mathrm{e}^-) - m\log(\mathrm{H}^+) \qquad [6]$$

The $-\log(\mathrm{e}^-)$ term in eqn [6] can be defined as pE (relative electron activity) in the same way pH is defined as $-\log(\mathrm{H}^+)$. Rearranging eqn [5] yields:

$$\mathrm{pE} + \mathrm{pH} = \log K - \log(\mathrm{Rd}) + \log(\mathrm{Ox}) \qquad [7]$$

Equation [7] can be used for plotting equilibrium solubility diagrams, because the pE + pH sum can be used as one axis and species activity as the other axis. $E_h$ is the more-common expression of soil redox potential, since it is readily determined with a platinum electrode. Soil pE can be calculated from $E_h$ using eqn [8]:

$$\mathrm{pE} = \frac{E_h(\mathrm{mV})}{59} \qquad [8]$$

## Normal Limits of pH and Redox Potential ($E_h$) in the Environment

Carbon dioxide from the atmosphere and organic acids formed by decaying organic matter are the two sources of acidity in most near-surface environments, resulting in acid pHs in the range 5–6. Lower pHs are found in acid sulfate soils, where values as low as 4.0 are sometimes recorded. As a result, a pH of 4 can be selected as the usual lower limit of pHs in natural environments (**Figure 1**).

The upper end of the pH limits is associated with $\mathrm{CO_2}$-free water in contact with carbonate rocks and can acquire a pH of 10 and, in contact with some silicates, a pH up to nearly 11. But since most surface waters emit $\mathrm{CO_2}$ to the atmosphere, such alkalinities are not attained. A reasonable upper limit of pH in most near-surface environments is near 9.0.

Oxygen is the strongest oxidizing agent commonly found in nature. Stronger agents than this cannot persist, because they react with water to liberate oxygen. Thus the upper limit of redox potentials is defined by the reaction:

$$\mathrm{H_2O} \Leftrightarrow \frac{1}{2}\mathrm{O_2} + 2\mathrm{H}^+ + 2\mathrm{e}^- \qquad E^0 = +1.23\ \mathrm{V}$$

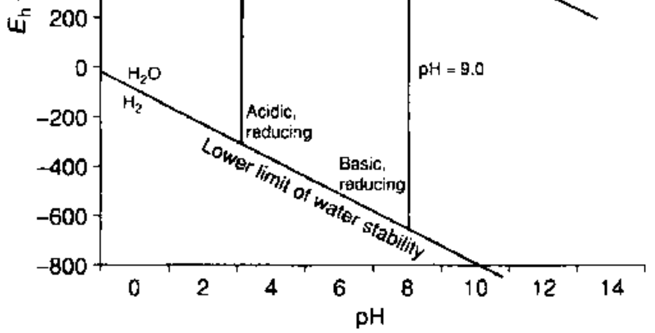


**Figure 1** The $E_h$–pH range found in soil surface environments. Adapted from Krauskoft KB (1967) *Introduction to Geochemistry*. New York, NY: McGraw Hill.

The potential of this half-reaction, which reflects the upper limitations of water stability, clearly depends on the pH, as shown by the equation:

$$E = +1.23 + 0.03\log[O_2]^{1/2}[H^+]^2$$

The lower limit of redox potential is that of the hydrogen reaction,

$$H_2 \Leftrightarrow 2H^+ + 2e^- \qquad E^0 = 0.00\,\mathrm{V}$$

For which

$$\begin{aligned} E &= 0.00 + 0.03\log[H^+]^2 - 0.03\log[H_2] \\ &= -0.059\,\mathrm{pH} - 0.03\log[H_2] \end{aligned}$$

Since the pressure of hydrogen in surface environments could not exceed 101 kPa, the maximum possible reducing potential in the presence of water would be:

$$E_h = -0.059\,\mathrm{pH} - 0.03\log(1) = -0.059\,\mathrm{pH}$$

## Aerobic and/or Anaerobic Redox Potential Range in the Environment

The $E_h$ scale used to quantify reduction intensity (or oxidation intensity) in chemical and biological systems is shown in **Figure 2**. The scale is in electromotive force units or volts (or millivolts) and can be divided into zones ranging from oxidized (where the aerobes function) to highly reduced (where the methane producers and sulfate reducers function).

Typically in wetland soils, $E_h$ values range from −300 to 700 mV, with a total range of approximately 1000 mV. In aerobic soils where the $O_2/H_2O$ redox couple functions, the $E_h$ range is between 300 and 700 mV.

## Redox Couples in Wetlands

Although several redox couples function in wetlands soils, the following are the most common reduction reactions involving a specific redox couple:

$$\begin{gathered} O_2 + 4e^- + 4H^+ \rightarrow 2H_2O \\ 2NO_3^- + 10e^- + 12H^+ \rightarrow N_2 + 6H_2O \\ MnO_2 + 2e^- + 4H^+ \rightarrow Mn^{2+} + 2H_2O \\ Fe(OH)_3 + e^- + 3H^+ \rightarrow Fe^{+2} + 3H_2O \\ SO_4^{2-} + 8e^- + 8H^+ \rightarrow S^{2-} + 4H_2O \\ CO_2 + 8e^- + 8H^+ \rightarrow CH_4 + 2H_2O \end{gathered}$$

The various inorganic redox systems found in soil and sediment become unstable at critical redox potentials (**Figure 3**). Sequentially following flooding, oxygen is reduced first, followed by nitrate and oxidized manganese compounds, and then ferric iron compounds. After the reduction of ferric iron, the next redox compound to become unstable is sulfate, followed by the reduction of carbon dioxide to methane.

Soil redox potential represents an indication of the oxidation–reduction status of the various redox couples. For example, a redox potential of 0 mV indicates that oxygen and nitrate are not likely to be present and that the bioreducible iron and manganese compounds are in a reduced state. At this same potential, however, sulfate is stable in the soil with no production of sulfide, which is toxic to plants. A redox potential of +400 mV indicates that oxygen may be present even though there may be excess water.

Redox potential measurements in pure systems can be used to predict the ionic distribution between chemical species which may interact with the transfer of electrons, such as ferrous and ferric iron or nitrite and nitrate nitrogen. However, in natural systems, there are many redox couples present, and not all redox couples are chemically interactive with others. Unless the concentration of a given redox couple is relatively high, inert electrodes (generally platinum) used for redox measurements are not specific for a specific redox couple. The redox electrode responds to the collective electrochemical potentials of all redox couples present.

The measured redox potential in soil is generally a mixed potential which reflects a weighted average of the potentials contributed by each of the redox couples present. As a result of the continuous

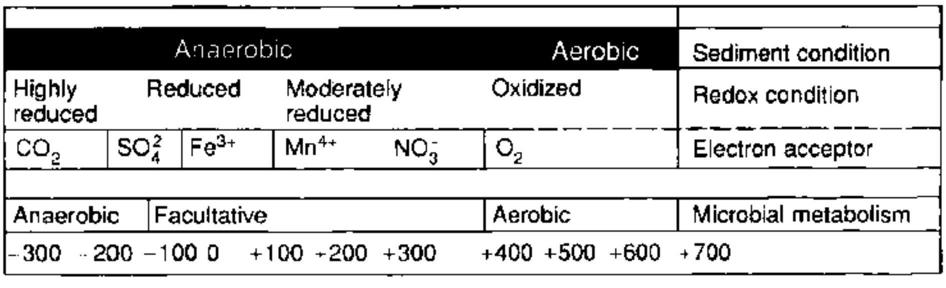


**Figure 2** Redox potential range in soil and sediment, showing the microbial metabolism process and electron acceptor.

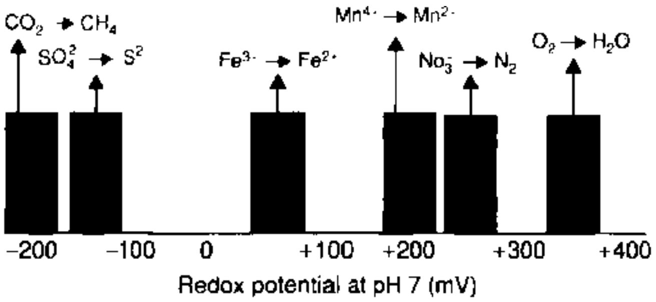


**Figure 3** Critical redox potential for the transformation of redox couples.

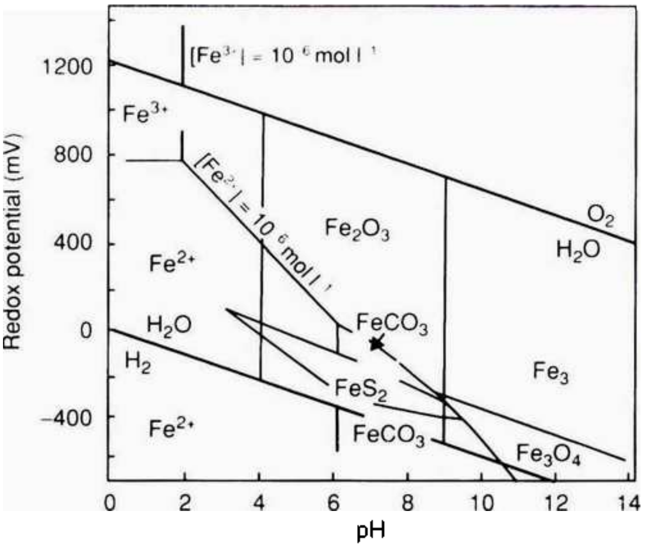


**Figure 4** Redox–pH effect on stability of iron minerals. Adapted from Krauskoft KB (1967) *Introduction to Geochemistry*. New York, NY: McGraw Hill.

addition of organic matter, which oxidizes and serves as an electron donor, a redox equilibrium is almost never attained in a natural system.

With the theoretical limitations involved in the use of redox potentials to quantitatively describe a specific ionic distribution in a mixed system, redox measurements can be successfully applied to characterize the oxidation–reduction transformations of many elements, including heavy metals and plant nutrients in soil.

## Predicting Mineral Stability

$E_h$ and pH can also be used for plotting stability fields for specific minerals. For example, the $E_h$–pH stability field for iron in **Figure 4** can be interpreted that hematite ($Fe_2O_3$) is a stable iron mineral under oxidizing environments. Under reducing environments, a stable mineral may be pyrite. If dissolved sulfur is low, siderite ($FeCO_3$) may be the predominant mineral under reducing conditions.

## Intensity and Capacity of Reduction

Redox potential measurement is only one measure of intensity of reduction in soils and sediment, as described below.

### Intensity

The ease of reduction is represented by the free energy of reduction or equivalent electromotive forces of the reaction (oxidation–reduction potential or redox potential).

The reduction processes that take place in a soil as a result of the oxidation of organic matter utilizing alternate electron acceptors of nitrate, manganic manganese, ferric iron, sulfate, and carbon dioxide do not proceed together. As indicated previously, nitrate is utilized first after oxygen disappears, followed by manganic manganese, ferric iron, sulfate, and finally carbon dioxide. For this reason the soil must be considerably more reduced for sulfate to be reduced than for nitrate to be reduced. If nitrate and sulfate and a suitable organic substrate are all present in a flooded, oxygen-deficient soil at the same time, the denitrifiers will reduce all of the nitrate before any of the sulfate is reduced by the sulfate reducers. The intensity of reduction is different for the two reduction processes. The soil is not very reduced (low reduction intensity or low electron pressure) for the reduction of nitrate, while it is very reduced during the reduction of sulfate (high reduction intensity or high electron pressure).

This intensity of reduction can be thought of in terms of the pressure of the electrons that the microorganisms need to dispense with as they carry out oxidation of the energy source. It does not require much electron pressure for the nitrate to accept electrons (and even less electron pressure for oxygen, if it is present, to acceptor electrons). This is why a large energy yield can be obtained from the oxidation of the energy source, since not so much of the possible energy from the oxidation process is utilized in forcing the electron acceptor to accept the electrons from the organic matter oxidation.

Electron acceptors that are difficult to reduce require more electron pressure to force the electrons to move to the electron acceptor. For example, sulfate reducers in soil and sediment can oxidize organic substrates, but a great deal of the energy derived from the oxidation is required to force the sulfate ion to accept the electrons and be converted to sulfide. This leaves less energy than the microorganism derives from the reaction that it can use for its energy needs. A great intensity of reduction is required for sulfate to accept the electrons produced during the oxidation of the energy source.

This approach to determine the intensity of reduction is called the oxidation–reduction or redox potential of the soil. Measurement of redox potential is conducted by placing a platinum electrode in the soil, connecting it to a suitable circuit that balances and measures the electron pressure from the reduction processes.

Reduction capacity described below differs from intensity of reduction.

### Capacity

The total amount of electrons accepted by oxidants or number and redox system undergoing reduction is the 'capacity.' The capacity can be best described in terms of its $O_2$ equivalent (not redox potential measurement).

The capacity factor of the various redox systems varies from one soil to another. The amount of oxygen present in the soil at the time of flooding is usually low. The amount of $NO_3^-$ present in the soil is dependent on soil organic matter and can be variable. Biologically reducible Mn can be present in soils at varying levels (less than 100 to 1000 mg kg$^{-1}$), while reducible Fe can be in the range of 500–3000 mg kg$^{-1}$ of soil.

Although the critical redox potential at which the inorganic redox systems become unstable provides valuable information, it does not provide any indication of the total capacity of the system to accept electrons or support microbial respiration. Capacity factor is equivalent to the total amount of electrons accepted by the oxidants in support of microbial respiratory activity. The redox couple present in lowest quantity has the smallest capacity for supporting microbial respiration. In mineral-flooded soils, reducible Mn and Fe primarily support the organic matter decomposition. In estuarine areas that receive sea water containing large amounts of $SO_4^{2-}$, the reduction of $SO_4^{2-}$ to $S^{2-}$ supports the microbial respiratory activities.

## Redox Condition and Metal/Metalloid Chemistry

Redox conditions of soil and sediment deposits also govern heavy metal chemistry (**Figure 5**). For example, chromium can exist in several oxidation states from Cr(0), the metallic form, to $Cr^{6+}$. The most stable oxidation states of chromium in the environment are $Cr^{3+}$ and $Cr^{6+}$. Besides the elemental metallic form, which is extensively used in alloys, chromium has three important valence forms: $Cr^{2+}$, $Cr^{3+}$, and $Cr^{6+}$.

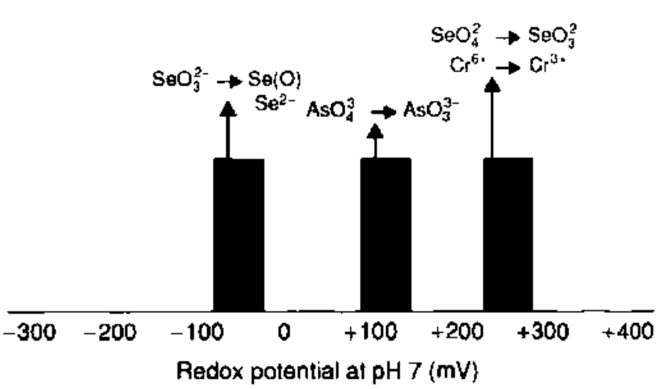


**Figure 5** Redox potential as related to metal and/or metalloid chemistry or transformation.

The trivalent chromic $Cr^{3+}$ and the tetravalent dichromate $Cr^{6+}$ are the most important forms in the environmental chemistry of waters and sediments. The presence of chromium $Cr^{6+}$ is of particular importance, because in this oxidation state Cr is water-soluble and extremely toxic. They are also the only forms that undergo valence changes in the $E_h$–pH ranges encountered in natural systems. The solubility and potential toxicity of the chromium that enters into aquatic or sediment columns are governed to a large extent by oxidation–reduction reactions. $Cr^{6+}$ reduces to $Cr^{3+}$ at approximately +300 mV (**Figure 5**).

Oxidation–reduction processes also play a role in arsenic (As) and selenium (Se) chemistry (**Figure 5**). For the pH and redox conditions in most soils and sediments, As exists as an oxyanion in the arsenate ($H_2AsO_4^-$, and $HAsO_4^{2-}$, $As^{5+}$) or arsenite [$H_3AsO_3$; $As^{3+}$] form. Monomethylarsenic acid (MMAA) and dimethylarsenic acid (DMAA) are important organoarsenic chemical species. Se occurs as selenate ($SeO_4^{2-}$, $Se^{6+}$), selenite ($HSeO_3^-$ and $SeO_3^{2-}$; $Se^{6+}$), and elemental selenium (Se(0)), but selenide (DMSe) is the most important. At higher soil redox levels (+200 to +500 mV), $As^{5+}$ is the predominant As species present. The reduction of $As^{5+}$ to $As^{3+}$ occurs at redox levels corresponding within the nitrate-reducing zone of soils characterized by a soil redox level of approximately +300 mV. When Fe reduction starts and soil redox levels drop below +50 mV, selenite is reduced to elemental Se or metal selenides. Reoxidation of reduced As and Se species occurs at similar soil redox levels to those identified for reduction reactions. Both microbial-mediated and chemical-oxidation processes are thought to be involved in the oxidation of reduced As and Se species.

## Measurement of Redox Potential

Inert metal electrodes (usually platinum) are widely used to measure the oxidation–reduction potentials of chemical systems. The inert metal electrode is used in conjunction with a reference electrode to form a complete cell. The most widely used reference electrodes for redox potential measurement are the calomel (mercury–mercurous chloride) and the silver/chloride electrode. The redox potential ($E_h$) is obtained from the potential of the cell ($E$) by adding the appropriate value for the reference electrode potential corrected for liquid junction effects. The potential measured with the platinum electrode and the secondary reference electrode must be corrected to give true $E_h$ or redox potential.

The platinum electrode referenced to a saturated calomel electrode will give $E_h$ or millivolt readings, provided corrections are made for the potential of the calomel electrode.

Performance of the platinum electrodes is checked by immersing the platinum electrode along with the reference calomel electrode in pH 4.00 and pH 7.00 buffer to which a small amount of quinhydrone has been added. The reading on the $E_h$ meter in millivolts should agree with that shown in **Table 1**. The meter reading is corrected relative to hydrogen by adding the calomel potential (**Table 1**). This value is the quinhydrone $E_h$ in millivolts relative to hydrogen. For example, a meter reading of 218 mV at 25°C for quinhydrone in pH 4.00 buffer gives an $E_h$ of 462.3 mV:

$$E_h = \text{meter reading} + \text{calomel potential}$$
$$E_h = 218.0 + 244.3 = 462.3\ \text{mV}$$

An electrode is considered satisfactory if it does not deviate from the above values by more than 5 mV.

When making soil $E_h$ measurements, the calomel potential must be added to the meter reading (millivolts). The procedure is as follows:

The potential is measured by inserting the electrodes in the substance to be measured and reading the millivolt scale on the pH meter. Since the calomel electrode is used as a reference, +244.3 mV is added to the recorded potential for reference to the hydrogen electrode.

**Table 1** Effect of temperature on the standard potential of a calomel half-cell and platinum electrode

| *Temperature (°C)* | *Calomel potential (mV)* | *Meter reading*[a] *pH 4 (mV)* | *pH 7 (mV)* |
|---|---|---|---|
| 5 | 257.0 | 236.0 | 66.5 |
| 10 | 253.9 | 231.6 | 60.1 |
| 15 | 250.8 | 227.3 | 53.7 |
| 20 | 247.3 | 222.7 | 47.3 |
| 25 | 244.3 | 218.0 | 40.8 |
| 30 | 240.9 | 213.4 | 34.3 |
| 35 | 237.5 | 208.5 | 27.7 |
| 40 | 234.1 | 203.5 | 21.0 |
| 45 | 230.6 | 198.4 | 14.2 |

[a] Meter reading with platinum and calomel electrode immersed in buffer saturated with quinhydrone.

## Further Reading

Bass Becking LGM, Kaplan IR, and Moore D (1960) Limits of the natural environment in terms of pH and oxidation–reduction potentials. *Journal of Geology* 68: 243–286.

Faulkner SP, Patrick WH Jr, and Gambrell RP (1989) Field techniques for measuring wetland soil parameters. *Soil Science Society of America Journal* 53: 883–890.

Garrels RM and Christ CM (1985) *Solutions, Minerals, and Equilibria*. New York: Freeman and Cooper.

Grenthe I, Stumm W, Lacksuharju M, Nilsson AC, and Wikbers P (1992) Redox potentials and redox reactions in deep groundwater systems. *Chemical Geology* 98: 131–150.

Patrick WH, Gambrell RP, and Faulkner SP (1996) Redox and measurement of Soils. In: Sparks DL *et al.* (eds) *Methods of Soil Analysis*, part 3, *Chemical Methods*, pp. 1255–1273. Soil Science Society of America Box Series No. 5. Madison, WI: Soil Science Society of America.

Ponnamperuma FN (1972) The chemistry of submerged soils. *Advances in Agronomy* 24: 29–96.

Reddy KR and Delaune R (2004) *Biogeochemistry of Wetlands*. CRC (in press).

# REDOX REACTIONS, KINETICS

**P S Nico**, California State University–Stanislaus, Turlock, CA, USA
**S Fendorf**, Stanford University, Stanford, CA USA

Oxidation and reduction (redox) reactions constitute a diverse, complex, and extremely important class of reactions in soils. Redox reactions influence properties, such as solubility and reactivity, of both organic and inorganic species/phases. For example, the two environmentally important oxidation states of Cr, Cr(III) and Cr(VI), have dramatically different properties. Cr(III) is sparingly soluble, tending to precipitate as Cr or Fe/Cr oxyhydroxides, while Cr(VI) exists as the soluble and quite mobile $CrO_4^{2-}$ oxyanion. Similarly, properties of fulvic and humic acids, including their contrasting solubilities and sorption behavior, are fundamentally a consequence of their oxidation state. Moreover, all biological life is dependent on redox reactions proceeding at disequilibrium.

Because of the dramatic effect that oxidation state can have on the properties of important compounds, soil chemists are often interested in being able to understand, model, and predict the redox behavior of a certain species under various conditions. We therefore need two kinds of information: thermodynamic and kinetic. Thermodynamics describes the overall path and time-independent energetics of a reaction. Kinetics describes the time-dependent behavior of a reaction. Thermodynamic data are generally more accessible; they answer unequivocally whether a reaction is 'favored' (i.e., possible) under certain conditions. However, reaction kinetics dictate the rate at which a reaction will proceed; reactions that do not proceed 'instantaneously' are often termed 'kinetically controlled.' For example, the oxidation of carbon as diamond to $CO_2$ by atmospheric $O_2$ is thermodynamically favored (by ~400 kJ/mol$^{-1}$) but proceeds appreciably only on geologic time scales. As we study kinetics, it is crucial to keep in mind that thermodynamic driving forces do not dictate reaction kinetics.

The vast majority of the redox reactions of interest to the soil chemist are kinetically controlled, and all life depends on kinetic constraints maintaining disequilibrium. Because of this, the study of redox kinetics is of central importance to understanding almost any soils system. Kinetic data are usually combined into a mathematical description of the reaction kinetics known as the kinetic rate law. A complete rate law contains terms for all the variables in a system that affect the time-course of the reaction. These factors include the chemical species involved in the reaction as well as other factors such as temperature, pressure, or ionic strength. The rate law also contains a constant that is an indication of the reaction's relative speed, known as the rate constant. The initial form and some of the important terms in a rate law can occasionally be deduced from equilibrium data or from a theoretical description of the reaction, but complete rate laws, other than the simplest gas phase reactions, are experimentally determined. Rate laws are essential for a variety of purposes, including modeling experimental data, understanding the detailed molecular steps involved in the reaction (the reaction mechanism) or, most importantly, to model and predict actual field behavior.

## Kinetic Rate Laws

We mathematically define the rate of a reaction in terms of how quickly a given reactant is consumed or product is formed. In order to do this we must already have good equilibrium data on the reaction that tells us the reactants, the products, and the stoichiometric relationships between these species. For example, a reaction like that described in eqn [1], where the reactants $A$ and $B$ combine to form products $C$ and $D$, and the stoichiomteric coefficients of each reactant or product, are given by a, b, c, or d respectively.

$$\mathrm{a}A + \mathrm{b}B \rightleftarrows \mathrm{c}C + \mathrm{d}D \qquad [1]$$

The rate can be defined based on any of the reactants or products as follows, where square brackets denote the concentration (and not activities) of a species.

$$\text{Rate} = -\frac{1}{\mathrm{a}}\frac{\mathrm{d}[A]}{\mathrm{d}t} = -\frac{1}{\mathrm{b}}\frac{\mathrm{d}[B]}{\mathrm{d}t} = \frac{1}{\mathrm{c}}\frac{\mathrm{d}[C]}{\mathrm{d}t} = \frac{1}{\mathrm{d}}\frac{\mathrm{d}[D]}{\mathrm{d}t} \qquad [2]$$

Eqn [2] is valid as long as the stoichiometric coefficients in eqn [1] are accurate under the given conditions (i.e., there are no side reactions that consume or produce species within the reaction). For cases involving side reactions, the proportionality of the reaction/products would not depend solely on the coefficients in eqn [1].

The rate as defined in eqn [2] is the difference between the rate of the forward reaction (eqn [1] read from left to right, which will be a function of, among other things, the concentration of $A$ and $B$) and the reverse reaction (eqn [1] read from right to left, which will be a function of the concentrations of $C$ and $D$).

$$\text{Rate}_{\text{net}} = \text{Rate}_{\text{forward}}([A], [B]) - \text{Rate}_{\text{reverse}}([C], [D]) \qquad [3]$$

Equilibrium is the point where the concentrations of reactions and products remain constant, or, in other words, that $\text{Rate}_{\text{net}}$ is zero. We can see from eqn [3] that equilibrium does not mean that eqn [1] has ceased to occur but simply that the $\text{Rate}_{\text{forward}}$ is equal to the $\text{Rate}_{\text{reverse}}$. As long as eqn [1] is far from equilibrium, we can neglect the rate of the reverse reaction and concentrate solely on the forward reaction. As stated already, the $\text{Rate}_{\text{forward}}$ will be a function of the concentrations of $A$ and $B$, known as the rate law. The exact nature of this function must be determined experimentally, but for many reactions the rate law takes the form shown in eqn [4], where $[A]$ and $[B]$ are the concentrations of reactants $A$ and $B$; concentration units of molarity (M or $\text{mol l}^{-1}$) are usually used for dissolved species. Solid species can also be reported in terms of molarity or in mass ($\text{mg l}^{-1}$) or surface area ($\text{m}^2\ \text{l}^{-1}$) concentrations. The constant $k$ is an experimentally determined proportionality factor known as the rate constant. The constants $\alpha$ and $\beta$ are also experimentally determined and are denoted as the reaction order. If $\alpha$ is equal to 1 and $\beta$ is equal to 2, then the reaction is said to be first-order in $A$, second-order in B, and third-order overall. The units of $k$ are dictated by the reaction order such that the units of the rate will come out appropriately; usually rates are reported in units of concentration per time, such as $\text{mol l}^{-1}\ \text{s}^{-1}$.

$$\text{Rate}_{\text{forward}} = k[A]^{\alpha}[B]^{\beta} \qquad [4]$$

If eqn [1] was an elementary reaction, the coefficients in eqn [1] (a and b) would be equal to the $\alpha$ and $\beta$ of eqn [4]. Unfortunately, most soil systems are sufficiently complex that we must deal in overall reactions, meaning reactions whose reactants, products, and stoichiometry have been determined macroscopically. In this case, there is no theoretical link between the coefficients in eqn [1] and the exponents in eqn [4]. In fact, the exponents in eqn [4] need not even be whole numbers.

## Integrated Rate Laws

Remember that eqn [4] is in fact a differential equation of the form shown in eqn [5]. If we were able to integrate eqn [5] we would be able to calculate the concentrations of $A$ or $B$ at any given time.

$$\text{Rate}_{\text{forward}} = -\frac{1}{\text{a}}\frac{\text{d}[A]}{\text{d}t} = -\frac{1}{\text{b}}\frac{\text{d}[B]}{\text{d}t} = k[A]^{\alpha}[B]^{\beta} \qquad [5]$$

The integration of a general equation like eqn [5] is not trivial. For this reason, let us first consider a simpler system. Eqn [6] shows the rate law for a reaction that depends only on the concentration of $A$ and is first-order in $A$.

$$\text{Rate} = -\frac{\text{d}[A]}{\text{d}T} = k[A] \qquad [6]$$

Eqn [6] is easily integrated to yield eqn [7], where $[A]_0$ is the initial concentration of $A$, $[A]_t$ is the concentration of $A$ at any time t, and $k$ is the first-order rate constant.

$$[A]_t = [A]_0 e^{-kt} \qquad [7]$$

The integrated rate laws for zero- and second-order, single-species rate laws are shown in Table 1.

**Table 1** Integrated rate law expressions

| *Reaction order* | *Rate law* | *Integrated rate law* | *Units of k* | *Linear plot* |
|---|---|---|---|---|
| 0 | $-\frac{\text{d}[A]}{\text{d}t} = k$ [8] | $[A]_t = [A]_0 - kt$ [9] | $\text{mol l}^{-1}\text{s}^{-1}$ | $[A]_t$ versus $t$ |
| 1 | $-\frac{\text{d}[A]}{\text{d}t} = k[A]$ [10] | $[A]_t = [A]_0 e^{-kt}$ [11] | $\text{s}^{-1}$ | $\ln[A]_t$ versus $t$ |
| 2 | $-\frac{\text{d}[A]}{\text{d}t} = k[A]^2$ [12] | $\frac{1}{[A]_t} = \frac{1}{[A]_0} + kt$ [13] | $\text{mol}^{-1}\ \text{l s}^{-1}$ | $\frac{1}{[A]_t}$ versus $t$ |
| 2 | $-\frac{\text{d}[A]}{\text{d}t} = k[A][B]$ [14] | $\frac{1}{[B]_0 - [A]_0}\ln\frac{[A]_0[B]_t}{[B]_0[A]_t} = kt$ [15]<br>$\frac{1}{[A]_t} = \frac{1}{[A]_0} + bkt$ [16] | $\text{mol}^{-1}\ \text{l s}^{-1}$ | |
| $\alpha + \beta + \gamma$ | $-\frac{1}{\text{a}}\frac{\text{d}[A]}{\text{d}t} = k[A]^{\alpha}[B]^{\beta}[C]^{\gamma}$ [17] | None | $\text{mol}^{1-(\alpha+\beta+\gamma)}\ \text{l}^{(\alpha+\beta+\gamma)-1}\ \text{s}^{-1}$ | |

Reactions whose rates are controlled solely by one species and are exactly zero; first-, or second-order do exist, and sometimes the order of a reaction can be identified solely by collecting concentration data over time for a single reactant and comparing to one of the simple rate laws. As seen in Table 1, if the reaction is first-order in $A$ then a plot of the $\ln[A]_t$ versus $t$ should be a straight line. Similarly, if the reaction is second-order in $[A]$, then a plot of $[A]_t^{-1}$ versus $t$ should be a straight line.

Unfortunately, the majority of reactions of interest do not fall neatly into one of the easily integrateable rate laws, and it is more common that we must use mathematics or experimental design to simplify a more complex rate law, like eqn [17]. The most common of these techniques are discussed below, but a clever experimentalist or mathematician can always devise new approaches.

## Initial Rates

The technique of initial rates can be used to determine the order of an individual reactant within a complex rate expression. The initial rate is the linear reaction rate at the beginning of a reaction before a significant amount (<10%) of any of the reactants has been consumed and before there is any appreciable reverse reaction. This is expressed mathematically in eqn [18].

$$\text{Rate}_{\text{initial}} = -\frac{1}{a}\frac{d[A]}{dt} \approx -\frac{1}{a}\frac{\Delta[A]}{\Delta t} \approx k[A]^{\alpha}[B]^{\beta}[C]^{\chi} \quad [18]$$

For example, if we wanted to determine $\alpha$, the order of $A$ in eqn [17], we need only to measure the initial rate at two different $A$ concentrations. We can then take the ratio of these two rates as shown in eqn [19].

$$\frac{\text{Rate}_1}{\text{Rate}_2} = \frac{k[A_1]^{\alpha}[B]^{\beta}[C]^{\chi}}{k[A_2]^{\alpha}[B]^{\beta}[C]^{\chi}} \quad [19]$$

Since the rate constant, $k$, and the concentration of $B$ and $C$ should be equal in both equations, eqn [19] reduces to eqn [20].

$$\frac{\text{Rate}_1}{\text{Rate}_2} = \frac{[A_1]^{\alpha}}{[A_2]^{\alpha}} = \left(\frac{[A_1]}{[A_2]}\right)^{\alpha} \quad [20]$$

$\alpha$ can be determined by taking the log of both sides and rearranging, to give eqn [21].

$$\alpha = \frac{\text{Log}\left(\frac{\text{Rate}_1}{\text{Rate}_2}\right)}{\text{Log}\left(\frac{[A_1]}{[A_2]}\right)} \quad [21]$$

This process can be repeated to determine the order of any number of reactants. Because these are reaction orders for a complex overall reaction, $\alpha$, $\beta$, and $\chi$ need not be whole numbers.

Although it is true that in theory one needs only to determine the initial rate at two different concentrations of $A$, it is good practice to employ several different concentrations to make sure that the orders calculated with any pair of $A$ concentrations are the same (i.e., the reaction rate is determined at varying concentrations of $A$ while all other concentrations are maintained at the same value). In fact, one can rearrange eqn [18] to give eqn [22]. Using this equation a plot of $\log(\text{Rate}_{\text{initial}})$ versus $\log[A_{\text{initial}}]$ should give a straight line with slope $\alpha$. The analogous equation can be used to determine the remaining reaction orders.

$$\text{Log}(\text{Rate}_{\text{initial}}) = \alpha \log[A_{\text{initial}}] + \log(k[B]^{\beta}[C]^{\chi}) \quad [22]$$

Once the orders of all the various reactants have been determined, the rate constant can be calculated simply by rearranging any of the expressions of the form of eqn [17] into eqn [23].

$$k = \frac{\text{Rate}_{\text{initial}}}{[A]^{\alpha}[B]^{\beta}[C]^{\chi}} \quad [23]$$

The rate constant should be the same, or very similar. The rate constants should also be equal (within experimental error) or the reaction is incorrectly modeled and the calculated orders are not valid.

The initial rate approach was used to determine the orders for the reactants involved in the reduction of birnessite ($MnO_{1.7}$) to Mn(II) by catechol ($C_6H_6(OH)_2$). According to the overall reaction equation (eqn [24]), the general rate expression should contain three factors, birnessite, catechol, and $H^+$, according to eqn [25].

$$\begin{aligned} &MnO_{1.7(s)} + 0.7C_6H_6(OH)_{2(aq)} + 2H^+ \\ &\quad \rightarrow Mn(II)_{(aq)} + 0.7C_6H_4O_{2(aq)} + 1.7H_2O \quad [24] \end{aligned}$$

$$\frac{d[Mn(II)]}{dt} = k[MnO_{1.7}]^{\alpha}[C_6H_6(OH)_2]^{\beta}[H^+]^{\chi} \quad [25]$$

Each reactant was varied independently and the initial rates were measured. Plots of initial rates versus $\log[C_6H_6(OH)_2]$ or $\log[MnO_{1.7}]$ were linear, with slopes equal to one indicating that the reaction was first-order in both catechol and birnessite. On the other hand, plots of initial rates versus $\log[H^+]$ were flat, i.e., slope of zero, indicating a zero-order dependence on $H^+$ concentration. This information allowed the rate law for this reaction to be simplified to eqn [26].

$$\frac{d[\mathrm{Mn(II)}]}{dt} = k[\mathrm{MnO_4^-}][\mathrm{C_6H_6(OH)_2}] \qquad [26]$$

## Pseudo *n*th-Order Reactions

Some reactions are so rapid as to preclude the collection of initial rate data. In these situations, we can reduce higher-order reactions to 'pseudo' lower-order reactions by adjusting the concentrations of the various reactants; this is often referred to as the method of isolation. Returning to eqn [17], if we set the concentration of $B$ and $C$ such that they are much greater than the concentration of $A$, we can assume, after taking into account appropriate stoichiometries, that $A$ can be entirely consumed without significantly altering the concentration of $B$ or $C$. Thus, eqn [17] reduces to eqn [27], where $k'$ is the 'pseudo $n$ order rate constant.' ($k'$ is also frequently denoted $k_{\mathrm{obs}}$ because it is the observed (measured) rate constant for the specific experimental conditions employed.)

$$-\frac{1}{a}\frac{d[A]}{dt} = k'[A]^n \qquad [27]$$

Using this approach, we can reduce a rate expression of any order into one with only one or two reactants of concern and determine if it conforms to one of the simple rate expressions in **Table 1**, in other words that $n$ is ≅ 0, 1, or 2. This was the case for the reduction of carbamate pesticides by Fe(II). The initially postulated rate expression for this process was eqn [28], but by monitoring the reduction of the carbamate pesticide in the presence of large excesses of Fe(II), eqn [28] simplifies to eqn [29].

$$-\frac{d[\mathrm{pesticide}]}{dt} = k[\mathrm{Fe(II)}][\mathrm{pesticide}] \qquad [28]$$

$$-\frac{d[\mathrm{pesticide}]}{dt} = k_{\mathrm{obs}}[\mathrm{pesticide}] \qquad [29]$$

Plots of ln[pesticide] versus time were found to be linear, indicating that eqns [29] and [28] are indeed first-order with respect to pesticide concentration.

## Second-Order, Two-Reactant Rate Laws

If we suspect that we a have a rate law in the form of eqn [14], we can model the data and determine the rate constant, $k$, in one of two ways. If it is relatively simple to measure both $A$ and $B$ during the course of reaction, it is possible to fit the data to eqn [15] using $k$ as a fitting parameter. Alternatively, if it is difficult to measure $A$ and $B$ simultaneously, but we know the stoichiometric relationship between $A$ and $B$, we can use that relationship to simplify eqn [15]. We can substitute an expression of the form $[B]_t = [B]_0 - ab^{-1}[A]_t$ into eqn [15], where $a$ and $b$ are the stoichiometric coefficients for $A$ and $B$ analogous to those in eqn [1]. Or, more elegantly, we can design our reaction such that $A$ and $B$ are initially present in their stoichiometric ratios according to eqn [30].

$$\frac{[A_0]}{[B_0]} = \frac{a}{b} \qquad [30]$$

When eqn [30] is valid, then the integrated form of eqn [14] is simply eqn [16]. This expression is similar to eqn [13] except for the stoichimetric factor $b$.

This approach was employed to determine the rate constant for the reduction of Cr(VI) by Fe(II). The rate expression for this reaction was postulated to be first-order in each reactant and second-order overall (eqn [31]). Three Fe(II) ions are necessary to provide the three electrons needed to reduce one Cr(VI) to Cr(III). This leads to an Fe(II) to Cr(VI) stoichiometry of 1:3, respectively. By controlling the initial reaction conditions such that $[\mathrm{Fe(II)_0}] = 3[\mathrm{Cr(VI)_0}]$, the researchers were able to use eqn [32] to fit their data and obtain a value for $k$.

$$-\frac{d[\mathrm{Cr(VI)}]}{dt} = k[\mathrm{Fe(II)}][\mathrm{Cr(VI)}] \qquad [31]$$

$$\frac{1}{[\mathrm{Cr(VI)}]_t} = \frac{1}{[\mathrm{Cr(VI)}]_0} + 3kt \qquad [32]$$

## Reaction Diagrams and Activation Energies

The progress of a reaction can be represented as in **Figure 1**, in which the $y$-axis represents energy and the $x$-axis is what is known as the reaction coordinate, and is simply a way of showing the progress of the reaction from reactants to products. **Figure 1** represents what is known as a reaction diagram; this can be a very effective and concise way of summarizing the energetics of a reaction. For a reaction to be thermodynamically favored, the total energy of the products must be less than the total energy of the reactants, or, in other words, the products must be more stable than the reactants. The energy barrier shown between the reactants and the products is known as the activation energy ($E_a$). The magnitude of this barrier (as represented by its height in **Figure 1**) determines the rate of the reaction. Reactions with smaller activation

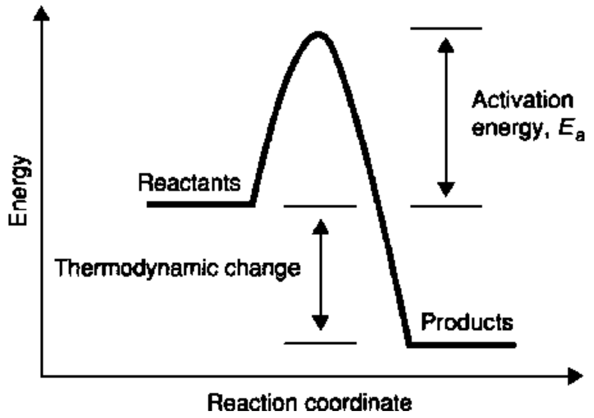


**Figure 1** Reaction diagram for generalized thermodynamically favored but activation energy-limited reaction.

energies will proceed rapidly from reactants to products as compared with those with larger activation energies. This is the reason why reactions can be thermodynamically favored but kinetically unimportant (owing to huge activation energies).

Reactants gain the energy necessary to overcome the activation energy by colliding with each other. For reactions with very small activation energies, reactants will form products virtually every time there is a collision. These reactions are known as diffusion- or transport-controlled reactions because the rate of reaction is only limited by how quickly the reactants can approach one another through diffusion. At normal temperatures, homogeneous aqueous diffusion-controlled reactions are too rapid to be of much interest to soil chemists. On the other hand, for heterogeneous reactions, in which reactants must diffuse toward a surface or into and out of pore spaces, rates of diffusion are much slower. Under these conditions it is possible to have a relatively slow, but diffusion-controlled reaction. Diffusion-controlled reactions generally obey rate laws of a different form than the ones discussed here.

For reactions with larger activation energies, only a small percentage of collisions will have enough energy to push the reactants over the activation energy barrier. For these reactions, it is the activation energy that determines the magnitude of the rate constant, $k$. The mathematical relationship between the activation energy and the rate constant was developed by Savante Arrhenius and is known as the Arrhenius equation (eqn [33]), where $E_a$ is the molar activation energy in $\mathrm{J\,mol^{-1}}$, $R$ is the universal gas constant ($8.3145\,\mathrm{J\,K^{-1}\,mol^{-1}}$), $T$ is temperature in K, and $A$ is a proportionality constant known as the preexponential factor.

$$k = Ae^{-\frac{E_a}{RT}} \qquad [33]$$

Many people are qualitatively familiar with the observation that most reactions tend to proceed more rapidly at higher temperatures, but it is the Arrhenius equation that allows this temperature dependence to be quantified.

The activation energy of a reaction can be determined by measuring $k$ at several different temperatures and using the Arrhenius equation. A plot of $\ln(k)$ versus $T^{-1}$ should yield a straight line with slope, $-E_a R^{-1}$ according to eqn [34], which is obtained simply by taking the natural log of both sides of eqn [33] and rearranging.

$$\mathrm{Ln}(k) = \mathrm{Ln}(A) - \frac{E_a}{\mathrm{RT}} \qquad [34]$$

Similarly, two temperature points can be used to calculate $E_a$ according to eqn [35].

$$E_a = \frac{\mathrm{Ln}(k_1) - \mathrm{Ln}(k_2)}{\left(\frac{1}{T_2} - \frac{1}{T_1}\right)} \qquad [35]$$

Because both of these approaches are simple rearrangements of the Arrhenius equation, they are theoretically equivalent. Once the activation energy has been determined we can model and predict the temperature dependence of a reaction. The magnitude of $E_a$ can help determine whether the reaction is diffusion-controlled or activation energy-controlled. Diffusion-controlled reactions tend to have an $E_a$ less than $\sim 40\,\mathrm{kJ\,mol^{-1}}$.

## Reaction Mechanism (Rate-Limiting Step and Preequilibria)

Understanding reaction kinetics has additional benefits beyond simply being able to model the time-dependent progress of a given reaction. The kinetics of a reaction can often be used to make conclusions regarding the mechanism of that reaction, although this must be done with caution. The mechanism of a reaction is the series of elementary reactions that combine to give the complete reaction. With complex soil chemical reactions, it is rarely, if ever, possible to break a reaction mechanism into all of its elementary reactions. This being said, all metal redox reaction must progress through three stages: (1) formation of a prereaction complex; (2) electron transfer; and (3) separation of postreaction complex. In organic redox reactions, the second step often involves the transfer of an entire atom or group of atoms instead of simply an electron, but the fundamental process is similar.

The individual steps that make up a reaction mechanism usually have different activation energies and therefore different rates. It is unlikely that for a given redox reaction each of the three stages outlined above will have the same rate constant. In fact, it is often the case that the activation energy of one of the steps will be much greater than that of the others and therefore will be much slower. The slowest step is known as the rate-limiting or rate-determining step (it is the 'bottle-neck' in the reaction progression). If the rate-determining step is the first step in a reaction mechanism, none of the following reactions or intermediate species will appear in the rate law. This effect can be seen in the reduction of carbon tetrachloride (CCl$_4$) by surface-absorbed Fe(II); the overall reaction is as seen in eqn [36].

$$CCl_4 + 2Fe(II) + H^+ \rightarrow CHCl_3 + Cl^- + 2Fe(III) \qquad [36]$$

$$\frac{d[CHCl_3]}{dt} = k[CCl_4][Fe(II)] \qquad [37]$$

The experimentally determined rate expression for this reaction was found to be first-order in carbon tetrachloride, second-order in Fe(II), and zero-order in H$^+$ (eqn [37]). The absence of H$^+$ in the rate law can be explained by postulating the following mechanism, in which eqn [38] is assumed to be the rate-limiting step.

$$CCl_{4(sorbed)} + 2Fe(II)_{(sorbed)} \rightarrow CCl^-_{3(aq)} + Cl + 2Fe(III)_{(sorbed)} \qquad [38]$$

$$CCl_{3^-(aq)} + H^+_{(aq)} \rightarrow CHCl_{3(aq)} \qquad [39]$$

However, if there are fast reactions before the rate-determining step, then these reactions will appear in the rate law. Because these early reactions are so much faster – both forward and reverse reactions – than the rate-limiting step, we can assume that they always remain at equilibrium throughout the course of the reaction. Acid–base reactions and metal speciation reactions are generally rapid and are frequently treated as preequilibria. Similarly, surface absorption can sometimes be treated as a preequilibrium, as was done in the case of the oxidation of fulvic acid by an Mn-oxide. The proposed reaction mechanism involved rapid absorption of the fulvic acid on to the Mn-oxide followed by a slow electron transfer step (eqns [40] and [41]), where Mn≡ denotes a Mn-oxide surface site.

$$Fulvic_{(aq)} + Mn{\equiv}OH \rightarrow Mn{\equiv}Fulvic(rapid) \qquad [40]$$

$$Mn{\equiv}Fulvic \rightarrow Mn(II) + oxidized\ Fulvic\ (slow) \qquad [41]$$

If eqn [41] is the rate-limiting step then the rate equation for this process should be like that seen in eqns [42], and [40] can be treated as an equilibrium expression (eqn [43]).

$$\frac{d[Mn(II)]}{dt} = k_{41}[Mn{\equiv}Fulvic] \qquad [42]$$

$$K_{40} = \frac{[Mn{\equiv}Fulvic]}{[Fulvic][Mn{\equiv}OH]} \qquad [43]$$

Combining eqns [42] and [43] and the equality that the total oxide surface sites must be equal to the free sites plus the fulvic occupied sites, $[Mn{\equiv}_{tot}] = [Mn{\equiv}OH] + [Mn{\equiv}fulvic]$, it was possible to obtain a rate expression that matched the data (eqn [44]).

$$\frac{d[Mn(II)]}{dt} = k_{41}K_{40}[Mn\equiv_{tot}]\frac{[Fulvic]}{1 + K_{abc}[Fulvic]} \qquad [44]$$

Although this equation may initially seem more complicated, it is practically speaking much simpler because it is based on total reactant concentrations that are much easier to measure.

## Biologically Controlled Redox Reactions

Many important redox transformations in soils are controlled by enzymes produced by microorganisms. Enzymatically controlled reactions are frequently modeled using either Michaelis–Menten kinetics or a derivative of Michaelis–Menten kinetics known as Monod kinetics. The proposed enzyme-controlled reaction mechanism used to develop Michaelis–Menten kinetics is shown in eqns [45] and [46], where E is the enzyme, S is the metal or organic molecule of interest known as the substrate, ES is the enzyme substrate complex, and P is the product.

$$E + S \rightleftarrows ES \qquad [45]$$

$$ES \rightarrow P\ and\ E \qquad [46]$$

To convert this qualitative mechanism into a quantitative rate law, we must make an assumption known as the steady-state assumption. This assumption is similar but not identical to the preequilibrium situation described above. In the steady-state assumption, we assume that certain species are present at relatively constant values throughout the course of the reaction, although not necessarily their equilibrium values. To develop a steady-state situation we must assume that the substrate binds rapidly and reversibly to the enzyme – eqn [45] is rapid in both the forward and reverse directions. Next it is assumed that the transformation of the enzyme substrate complex into the product (eqn [46]) is a first-order process and is the rate-limiting step in the overall reaction (eqn [48]). These assumptions allow us to assume that ES is present in a steady state or, in other words, that the rate of ES formation is equal to the rate of ES destruction. Eqn [47] shows this mathematically where the equation for ES formation is shown on the left and is set equal to the sum of the equations for ES destruction.

$$k_{45}[\mathrm{E}][\mathrm{S}] = k_{-45}[\mathrm{ES}] + k_{46}[\mathrm{ES}] \qquad [47]$$

$$\mathrm{Rate} = \frac{d[\mathrm{P}]}{dt} = k_{46}[\mathrm{ES}] \qquad [48]$$

If we mathematically combine these two assumptions by substituting in for ES, we derive the basic Michaelis–Menten equation (eqn [49]).

$$\mathrm{Rate} = \frac{k_{222}[\mathrm{E}]_{\mathrm{tot}}[\mathrm{S}]_{\mathrm{tot}}}{K_{\mathrm{m}} + [\mathrm{S}]_{\mathrm{tot}}} \qquad [49]$$

$K_{\mathrm{m}}$ is called the Michaelis–Menten or the Michaelis constant and $[\mathrm{E}]_{\mathrm{tot}}$ and $[\mathrm{S}]_{\mathrm{tot}}$ are the total enzyme and substrate concentrations, respectively.

The development of Monod kinetics is similar to that for Michaelis–Menten, with the insertion of $X$ as the total cell concentration and $K_s$ as the half saturation constant.

$$\mathrm{Rate} = \frac{k[X][\mathrm{S}]_{\mathrm{tot}}}{K_s + [\mathrm{S}]_{\mathrm{tot}}} \qquad [50]$$

Because it is frequently more feasible to control/measure cell concentration than enzyme concentration, Monod kinetics often have a practical advantage over Michaelis–Menten kinetics.

## Conclusion

It is important to remind ourselves of primary assumptions made throughout this chapter. First, we have almost exclusively dealt with reactions that are far from equilibrium and were therefore able to ignore reverse reactions. For systems closer to equilibrium these reactions must be included. Although this complicates the mathematics, the general approach for developing the rate laws remains the same. Second, because these rate laws are empirically derived relationships, they are only rigorously applicable to systems with reactant concentrations, temperatures, ionic strengths, and other variables within the ranges used experimentally to develop the rate laws. They can sometimes be applied to slightly different systems, but this must always be done with caution. Lastly, there are many more detailed and complex rate laws that have been developed for both abiotic and biotic redox reactions, but they are beyond the scope of this chapter. The interested reader is referred to the further reading as a place to start.

## List of Technical Nomenclature

| | |
|---|---|
| $X$ | Concentration of X (mol $\mathrm{L}^{-1}$) |

*See also:* **Anaerobic Soils; Redox Potential**

## Further Reading

Amonette JE, Workmann DJ, Kennedy DW, Fruchter JS, and Gorby YA (2000) Dechlorination of carbon tetrachloride by Fe(II) associated with goethite. *Environmental Science Technology* 34: 4606–4613.

Buerge IJ and Hug SJ (1999) Influence of mineral surfaces on chromium(VI) reduction by iron(II). *Environmental Science Technology* 33: 4285–4291.

Lasaga AC and Kirkpatrick RJ (eds) (1981) Kinetics of geochemical processes. In: *Reviews in Mineralogy*, vol. 8. Washington, DC: Mineralogical Society of America.

Matocha CJ, Sparks DL, Amonette JE, and Kukkadapu RK (2001) Kinetics and mechanism of birnessite reduction by catechol. *Soil Science Society of America Journal* 65: 58–66.

Mortimer RG (1993) *Physical Chemistry*. Redwood City, CA: Benjamin/Cummings.

Sparks DL (1989) *Kinetics of Soil Chemical Processes*. San Diego, CA: Academic Press.

Sparks DL (1995) *Environmental Soil Chemistry*. San Diego, CA: Academic Press.

Sposito G (1994) *Chemical Equilibria and Kinetics on Soils*. New York: Oxford University Press.

Strathmann TJ and Stone AT (2002) Reduction of the pesticide oxamyl and methomyl by Fe(II): effect of pH and inorganic ligands. *Environmental Science Technology* 36: 653–661.

Waite TE, Wrigley IC, and Szymczak R (1988) Photoassisted dissolution of a collodial manganese oxide in the presence of fulvic acid. *Environmental Science Technology* 22: 778–785.

# REMEDIATION OF POLLUTED SOILS

**E Lombi and R E Hamon**, CSIRO Land and Water, Glen Osmond, SA, Australia

## Introduction

Public concern about environmental pollution has increased substantially since the beginning of the twentieth century with growing awareness of the stresses being placed on the environment due to activities arising from both industrial activity and intensive agriculture, and as a consequence of high-profile incidents that have highlighted the risks of pollution for human health. Examples of such incidents include industrial pollution at Love Canal, USA, the chemical factory explosion at Bhopal, India, and the nuclear near-accident at Three Mile Island, USA. Pollutants tend to have longer residence times in soil than in water or air. The soil acts as a sink or a filter in which pollutants are accumulated rapidly but are depleted slowly. The decontamination of polluted soils is becoming an issue of societal and political interest, because of its importance for environmental protection and human health, and its economical relevance. New technologies based on physical, chemical, and biological processes to reclaim polluted soils have been developed during the last decades. This article briefly describes the available technologies for the remediation of organic and inorganic contaminants in soils. A synopsis of these technologies is depicted in **Figure 1**.

### Containment Versus Clean-Up Strategies

In terms of managing soil pollution, there are two types of strategy. The first, which at the beginning of the twenty-first century is the preferred option for many regulatory agencies, has the remediation goal expressed as a maximum allowable total concentration of pollutant in the soil. To comply with this approach, the total concentration of pollutant in soil must be lowered and this can be achieved using the group of technologies termed 'clean-up' technologies (**Figure 1**). The second strategy does not require a reduction in the total concentration of a pollutant in soil, but aims to manage exposure to the pollutant, and hence environmental and health risks, through technologies that lower pollutant mobility and bioavailability. These are termed 'containment' technologies (**Figure 1**).

### *In Situ* Versus *Ex Situ* Strategies

*In situ* remediation refers to a soil-treatment regime in which the soil remains confined to its original location throughout the remediation process. Some earth-moving activities may be required in order to ensure complete exposure of the soil to the remedial treatment. Protocols that minimize off-site transport of dusts, vapors, or contaminated liquors are critical. *In situ* strategies have advantages in that costs, and environmental and health risks associated with excavation and off-site transport of polluted soil are reduced. However *ex situ* strategies (transport of

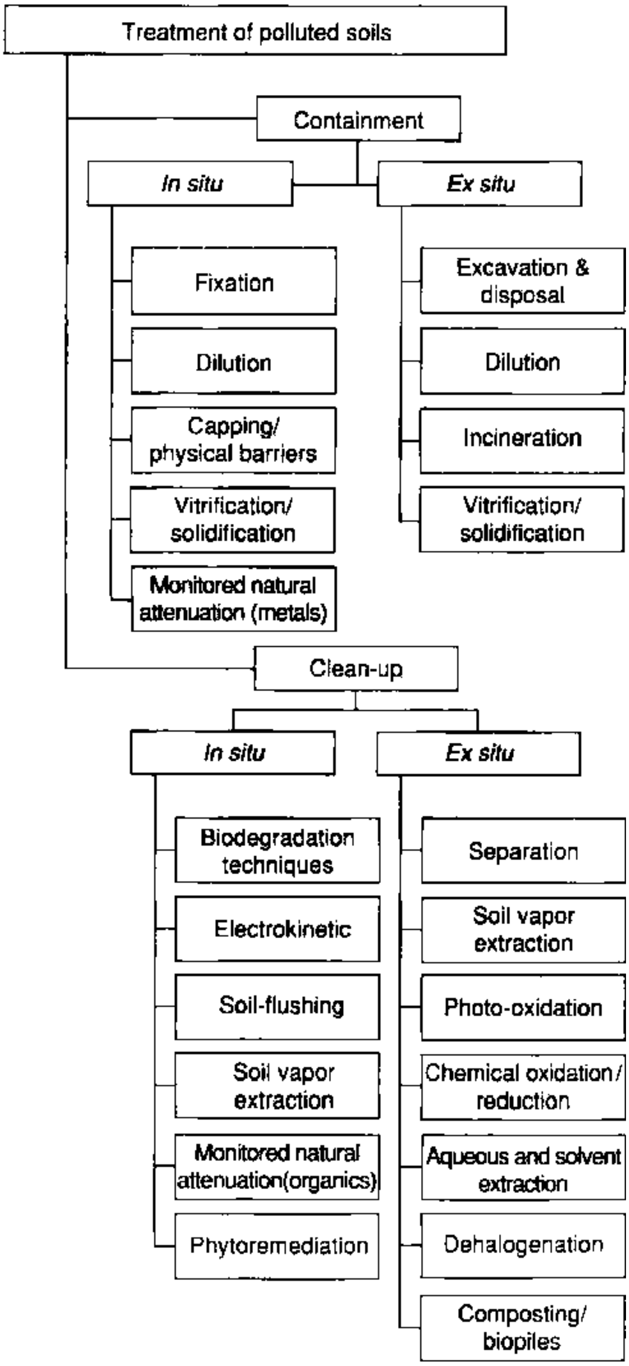


**Figure 1** Decision tree of the alternatives available to treat polluted soils.

polluted soils to an external location) may be unavoidable because of impracticalities in establishing essential treatment infrastructure at the polluted site, or due to the site being positioned in a high-risk area.

## Physical Treatment Technologies

### Excavation and Off-site Disposal (*Ex Situ*, Requires Ongoing Monitoring)

Applicable to both inorganic and organic pollutants, excavation and off-site disposal entails removing the polluted soil from a high-value or vulnerable site to a disposal facility where potential pollutant exposure routes can be more effectively managed. Some pretreatment of the soil may be required to reduce the concentration of readily mobile pollutants in order to meet disposal-facility restrictions. Precautions associated with the excavation and transport of polluted soils need to be taken. This is a rapid method for managing polluted sites which has been widely employed in the past but potentially just relocates the problem and hence is not a remediation treatment *per se*. With landfill space at a premium, leading to rapidly rising disposal costs, this strategy is becoming less acceptable to industry, regulatory bodies, and the public. Ongoing monitoring of the disposal facility to ensure no off-site movement of contaminants is essential.

### Dilution (*In Situ* or *Ex Situ*)

Applicable to both inorganic pollutants, including nutrients, and organic pollutants, dilution of soil pollution can be achieved through soil-mixing, deep tillage, or fracturing. It may be effective in reducing the potential contaminant exposure dose at sites in which pollutants are present in low levels or in which contamination is present in small, dispersed 'hot spots.' This method increases the volume of contaminated soil but may be appropriate for management of soils that have been polluted by nutrient oversupply (e.g., soils receiving dairy/piggery effluents). However, dilution is unlikely to meet with regulatory or community acceptance with respect to other types of pollutants.

### Capping and Encapsulation (*In Situ*, Requires Ongoing Monitoring)

Applicable to both inorganic and organic pollutants, containment technologies aim to eliminate exposure of humans and the environment to the polluted soil. Lateral or vertical movement of the contaminant (due to water percolation, erosion, or plant root acquisition) is eliminated through the construction of highly reactive and/or impermeable subsurface barriers and surface caps. Both inorganic and organic pollutants may be managed using the monitored natural attenuation strategy. However, as with excavation and off-site disposal, the soil is not cleaned of pollutants, so this is not a permanent remediation solution. Barriers may fail as a result of weathering and/or chemical degradation, settlement, cracking, erosion, poor drainage away from the surface cap, or penetration by deep-rooted vegetation. There is therefore a need for ongoing monitoring of the site to ensure that the barriers and caps have not been breached. Introduction of barriers constructed of materials that either sorb the contaminant or chemically react with it can provide an extra degree of protection against inadvertent contaminant migration. Mobile contaminants, which might otherwise escape through ruptures in the barrier, may instead diffuse toward it and be chemically immobilized.

### Separation–Fractionation (*Ex Situ*)

Applicable to both inorganic pollutants, including radionuclides, and organic pollutants, separation and fractionation is based on the principle that contaminants are frequently associated with specific particle-size fractions within soils and uses sieving or sedimentation to separate and remove the contaminated fraction from excavated soils. For example, the most reactive components of soils are generally found within the clay-size fraction ($<2\ \mu m$) due to the high relative surface area of this fraction. Contaminants entering the soil in soluble forms therefore typically partition to, and hence are concentrated in, the clay fraction. Separation of this fraction from the soil produces a highly polluted residue which requires further treatment (e.g., see Incineration, below) or disposal. However, the benefit is that the mass of material requiring further treatment is significantly reduced compared with the original soil. Contaminants may also be present in the form of discrete particles (e.g., lead (Pb) pellets from firing ranges), allowing for removal from the soil using physical sieving techniques. Immiscible organic pollutants may also be isolated from the soil using gravity separation methods.

### Electrokinetic Separation (*In Situ*)

Applicable to both inorganic and polar organic pollutants, electrokinetic technology uses the generation of a low-intensity current between electrodes placed in the soil to mobilize and concentrate anionic and cationic pollutants toward the anode and cathode, respectively. The contaminants are then recovered through processes such as electroplating, precipitation, and coprecipitation at the electrode, concentration on ion-exchange resins or physical extraction of the aqueous phase. Soil types amenable to this treatment are those of low permeability (high clay content) which are poorly drained. This is a new technology, but limited pilot trials show promising results as of 2003.

### Soil-Flushing (*In Situ*)

Applicable to both inorganic pollutants, including radionuclides, and organic pollutants, in soil-flushing an aqueous solution (water or water containing contaminant-solubilizing additives) is circulated through the contaminated soil. The pollutant is desorbed from the soil and retrieved from the fluid. The operation usually takes the form of a continuous cycle, with the purified fluid being reinjected back into the contaminated soil, circulated, reextracted, and repurified. This technology performs poorly in soils with low permeability and with pollutants that partition strongly to the soil surface. For soils in which the pollutant has been present for an extended period of time, natural attenuation processes can also markedly reduce pollutant mobility, requiring many flushing cycles to be completed before pollutant concentrations are significantly reduced. These factors can make application of this technology impractical in terms of the timescales needed to complete remediation and associated costs.

### Vitrification (*In Situ* or *Ex Situ*, Requires Ongoing Monitoring)

Applicable primarily to inorganic pollutants, especially radionuclides, in vitrification an electric current is used to melt the soil at temperatures up to 2000°C. The melted soil cools to form a chemically stable, leach-resistant, glass-like solid which encases and entraps inorganic pollutants. Organic pollutants are destroyed through pyrolysis at the high temperatures attained. However, this procedure destroys the soil, thereby limiting future site use. Adoption of this method is likely to be limited to inorganic contaminants under extreme cases where pollutant risks are very high, such as may be the case for radionuclide-polluted soils. Given its use for high-risk pollutants, it is important to institute an ongoing monitoring protocol for the solidified mass to ensure that contaminants are not lost over time through leaching or chemical weathering processes.

### Solidification (*In Situ* or *Ex Situ*, Requires Ongoing Monitoring)

Applicable primarily to inorganic pollutants, including radionuclides, solidification involves introduction to the polluted soil of a leach- or weathering-resistant material through either mixing or injection which is then allowed to cure to form a solid product. The material acts to prevent pollutant exposure and migration by physically encapsulating the pollutants within a nontoxic, impermeable coating. A wide range of materials have been used in solidification technologies, including: cement, fly ash, bitumen, emulsified asphalt, and polyethylene. The resulting product is not soil but can often be used as a building or construction material. Ongoing monitoring of the final product is necessary to assess contaminant leakage.

### Incineration (*Ex Situ*)

Applicable primarily to organic pollutants and pathogens, excavated soil is transferred to a contained facility and heated under controlled conditions in an incinerator. Organic pollutants are vaporized and collected, or destroyed through pyrolysis. Some volatile metals (e.g., mercury (Hg), cadmium (Cd), Pb) may also be released and captured from the exhaust gases. For soils or residues generated from separation–fractionation treatments containing very high concentrations of inorganic pollutants, it may be economically viable to roast or smelt the soil and recover the metal or nutrient for reuse.

### Soil Vapor Extraction (*In situ* or *Ex situ*)

Applicable to organic pollutants, in soil vapor extraction, volatile organic contaminants (VOCs) are mobilized and removed from soils through application of a pressure–concentration gradient using a vacuum. Extraction wells are drilled into the soil (*in situ*) or the soil is excavated and placed over a network of perforated pipes (*ex situ*) and the organic pollutants contained within gases extracted following application of the vacuum are collected and destroyed by processes such as pyrolysis or photo-oxidation. Volatilization of the organic contaminants from the soil can be thermally enhanced by steam or hot-air injection, or electrical or electromagnetic heating techniques.

### Photo-oxidation (*Ex Situ*)

Applicable to organic pollutants, vacuum-extracted organic pollutants (see Soil vapor extraction, above) are condensed, mixed with catalyst, and fed into a reactor in photo-oxidation. Ultraviolet light (sunlight or artificial) activates the catalyst, causing oxidation of the organic compounds into simple inorganic products, mainly carbon dioxide and water.

## Chemical Treatment Technologies

### Monitored Natural Attenuation (*In Situ*, Requires Ongoing Monitoring)

Both inorganic and organic pollutants can be managed by a monitoring natural attenuation strategy. The mobility and biological availability of pollutants can decrease with time of contact of pollutants with soils due to chemical reactions and/or sorption which occur over the long term. The rates of these reactions depend upon the chemical characteristics of both the soil and the pollutant. The rates are difficult to predict, but are possible to measure. For low-risk polluted soils in which there is evidence that natural attenuation is occurring at a satisfactory rate, it may be appropriate to rank these soils as lower priority in terms of requirement for other remedial treatments. Even if contaminant solubility is not of immediate concern, action to reduce movement of polluted soil through erosion needs to be considered to prevent physical transport of contaminants off-site. Factors affecting the potential reversibility of the natural attenuation process are also unclear, so there is an ongoing requirement for monitoring of any site for which natural attenuation is enlisted as the primary 'treatment' strategy.

### Fixation–Stabilization (*In Situ* or *Ex Situ*, Requires Ongoing Monitoring)

Applicable primarily to inorganic pollutants, including nutrients. The availability or toxicity of a pollutant is not necessarily related to its total concentration in the soil. This is because contaminants can exist in different chemical forms in soils, and it is primarily the aqueous and easily exchangeable forms which are both mobile and biologically available. Fixation–stabilization technologies aim to decrease the concentration of the aqueous and exchangeable forms of pollutants in soils through the application of nontoxic reactive agents. Agents that have been tested for this purpose include: lime, soluble phosphate compounds, fly ash, clays such as zeolite and also those rich in iron and aluminum oxides (e.g., red mud, steel shots, red gypsum, etc.), as well as mixtures of these materials. Many such agents are by-products from processing industries and so are inexpensive. The agents act either by encouraging the formation of insoluble precipitates of inorganic pollutants or by increasing the amount of pollutant sorbed to the solid phase, thereby decreasing the aqueous concentration of the pollutant. An advantage of this method is that, providing pollutant mobility and availability can be decreased to a satisfactory level, the soil can retain many components of its original functionality. Limitations to use of this technology include soils contaminated with mixed pollutants for which a chemical strategy to manage one of the pollutants is incompatible with the chemistry required to manage another. For example, increasing soil pH can significantly reduce the mobility of metals such as Pb, zinc (Zn), and nickel (Ni) but could result in an increase in the mobility of other co-contaminants such as arsenic (As) and cyanide. Moreover, as with the natural attenuation strategy, care is needed to prevent physical transport of polluted soil offsite, and an ongoing monitoring program may be necessary to assess the continued stability of the fixation treatment.

### Chemical Oxidation–Reduction (*Ex Situ*, may Require Monitoring (Inorganics))

Applicable to some inorganic and organic pollutants, chemical oxidation–reduction uses the addition of oxidizing or reducing agents to convert pollutants to less toxic or inert compounds. For example, both Cr and As exist in more than one valence state. Cr(VI) is more toxic and mobile in soils than Cr(III), and As(III) is more toxic and mobile than As(V). Chemical reduction of Cr and oxidation of As can therefore decrease the pollutant toxicity. However, depending on conditions, there is always a risk of reconversion back to the more toxic valence state. Hence it is necessary to institute an ongoing monitoring program to detect this. Biological degradation processes are usually more cost-effective for organic pollutants than chemical oxidation methods and hence tend to be preferred.

### Aqueous and Solvent Extraction (*Ex Situ*)

Applicable to inorganic and organic pollutants, in the process of aqueous and solvent extraction, excavated soil is placed in a pile and leached with an extraction fluid, or soil and extraction fluid are mixed in a reactor vessel. The extraction fluid dissolves the pollutants and the fluid is collected and then treated to remove the pollutants. The excavated soil is often pretreated using physical separation techniques to minimize the volume of material required to be extracted. For inorganic pollutants, an acidic aqueous solution which may contain surfactants is typically used as the extraction fluid. Soil pH is neutralized with base following removal of the pollutant. For organic pollutants, the extraction fluid is likely to be an organic solvent. Traces of the solvent may remain in the soil following extraction, so toxicity of the solvent is an important consideration. Low extraction efficiencies for pollutants that are highly attenuated (see Monitored Natural Attenuation, above) can severely limit the performance of this technology.

### Dehalogenation (*Ex Situ*)

Applicable to halogenated organic pollutants, dehalogenation, the removal of halogen groups from organic pollutants, generally renders such compounds less toxic or nontoxic. Reagents that catalyze the removal of halogen groups are mixed with polluted soils, and the mixture is heat treated in a reactor vessel to facilitate the reaction. Due to the cost of reagents, this technology is generally uneconomical for highly polluted soils or large volumes of polluted soil.

## Biological Treatment Technologies

### Biodegradation Techniques (*In Situ*)

Applicable to organic pollutants, biodegradation techniques are bioremediation clean-up techniques which aim to decrease the total concentration of pollutant in soil through the degradation of organic compounds by microorganisms. The microorganisms involved, also called degraders, are stimulated to use the pollutants as a source of nutrients and energy (enhanced biodegradation). Generally, this is achieved by optimizing the environmental conditions in terms of oxygen, nutrients, and moisture, and controlling the temperature and pH. For instance, the technology known as bioventing aims to stimulate pollutant biodegradation in unsaturated soils by increasing the soil oxygen content. This is achieved by forcing air through the soil by either extraction (vacuum) or injection methods. Specific strains of degraders can also be introduced to the soil to enhance the degradation process in a technology known as bioaugmentation. Typically for this application, microorganisms from the site under remediation are collected and those that preferentially feed on the contaminants at the site are isolated, cultured, and reintroduced to the site as a means of rapidly increasing the population of degraders. Remediation trials have also been conducted using nonindigenous strains of degraders introduced to a polluted site. However, so far there is no conclusive evidence to indicate that the use of non-native microorganisms has been beneficial in the situations tested.

One of the advantages of biodegradation technologies is that they can typically be implemented at a low cost. However the process may require more time than other available treatments. The degradation rate is controlled by both environmental conditions and by the nature of the pollutant. For example, biological treatment of polyaromatic hydrocarbons (PAHs) is selective, and the less-degradable PAHs (of high molecular weight) may remain unchanged. Furthermore, some compounds may be broken down into more toxic by-products during the bioremediation process (e.g., trichloroethylene (TCE) to vinyl chloride). Therefore, continuous monitoring and additional precautions have to be implemented in these cases. Although not all organic compounds are amenable to biodegradation, bioremediation techniques have been used successfully to treat soils polluted by petroleum hydrocarbons, solvents, pesticides, wood preservatives, and other organic chemicals.

### Composting and Biopiles, and Slurry-Phase Biological Treatment (*Ex Situ*)

Applicable to organic pollutants, in this suite of technologies, the soil is excavated and mixed with appropriate amendments in order to stimulate microbial degradation of the pollutants. During composting, the microbial activity induces an increase in temperature that allows a faster degradation rate and can also be effective in reducing pathogen numbers. Piles of the material to be remediated can be aerated with blowers or vacuum pumps (aerated static-pile composting), or through periodic mixing with mobile equipment (windrow composting). Alternatively, the compost is placed in a reactor vessel where it is mixed and aerated (mechanically agitated in-vessel composting).

In the slurry-phase biological treatment, the excavated soil is mixed with water and various additives. The slurry is stirred to keep solids suspended and microorganisms in contact with the soil contaminants.

Limitation to this technology can be due to the large area needed for composting, the necessity to control release of VOCs, and the presence of toxic metals, which may impair microbial activity. However, this type of biological treatment has been used successfully to remediate soil contaminated by explosives and PAHs.

### Phytoremediation (*In Situ*)

Applicable to both inorganic pollutants, including radionuclides, and organic pollutants, phytoremediation refers to a group of technologies that use green or higher terrestrial plants for treating chemically or radioactively polluted soils. There are five fundamental processes by which plants can be used to remediate soil contaminated with organic or inorganic pollutants. These processes result in the plant-based containment or removal of the soil pollutant. Phytoimmobilization processes prevent the movement and transport of dissolved contaminants by using plants to decrease the mobility of pollutants in soils. Phytostabilization as defined here is the use of pollutant-tolerant plants to mechanically stabilize polluted soils to prevent bulk erosion and airborne transport of pollutants to other environments. In addition, leaching of pollutants may be reduced due to higher evapotranspiration rates relative to bare soils. Phytovolatilization relies on the joint effect of plants and microbes which, acting in concert, are able to convert some metals and organic pollutants to volatile or more available forms. Examples include removal of Se or volatile organic pollutants from soils. In a different approach, plant biochemistry has been modified by the insertion of bacterial genes that make the transformed plants able to volatilize Hg. Phytoextraction can be used to remediate both metallic and organic constituents from soil by direct uptake into plants and translocation to aboveground biomass. Two phytoextraction approaches are being developed, both aiming to extract toxic metals from soil. The first makes use of hyperaccumulator plants with exceptional metal-accumulating capacity (natural phytoextraction). The second uses high-biomass crops that are induced to take up large amounts of metals by enhancing the mobility of metals in soil with chemical treatments (chemically assisted phytoextraction). This latter method has recently been criticized because enhancing mobilization of toxic metals, to increase plant uptake, is also likely to increase the risk of leaching to groundwater. Phytodegradation refers to enhanced degradation of organic pollutants in soil induced by the action of rhizosphere microorganisms and by plant-derived enzymes. Furthermore, some organic compounds, depending on their chemical and physical characteristics, can be taken up by plants and partially or totally degraded by plant enzymes or endophytic microorganisms. Phytodegradation is also defined as plant-assisted biodegradation, because plant roots supply nutrients to microorganisms, which enhances their biological activities.

This group of technologies have a series of limitations as follows: the depth of the treatment zone is limited by the depth of the root system, high concentrations of the pollutants may be toxic to plants, the process may be seasonal, it is not effective for strongly sorbed contaminants, and may require long times to complete. On the other hand, these technologies are low-cost, environmentally friendly, and thought to be well accepted by the public.

## Selection of Remediation Technologies and Future Directions

Once a need for soil remediation is recognized, the best available technology is selected according to the nature, toxicity, and origin of the contaminant, the present and potential hazard related to the degree of contamination, the chemical and physical characteristics of the soil, the land use, the time available for remediation, the community acceptance, and a cost-benefit analysis. These different criteria vary widely in the technologies described above. For instance, the average cost of remediation with different technologies can differ by a factor of greater than 25 (**Figure 2**). The development of low-cost, low-input technologies, driven in industrialized countries by economic benefits, is also essential to tackle remediation of polluted soils in developing countries.

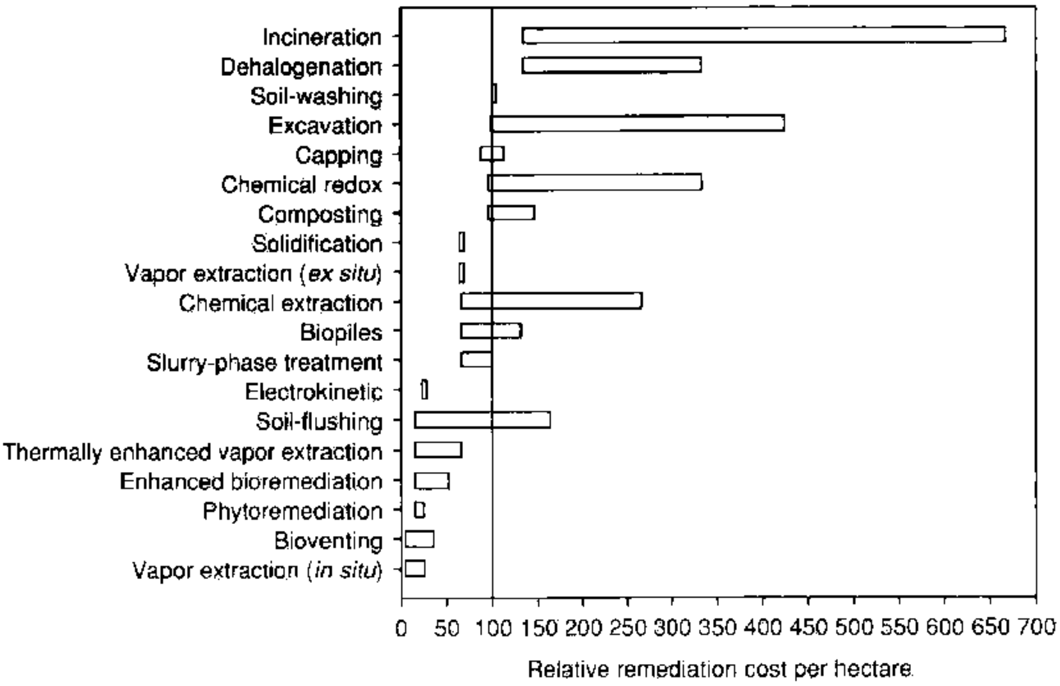


**Figure 2** Range of relative remediation costs for different technologies, expressed as a percentage of the cost for capping. Costs are calculated on the basis of the 1 ha area, 0.25 m soil depth, and soil density of 1.2 g cm$^{-3}$.

Some of the technologies described in this article are well established, whereas others are still being developed. Due to ever-increasing public concern for the risks posed by soil pollution, it is expected that a larger number of technologies will be available in the future. For example, rapid progress in the field of molecular biology will allow a better understanding of the mechanisms involved in bio- and phytoremediation. This knowledge is of fundamental importance in order to optimize these processes. A new field of investigation may also be opened by recent concerns related to the possibility of pathogen-contaminated soils.

*See also:* **Hydrocarbons**; **Pollutants:** Biodegradation; Effects on Microorganisms; **Salt-Affected Soils, Reclamation**

## Further Reading

Adriano DC, Bollag JM, Frankenberger WT, and Sims RC (eds) (1999) *Bioremediation of Contaminated Soils.* SSSA Monographs. Madison, WI: Soil Science Society of America.

Smith LA, Means JL, Means A *et al.* (eds) (1995) *Remedial Options for Metals-Contaminated Sites.* Battelle Memorial Institute. Boca Raton, FL: Lewis Publishers.

Huang PM and Iskandar IK (eds) (1999) *Soils and Groundwater Pollution and Remediation: Asia, Africa and Oceania.* Boca Raton, FL: Lewis Publishers.

Riser-Roberts E (1998) *Remediation of Petroleum Contaminated Soils: Biological, Physical, and Chemical Processes.* Boca Raton, FL: Lewis Publishers.

Wise DL, Tarantolo DJ, Inyang HI, and Cichon EJ (eds) (2000) *Remedial of Hazardous Waste Contaminated Soils.* New York: Marcel Dekker, Inc.

# REMOTE SENSING

Contents

## Organic Matter

**D K Morris**, Stennis Space Center, MS, USA
**C J Johannsen, S M Brouder, and G C Steinhardt**, Purdue University, West Lafayette, IN, USA

### Introduction

Remote sensing is the collection and interpretation of data obtained from an object at a remote vantage point without making physical contact with the object or subject. Electromagnetic energy, as shown in Figure 1, may be recorded by a variety of remote-sensing methods. The most common include color and color infrared photography, aerial and satellite imaging, and radar sensing.

Aerial photography records energy in the ultraviolet part of the spectrum (which is not used for agricultural analysis because it is absorbed by the atmosphere), the part visible to humans, and a portion of the near-infrared (NIR) spectrum. Optical sensors record energy in a digital format from the visible, reflective infrared, and thermal infrared spectra from both aerial and satellite platforms within the Earth's atmosphere. Microwave wavelengths are another form of energy important to remote sensing. Microwave energy is received on Earth and the responses recorded by a digital instrument. This is known as 'passive energy,' compared with energy from portions of the spectrum that are measured directly and thus called 'active energy.'

For more than three decades, remote sensing has been envisioned as a useful source of data for inventorying and monitoring the Earth's surface for many purposes. One can apply the use of remote sensing to disciplines in agriculture, hydrology, forestry, oceanography, weather, land use, environmental concerns, security purposes, and military operations.

In the 1950s it was demonstrated that infrared aerial photography could be used to detect vegetation stress. This led to a host of applications and, by the 1960s, the US Weather Bureau was launching satellites to observe cloud patterns around the globe. Weather satellites today are used to monitor not only clouds, but also changes in precipitation patterns by radar, lightning strikes through brightness sensors, and vegetation density changes. They also contribute to predictions of where and when hurricanes are expected to strike land areas.

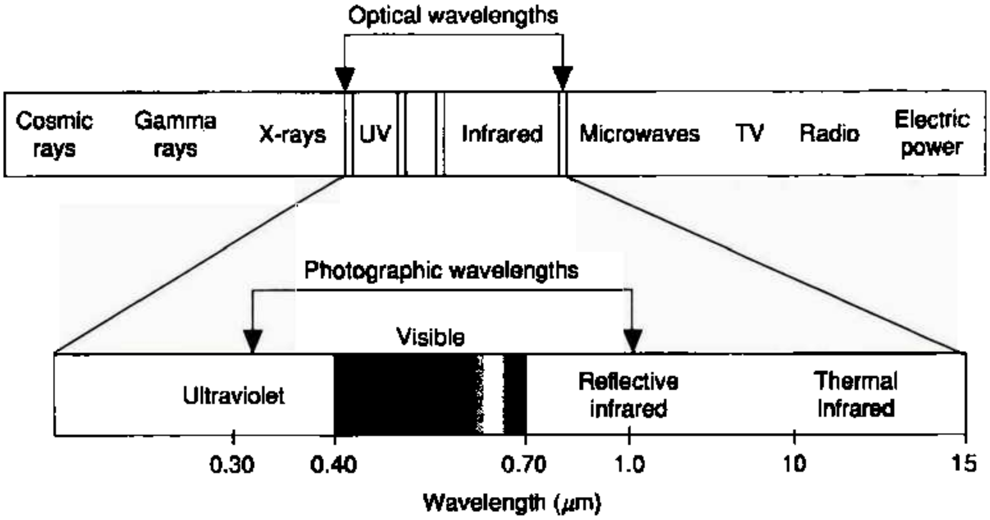


**Figure 1** The electromagnetic spectrum showing the wavelengths of energy that are important to remote sensing applications. These include energy measurements from the visible, reflective infrared, thermal infrared, and microwave regions. (Reproduced with permission from Landgrebe DA (2003) *Signal Theory Methods in Multispectral Remote Sensing*. New Jersey: John Wiley.)

However, until recently, satellite data have not been available at a satisfactory temporal frequency for agricultural purposes; that is, they have not had the ability to provide repetitive coverage on a daily basis or 5- to 10-day intervals. The necessary spatial resolution, which is the ability to identify the detail of the measurement at a meter or less, has been provided by commercial satellites only since 1999. Critical needs for remote sensing in agriculture are to provide spatial resolutions of 5–20 m, frequent coverage of 2–5 days, and rapid data delivery to the user.

Aircraft platforms help in rectifying some of the problems associated with satellite acquisition in that spatial resolution can be acquired in the 1- to 5-m range because of smaller sensor viewing areas. Further, aircraft or aerial data collection is not limited to a fixed schedule and/or orbit as is satellite data acquisition. Satellite data are obtained from fixed orbits, provide no data during cloud cover, and have higher cost considerations, all of which users need to consider when acquiring data within a narrow, specific timeframe.

Current trends in the use of remote sensing data in agriculture are driven by goals similar to those of site-specific crop management. The delineation of suitable management units is an important aspect of remote sensing relating to site-specific crop management practices, and remote sensing has great potential to provide soil and crop information for this purpose. One of the most useful aspects of remote sensing is its ability to generate images that show the spatial variation in fields caused by natural and cultural factors without having to depend on costly sampling intervals or geostatistical interpolations. A major challenge is the development of robust image-analysis techniques for delineating soil conditions; another is to provide the proper training to data providers and farmers/producers in the use of this new information.

Early remote sensing research used primarily two approaches: the first was the traditional photointerpretation or visual approach, in which the actual image or photograph was analyzed subjectively by human interpretation. Although this method gave users an overall impression of within-field variation, it also had two crucial shortcomings: first, visual interpretation of black-white or color paper prints depends largely on the interpreter's personal experience and his or her knowledge of field conditions. This leads to the possibility that the results will vary among analysts and over time. Second, the speed of visual interpretation may pose a problem because of the time required to do an exhaustive analysis of the image.

The second approach in the early days of remote sensing, which is more standardized and less labor-intensive than the visual approach, used computer-implemented pattern recognition techniques applied to an image to classify surface soil patterns into different groups or classes. Advances in computer and information technology have overcome concerns related to the large data volume that needs to be analyzed for a successful application.

## Mapping of Soils

There are many different soil classification systems used around the world and each is a product of the needs of people, the amount of data available, training, and the technology that was used. Remote sensing techniques are only beginning to show a positive influence on the way soils are mapped on the Earth's surface. A unified global effort in soil survey and mapping was the FAO/UNESCO Soil Map of the World produced in the late 1960s, with a revised legend in 1988. This effort produced a map at a scale of 1:5 000 000, but did not make use of spatial technologies.

The US Department of Agriculture (USDA) National Resources Conservation Service (NRCS), a partnership of the National Cooperative Soil Survey (NCSS) with local and state agencies and land grant universities within the USA, are generating soil maps and interpretations using Soil Taxonomy. For modern farming practices, these soil surveys typically do not have sufficient resolution to capture natural soil variability in enough detail to support 'precision farming' recommendations, which use technologies such as remote sensing, geographic information systems, and global positioning systems. With the increasing availability of both very high spatial and spectral resolution remote sensing data, the opportunity to redefine soil survey maps is an important benefit.

Another pressing problem related to the use of soil maps, yield-monitoring, or soil samples is that the data are collected at different spatial scales and therefore need to be adjusted in order to make relevant management decisions. Producers have long known that soil variability influences the productive potential of their fields. Because of labor and analysis costs, some highly commercialized producers use a system of soil sampling in small blocks (i.e., grids). These blocks are commonly 0.4 ha (1 acre) to 1 ha (2.5 acres) in size and can be either square or rectangular in shape. This type of sampling scheme requires that soil samples acquired at coarse and/or variant resolutions be interpolated to obtain average soil characterization values at a given finer resolution.

Since natural soil boundaries do not follow straight lines found within the grid system, a producer makes decisions as to which soil is most represented within the grid (**Figure 2**). Generally, geostatistical analysis is used for this interpolation, based on kriging or inverse-distance weighting. The drawbacks of geostatistical analysis include the need for a large number of samples at close sampling intervals and the assumption of stationary data variation, which is often not the case for soil properties. Remote sensing offers the potential to overcome these drawbacks.

## Remote Sensing for Soil Mapping

Soil scientists have used photography in producing soil maps since the early 1930s. During the first 50 years, much of the interest in photography was related to having a reliable local-to-regional base map. During the 1960s, multispectral remote sensing data were considered of limited use because the evaluated results were compared with existing soil maps generated by conventional methods. The low accuracy attributed to remotely sensed data when compared with existing soil maps was the result of the assumptions that (1) the conventionally prepared soil map is 100% correct, and (2) that the spectral response of the soil surface is determined by the same morphological features used to derive the soil survey maps.

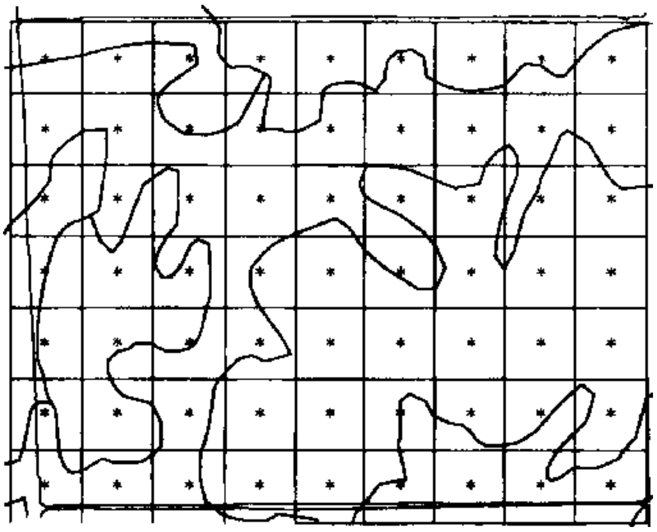


**Figure 2** A gridded soil-sampling pattern overlaid on a soil map to illustrate the difficulty in representing the natural soil patterns with this approach.

Soil survey mapping units are defined by the soil properties throughout the soil profile that affect management practices, such as drainage, and erosion control; whereas maps of spectral responses are based directly on the spectral properties of only the portion of the soil profile that is exposed at the surface. In laboratory studies, soil reflectance has been successfully correlated with soil properties such as organic carbon, moisture, and total nitrogen under controlled conditions.

With the more recent emphasis on intensive, precision farming, producers have been requesting more detailed field-level soil maps. Current attention is focused on the resolution to be mapped and the soil categories or soil management zones that a producer can use effectively in making management decisions.

We know that (1) soil properties such as organic matter and iron oxides are measured by reflectance or image response, (2) soil surface conditions such as surface roughness influence the response, and (3) remotely sensed imagery of different wavelengths enhances soil patterns so they can be more easily mapped. In summary, the responses of images can be related to soil properties such as organic matter content, since dark soils (higher organic matter) are less reflective in contrast with light-colored soils (lower organic matter). The vegetation spectral response can also be used to infer various soil conditions such as water-holding capacity, eroded locations, and nutrient differences.

With advances in remote sensing technologies (e.g., digital imaging), such information can now be directly integrated into a geographic information system. The approach commonly taken in studying the feasibility of using imaging devices for soil mapping has been the acquisition and analysis of imagery or other remotely sensed data over a specific area that is then supplemented by limited ground reference data.

There must be an awareness of changes in remote sensing imagery due to surface reflectance, sensor sensitivity, and spectral composition of the sunlight on the area being sensed. For example, if images are obtained over an area every hour throughout the day, the soil surface will have a different soil reflectance during each acquisition caused by the sun's angle. If the surface soil is moist and it is a windy day, the surface soil reflectance will change owing to the drying action (**Figure** 3).

Automatic pattern-recognition techniques used on aerial sensing data show that different classes of soil surface conditions can be mapped with reasonable accuracy. Since the launch of the Landsat 1 satellite in 1972, much attention has been directed to the relationship of soils and land use. 'Agrophysical units' are defined as geographic areas having definable or comparable agronomic and physical parameters such as soils, which reflect a range in agricultural use and management.

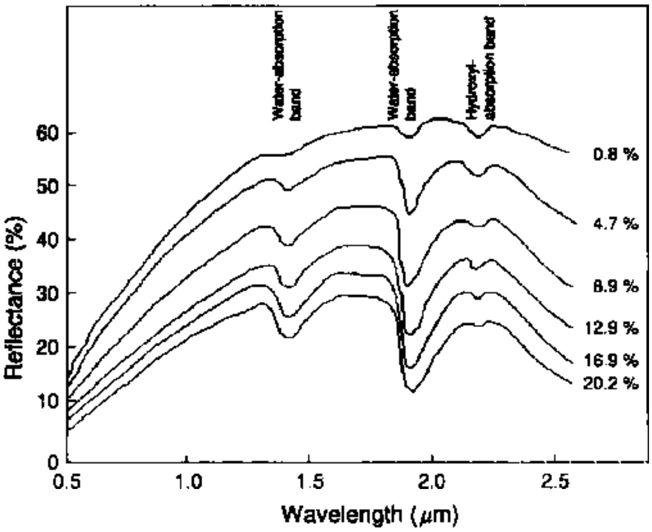


**Figure 3** Laboratory spectra of a silt loam soil under different moisture conditions. Soil reflectance decreases in the visible and near infrared region with an increase in soil moisture. This silt loam soil was dried in the laboratory to demonstrate the reflectance change from 0.5 to 2.5 μm. (Adapted from Bowers SA and Hanks RJ (1965) Reflection of radiant energy from soil. *Soil Science* 100: 130–138.)

Early work using visual interpretation of Landsat images provided overlays with acetate transparencies of soil map sheets to investigate relationships of Landsat multispectral data to soil patterns. Although the results vary by geographic regions, there is agreement that the use of satellite or airborne imagery decreases the time required in the soil survey and mapping period and therefore reduces costs. Spectral classification alone cannot distinguish between widely differing soils exhibiting similar spectral responses. By adding ancillary data, a more detailed delineation of soils can be provided as compared to information derived solely from Landsat data. However, these comparisons also reveal some discrepancies between spectral soil maps produced from Landsat data and conventional soil survey maps. One important soil property that has received much focus is soil organic matter.

## Organic Matter Detection

Quantitative relationships between soil color and organic matter content are not well quantified because of differences in soil moisture, iron oxides, and other soil properties. However, these relationships are of considerable practical importance in mapping and classifying soils, interpreting soil properties, and designing sensors that can detect soil properties.

Soils of higher organic matter content are generally less reflective or darker in color than those soils with lower organic matter content; therefore, soil color has long been associated with native soil fertility. Soils with thick, dark surface horizons are often separated from other soils in many soil classification systems, which emphasize the importance of these soils both as a medium for plant growth and also as an indicator of land value for agricultural and urban uses.

Accurate and timely information about soil organic matter is essential for agricultural production and environmental research. During the past three decades, high-altitude remote sensing coupled with laboratory-based reflectance spectroscopy has emerged as an important technology for monitoring the Earth's agricultural resources.

Although data from currently operational reflectance sensors such as Landsat, SPOT, and a host of other public and private sensors (**Table** 1) have proven useful for crop identification and wetland inventory, attempts to characterize soils have attained only limited success. The major difficulty in deriving soil information from remotely sensed data is the

**Table 1** Typical spatial resolution values of selected satellite sensors

| Satellite & sensor | Spatial resolution (m) |
|---|---|
| DigitalGlobe Quick Bird panchromatic | 0.6 |
| Space imaging IKONOS panchromatic | 1 |
| DigitalGlobe Quick Bird multispectral | 2.4 |
| Space imaging IKONOS multispectral | 4 |
| IRS-1C panchromatic | 5 |
| RADARSAT | 8–100 |
| SPOT panchromatic | 10 |
| Landsat 7 panchromatic | 15 |
| JERS-1 radar | 18 |
| SPOT multispectral | 20 |
| Seasat radar | 25 |
| Landsat thematic mapper | 30, 60, 120 |
| IRS-1B LISS-II | 36 |
| IRS-1A LISS-I | 72.5 |
| Landsat multispectral scanner | 80 |
| Advanced very high resolution radiometer (AVHRR) | 1100 |

complex nature of relationships between soil parameters and reflectance. Organic matter, moisture, texture, cation exchange capacity (CEC), mineral oxides, and surface conditions such as residue cover and soil crusting have demonstrated unique influences on soil reflectance. These relationships are wavelength-dependent; therefore, the soil parameters may affect reflectance differently in different regions of the spectrum.

In glaciated soil regions, high organic soils can be readily differentiated in high-resolution satellite images from soils formed under prairie grass vegetation and soils formed under forest vegetation (**Figure** 4).

Several problems have been encountered when viewing the soil with an electro-optic sensor. These problems can be reduced if one has collected reference information on ground conditions. Bare soil areas with high-residue content from the previous crop may mask soil spectral responses (**Figure** 5). The use of high spatial digital terrain models

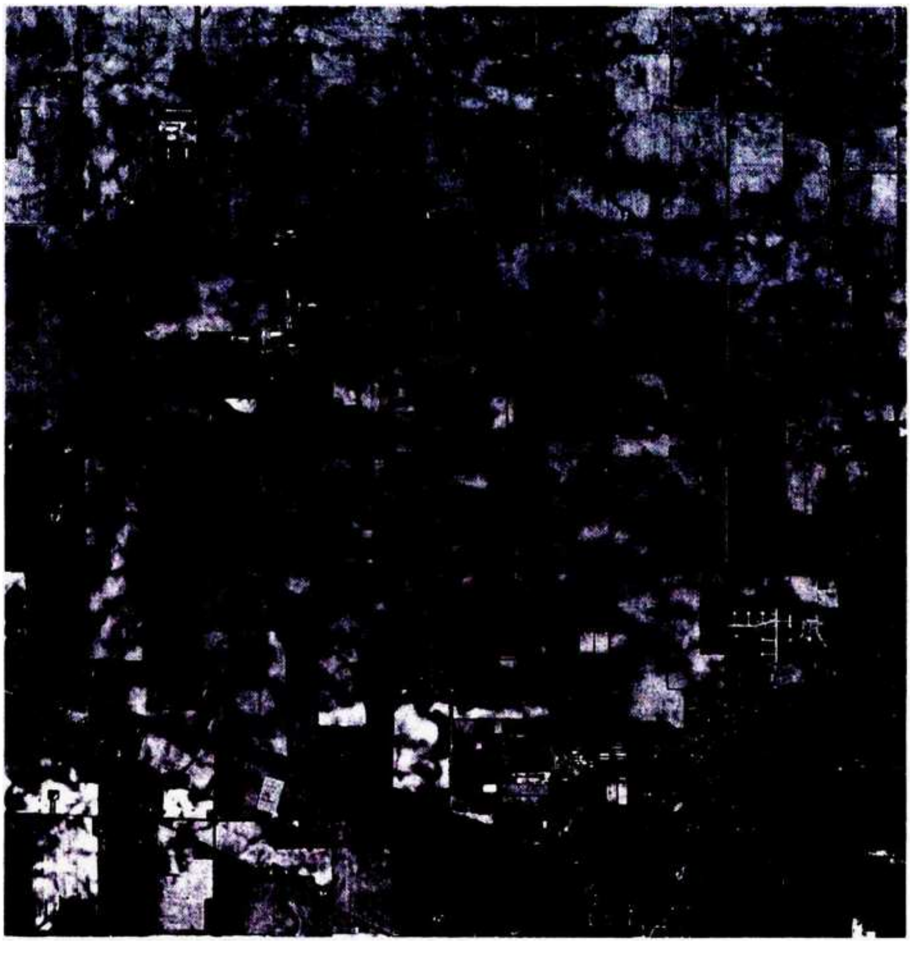


**Figure 4** (see color plate 54) Soil patterns as seen on an Ikonos color infrared image (24 May 2000) of the Purdue University Agronomy Center for Research and Education (ACRE, W. Lafayette, Indiana, USA) and surrounding area. The dark soils (Mollisols) are higher in organic matter and derived under grassland vegetation, while the lighter soils (Alfisols) have low surface soil organic matter and are derived under forest vegetation. Red shows healthy vegetation; white, streets, roadways, and buildings; dark green to black, areas of high organic matter and high surface moisture of soils; light-green areas, low surface soil organic matter, some of which are muted by high surface residues from the previous crop. © Space Imaging 2000

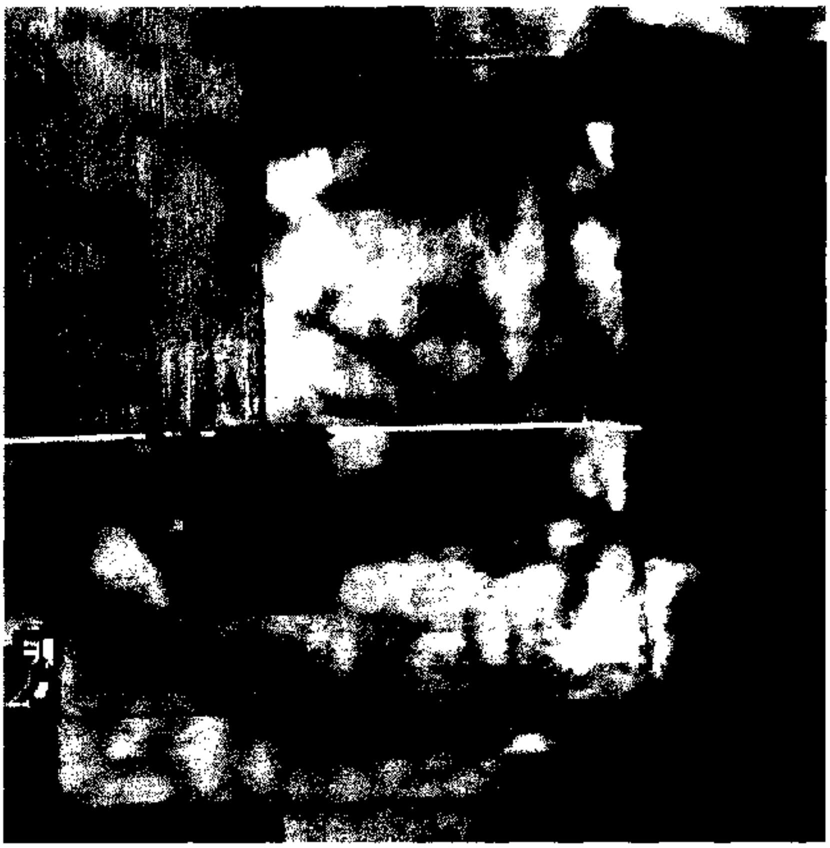


**Figure 5** (see color plate 55) Aerial color infrared image showing soil patterns over an eastern Indiana farming area. The darker areas are higher in organic matter content and were previously wetland areas. The field in the upper left does not have the same soil patterns, since residues from the previous year cover it. Red shows healthy vegetation such as grasses and bushes; maroon, trees (the color resulting from the leaves, branch spacing, and shadows); white, roads and buildings; dark green to black, areas of high organic matter; light green, low surface soil organic matter content. Muted lined patterns masking the normal colors are corn residues from the previous crop. Courtesy of Emerge.

(DEM) produced by remote microwave or laser techniques can provide a better understanding of soil formation and surface moisture movement and aid in interpreting surface soil organic matter content.

Sensing soil organic matter from airborne or satellite platforms requires reference data. Soil types whose classifications are driven by soil color and all of the previously mentioned parameters may have as much variation in soil properties from within a given soil type as they do between different soil types. Many different methods have been developed for the calibration of soil organic matter sensors. Most involve a given sequence of soils of about the same age, related parent material, and similar climatic conditions, but characterized by variations in topography and drainage. Laboratory studies have measured soil reflectance, attempting to quantify the relationship between soil color and organic matter content by comparing the soil spectra with soil Munsell values.

Most studies have concluded that NIR predictions of organic matter and soil moisture become less accurate as one considers soil samples from a wider geographic range. Low correlations between reflectance and organic carbon content have been attributed to the confounding effects of texture, mineralogy, and soil moisture. It is expected that high-resolution (temporal, spatial, and spectral) reflectance data will lead to a better understanding of the relationships between these and other soil properties and reflectance.

Investigations have attempted to link reflectance to specific organic matter components. Typically these studies fractionate organic matter into pools such as humic acid, fulvic acid, and the remaining mineral components; then each fraction is analyzed using a spectrophotometer. It has been shown that even small amounts of organic material may be particularly effective in masking mineral features in the shorter wavelengths. It appears that unfractionated soils and their mineral components are important influences on reflectance with regard to examining organic carbon content in relation to other soil properties.

**Table 2** Importance of spectral bands based on response of soil reflectance to organic carbon (OC) content, parent material, and other soil properties

| *Soils within similar parent materials (northern soils)* | | | *Soils across different parent materials (all soils)* |
|---|---|---|---|
| *Responds to OC content differences only (nm)* | *Responds to other properties. OC content effect is dominant (nm)* | *Responds to other properties that mask OC content effect (nm)* | *Responds to OC content and minimizes parent material effect (nm)* |
| 1065 | 405–1055 | 1775–1885 | 1955–1965 |
| 1085–1105 | 1075 | 1975–2205 | 2215 |
| 1125–1165 | 1115 | 2225–2255 | 2265 |
| 1955–1965 | 2325 | 2275 | 2285–2295 |
| 2215 | 2375 | | 2315–2495 |
| 2265 | 2425 | | |
| 2285–2295 | 2445–2485 | | |
| 2315 | | | |
| 2335–2365 | | | |
| 2385–2415 | | | |
| 2435 | | | |
| 2495 | | | |

Source: Henderson TL, Baumgardner MF, Franzmeier DP, Stott DE, and Coster DC (1992) High dimensional reflectance analysis of soil organic matter. *Soil Science Society of America Journal* 56: 865–872, with permission.

To examine specific effects of organic carbon on soil reflectance, researchers have tested differing spectral response in separable bands in terms of observable soil properties such as sand, silt, clay, iron oxides, magnesium oxides, and organic carbon. The separable bands of the inorganic fractions (in which the extractable organic matter and humic acid have been removed) are compared with the separable bands from the soils that are unfractionated. There is a direct relationship between organic carbon and soil reflectance in the set of bands in question: as soil organic carbon increases reflectance decreases. The organic matter contained within the soil is the physical factor responsible for spectral differences in this set of bands (Table 2).

In Table 2, the first column lists the bands that respond to organic carbon content only. In these bands, higher organic carbon content causes a decreased measure of reflectance. The first three of the 12 optimal bands are in the NIR portion of the spectrum (700–1300 nm), with the others in the middle-infrared (MIR) region (1300–2500 nm). The second column lists bands in which soil reflectance responds to both organic carbon and other soil properties. These bands include the entire visible range of the spectrum (400–700 nm) and some in the NIR and MIR region; they are useful because the organic carbon content masks the effect of the mineral soil properties. The wavelengths in column 3 should be avoided when classifying organic carbon content. Although the soil curves are separable in the bands listed, the spectral response in these bands is unrelated to organic carbon content, since no direct relationship between organic carbon content and the magnitude of the spectral response in those wavelengths exists.

## The Future

The hyperspectral format contains up to several hundred wavelength bands and provides the potential for differentiating more specific soil properties. Note that current studies have used only three to four bands of data, which are very broad bands. The hyperspectral bands would be very narrow in terms of the wavelengths covered. The spectral mapping of specific soil properties will provide further assistance to soil pedologists in the development of soil surveys.

Recent advances in precision farming technologies require an accurate representation of yield-limiting variables, which include soil properties. Producers have used soil type maps, as they are related to soil organic matter, to help in management practices such as fertility application, herbicide placement, and seed spacing. When coupled with the adoption of global positioning satellites and geographic information system technologies, accurate soil organic matter maps allow for greater precision in performing management practices.

Advances in remote sensing technology are providing a new way to 'see' soil properties and agricultural fields. The provision of geographically accurate soils data assists producers to adjust their application rates of fertilizers and pesticides according to soil properties. This helps producers to increase profits and to provide environmentally safer approaches to producing food.

*See also:* **Classification of Soils**; **Organic Matter:** Principles and Processes; **Remote Sensing:** Soil Moisture; **Statistics in Soil Science**

## Further Reading

Baumgardner MF, Silva LF, Biehl LL, and Stoner ER (1985) Reflectance properties of soils. *Advances in Agronomy* 38: 1–44.

Ben-Dor EN, Goldlshleger N, Benyamini Y, Agassi M, and Blumberg D (2003) The spectral reflectance properties of soil structural crusts in the 1.2- to 2.5-μm spectral range. *Soil Science Society of America Journal* 67: 289–299.

Buol SW, Hole FD, McCracken RJ, and Southard RJ (1997) *Soil Genesis and Classification.* Ames, IA: Iowa State University Press.

Ding G, Novak JM, Amarasiriwardena D, Hunt P, and Xing B (2002) Soil organic matter characteristics as affected by tillage management. *Soil Science Society of America Journal* 66: 421–429.

Fernandez RN, Schulze DG, Coffin DL, and Van Scoyoc GE (1988) Color, organic matter, and pesticide adsorption relationships in a soil landscape. *Soil Science Society of America Journal* 52: 1023–1026.

Henderson TL, Baumgardner MF, Franzmeier DP, Stott DE, and Coster DC (1992) High dimensional reflectance analysis of soil organic matter. *Soil Science Society of America Journal* 56: 865–872.

Hoffer RM and Johannsen CJ (1969) Ecological potentials in spectral signature analysis. In: Johnson P (ed.) *Remote Sensing in Ecology*, pp. 1–16. Athens: University of Georgia Press.

Jensen JR (2000) *Remote Sensing of the Environment: An Earth Resource Perspective*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

Johannsen CJ, Baumgardner MF, Willis PR, and Carter PG (1998) *Advances in Remote Sensing Technologies and Their Potential on Agriculture.* 1st International Conference on Geospatial Information in Agriculture and Forestry, Paper 2-1, Proceedings vol. 1: 6–11.

Kristof SJ (1971) Preliminary multispectral studies of soils. *Journal of Soil and Water Conservation* 26(1): 15–18.

Lillesand TM, Kiefer RW, and Chipman JW (2004) *Remote Sensing and Image Interpretation*, 5th edn. New York: John Wiley.

Moran SM, Inoue Y, and Barnes EM (1997) Opportunities and limitations for image-based remote sensing in precision crop management. *Remote Sensing of Environment* 61(3): 319–346.

National Research Council (1997) *Precision Agriculture in the 21st Century: Geospatial and Information Technologies in Crop Management.* Washington, DC: National Academy of Sciences.

Shonk JL, Gaultney LD, Schultz DG, and VanScoyc GE (1991) Spectroscopic sensing of organic matter content. *Transactions of the American Society of Agricultural Engineers* 34(5): 1978–1984.

Stoner ER and Baumgardner MF (1981) Characteristic variations in reflectance of surface soils. *Soil Science Society of America Journal* 45: 1161–1165.

# Soil Moisture

**T J Jackson**, USDA Agricultural Research Service, Beltsville, MD, USA

Soil moisture mapping using remote sensing is becoming a reality. Soil moisture has been difficult to measure and map using conventional ground-based technologies. Microwave remote sensing-based soil-moisture retrieval has been demonstrated using tower and aircraft instruments; however, the implementation of this approach on satellites has been limited by both technology and availability. Developments in both science and associated technologies now make the exploitation of the microwave region for soil moisture mapping feasible. A review of the basic principles, operational retrieval algorithms, aircraft applications, and current and future satellite missions is presented.

## Microwave Remote Sensing of Surface Soil Moisture

Microwave sensors, especially those operating at lower microwave frequencies (<6 GHz), are of particular value in soil moisture measurement because they respond to the amount of water in the soil. When this technique is used in a remote sensing instrument, the surface layer of soil contributes to the microwave sensor measurement.

Microwave remote sensing can provide a direct measurement of the surface soil moisture for a range of vegetation-cover conditions within reasonable error bounds. Two basic approaches are used, passive and active. In passive methods, the natural thermal emission of the land surface (or brightness temperature) is measured at microwave wavelengths, using very sensitive detectors. In active methods or radar, a microwave pulse is sent and received. The power of the received signal is compared to that which was sent to determine the backscattering coefficient.

The microwave region of the electromagnetic spectrum consists of frequencies of 0.3–30 GHz. This region is subdivided into bands, which are often referred to by a lettering system. Some of the relevant bands that are used in Earth remote sensing are: K (18–27 GHz), X (8–12 GHz), C (4–8 GHz), and L (1–2 GHz). Within these bands, only small ranges are protected for scientific applications, such as radioastronomy and passive sensing of the Earth's surface. Communications and other uses can contaminate the unprotected regions of this electromagnetic region. This contamination is called radiofrequency interference and it is becoming a greater problem each year.

A general advantage of microwave sensors (in contrast to visible and infrared) is that observations can be made through cloud cover, because the atmosphere is nearly transparent, particularly at frequencies <10 GHz. In addition, these measurements are not dependent on solar illumination and can be made at any time of the day or night. Instruments can be mounted on trucks, aircraft, and spacecraft for repetitive large-area observations.

Microwave sensors operating at very low microwave frequencies (<6 GHz) provide the best soil moisture information. At low frequencies, attenuation and scattering problems associated with the atmosphere and vegetation are less significant, the instruments respond to a deeper soil layer, and a higher sensitivity to soil water content is present.

As noted above, there are two microwave remote sensing approaches, passive and active. Each has unique capabilities and limitations that must be matched with specific types of applications. Currently, passive microwave remote sensing offers more robust and well-validated retrieval algorithms. However, it is limited by spatial resolution at satellite altitudes. Very large antenna sizes are required to achieve high spatial resolutions. Getting these into space at a reasonable cost has been a limitation for quite some time. Currently, products are limited to 50 km resolution; however, future systems will provide 10 km products. Passive systems can provide daily global coverage. Active microwave techniques can provide very high spatial resolution data from space (10–20 m) but suffer from weak soil moisture algorithms. The temporal repeat of these systems is also very limited, typically over 30 days.

## Reflectivity and Soil Moisture

Active and passive sensors provide different types of measurements, which are determined by various physical phenomena. However, both types of sensor provide information on the surface reflectivity. The essential relationship needed to utilize microwave remote sensing is a link between reflectivity and soil moisture.

Assuming that the Earth is a plane surface with surface geometric variations and volume discontinuities much less than the wavelength, only refraction and absorption of the media need to be considered. This situation permits the use of the Fresnel reflection equations as a model of the system. These equations predict the surface reflectivity ($r$) as a function of dielectric constant (k) and the viewing angle ($\theta$), based on the polarization of the sensor (horizontal-H or vertical-V).

$$r^{\mathrm{H}} = \left|\frac{\cos\theta - \sqrt{\mathrm{k} - \sin^2\theta}}{\cos\theta + \sqrt{\mathrm{k} - \sin^2\theta}}\right|^2 \qquad [1]$$

$$r^{\mathrm{V}} = \left|\frac{\mathrm{k}\cos\theta - \sqrt{\mathrm{k} - \sin^2\theta}}{\mathrm{k}\cos\theta + \sqrt{\mathrm{k} - \sin^2\theta}}\right|^2 \qquad [2]$$

The Fresnel equations apply when the two media at the interface each have uniform dielectric properties. From the reflectivity, the dielectric constant of the soil can be estimated. The dielectric constant of soil is a composite of the values of its components: air, soil, and water. Although the dielectric constant is a complex number, for soil mixtures the imaginary part is small and can be ignored for computational purposes without introducing significant error. Values of the real part of the dielectric constant for air and dry soil particles are 1 and 5, respectively. Therefore, a very dry soil has a dielectric constant of about 3. Water has a value of about 80 at low frequencies. The basic reason microwave remote sensing is capable of providing soil moisture information is this large difference between the dielectric constants of water and the other components. Because the dielectric constant is a volume property, the volumetric fraction of each component is involved in the computation of the mixture dielectric constant. The computation of the mixture dielectric constant (soil, air, and water) has been the subject of several studies, and different theories exist as to the exact form of the mixing equation.

The dielectric constant of water referred to above is that of free water, in which the molecules are able to rotate and align with an electrical field. However, not all the water in soil satisfies this condition. Some water in the soil has different properties. The initial water added to dry soil below a 'transition' water content is held more tightly by the soil particles and has the dielectric properties of frozen water (~3).

Based upon the discussion above, four components must be considered in computing the dielectric constant of soil: air, soil particles, bound water, and free water. Thus, in order to interpret the data correctly, some knowledge of the soil texture is needed. All dielectric mixing models incorporate this information. Based on an estimate of the mixture dielectric constant derived from the Fresnel equations and information on the soil texture, volumetric soil moisture can be estimated using an inversion of the model.

Various theories describe the reflection resulting from a soil profile with uniform or varying properties. The computations involve a nonlinear weighting that decays with depth. Some modeling studies suggest that this dominant depth is one-tenth the wavelength. Field experiments suggest that the contributing depth is about one-quarter the wavelength. At a frequency of 6 GHz this depth would be about 1.5 cm and at 1.4 GHz about 5 cm.

## Passive Microwave Techniques

Passive microwave remote sensing utilizes radiometers that measure the natural thermal radio emission within a narrow band centered on a particular frequency. The measurement provided is the brightness temperature in degrees Kelvin, $T_B$, which includes contributions from the atmosphere, reflected sky radiation, and the land surface. Atmospheric contributions are negligible at frequencies <10 GHz, and the following section assumes this to be the case. Galactic and cosmic radiation contribute to sky radiation and have a known value that varies very little in the frequency range used for observations of soil water content ($T_{\mathrm{sky}}$ ~4 K). The observed brightness temperature of a surface is equal to its emissivity ($e_{\mathrm{obs}}$) multiplied by its physical temperature ($T$). The observed emissivity is equal to 1 minus the reflectivity, which provides the link to the Fresnel equations and soil moisture for passive microwave remote sensing.

Based upon the above and neglecting atmospheric contributions, for a specific frequency ($f$), polarization ($p$), and viewing angle ($\theta$) the equation for $T_B$ is:

$$T_{B_{fp\theta}} = e_{\mathrm{obs}_{fp\theta}} T + (1 - e_{\mathrm{obs}_{fp\theta}}) T_{\mathrm{sky}} \qquad [3]$$

The second term of eqn [3] is about 2 K. A typical range of response for a soil is 60 K; therefore, the reflected sky contribution can be dropped for computational purposes. Eqn [3] can be rearranged as follows:

$$e_{\mathrm{obs}_{fp\theta}} = T_{B_{fp\theta}} / T \qquad [4]$$

If the physical temperature is estimated independently, the emissivity can be determined from $T_B$. The physical temperature can be estimated using surrogates based on satellite surface temperature, air-temperature observations, or forecast model predictions. As noted above, the derived emissivity is for a specific instrument configuration (frequency, polarization, and viewing angle).

For natural conditions, varying degrees of vegetation type and density are likely to be encountered. The presence of vegetation has a major impact on the microwave measurement. Vegetation reduces the sensitivity of the retrieval algorithm to changes in soil water content by attenuating the soil signal and by adding a microwave emission of its own to the microwave measurement. This attenuation increases with increasing microwave frequency, which is another important reason for using lower frequencies. Attenuation is characterized by the optical depth of the vegetation canopy. A widely used correction for these effects uses information on the vegetation type (parameter $b$) and the vegetation water content ($W_{\mathrm{veg}}$) as follows:

$$e_{\mathrm{surf}_{fp\theta}} = 1 + [e_{\mathrm{obs}_{fp\theta}} - 1]\exp[-2bW_{\mathrm{veg}} \cos\theta] \qquad [5]$$

The corrected emissivity represents that of the soil surface ($e_{\mathrm{surf}}$). Methods have been developed for estimating the vegetation-type parameter (typically derived from land cover) and vegetation water content using visible/near-infrared remote sensing.

After performing the vegetation correction, one can obtain an estimate of the emissivity of the soil surface, which includes the effects of surface roughness. These effects must also be removed in order to determine the equivalent smooth-surface soil emissivity, which is required for the Fresnel equation inversion. Techniques are available for doing this that use statistical characterization of the surface heights and land-use information.

**Figure 1** is a simulated set of curves showing the relationship between soil moisture and emissivity for bare soil and vegetated conditions. Vegetation decreases the sensitivity to soil moisture. Errors in the model, parameters, and the precision of measurements become more important and at some point will overwhelm the information in the observation.

## Active Microwave Methods

Active microwave sensors or radars measure the sent and received power to yield a variable called the backscattering coefficient ($\sigma^0$). The backscattering coefficient is related to the surface reflectivity. As described previously, the reflectivity is then used to determine surface soil moisture. For active techniques, the link between the measurement of the backscattering coefficient and the surface

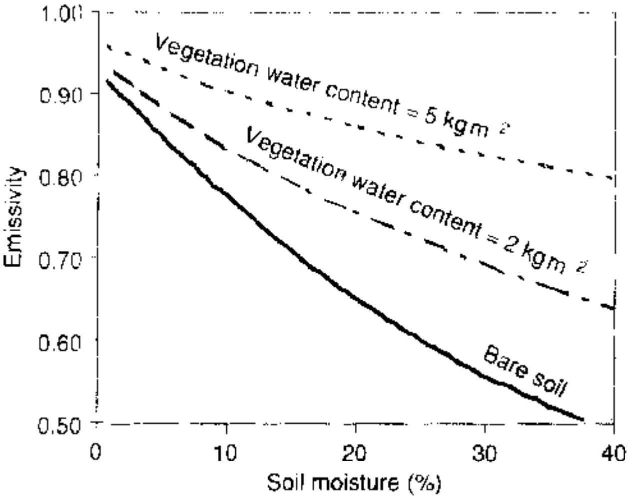


**Figure 1** Relationships between soil moisture and emissivity for bare soil and vegetation. The configuration is a passive microwave instrument operating at 1.4 GHZ, horizontal polarization, and a 40 degree viewing angle.

reflectivity is more complex than for passive methods. The geometric properties of the soil surface and vegetation have a greater effect on these measurements and simple correction procedures are difficult to develop.

The signals sent and received by a radar are usually linearly polarized, either horizontal (H) or vertical (V). Possible combinations are HH, VV, HV, and VH. Advanced multipolarization systems can make all of these measurements simultaneously.

For bare soils, all models that relate the backscattering coefficient to soil moisture require at least two soil parameters, the dielectric constant and the surface-height standard deviation. This means that, in order to invert these models, the surface-height standard deviation must be accurately estimated.

For a given sensor configuration of frequency and viewing angle, different results are obtained at different polarizations. Therefore, most approaches to determining soil moisture with active microwave methods utilize dual polarization measurements. With two independent measurements of two dependent variables, both the dielectric constant and the surface height standard deviation can be solved for. Algorithms incorporating this approach for bare soils are available in the literature. A robust approach to account for vegetation has not yet been developed and validated.

Active microwave sensors on aircraft and spacecraft typically employ synthetic aperture radar (SAR), which utilizes the motion of the platform to synthesize larger antennas. Exceptional spatial resolutions with footprints of about 20 m can be achieved from satellite altitudes.

## Aircraft Platforms and Application Example

Aircraft-based microwave instruments are especially useful in studies requiring the mapping of large areas. In most cases, they offer better spatial resolution (as low as several meters) than satellite systems, as well as more control over the frequency and timing of coverage. Many of the recent soil moisture studies have utilized the electronically scanned thinned-array radiometer (ESTAR). ESTAR is an L-band H-polarized passive microwave instrument that provides image products. It is also a prototype for a new type of sensor system that uses synthetic aperture antenna technology to solve the high-altitude spatial resolution problem described earlier.

Several active-microwave aircraft systems are in operation. One of the most widely used for studies related to soil moisture is the AIRSAR, operated by NASA Jet Propulsion Laboratory. The AIRSAR flies on a DC-8 aircraft and provides multipolarization data at frequencies of 5.2, 1.2, and 0.4 GHz (C, L, and P-band). It is very efficient for mapping large areas and provides exceptional spatial resolution, on the order of several meters.

Perhaps the most informative and widely studied aircraft data set is that obtained from Washita'92. This was a large-scale study of remote sensing and hydrology conducted using ESTAR over the Little Washita watershed in southwestern Oklahoma, USA. Passive and active microwave observations were made over a 9-day period in June 1992. The watershed was saturated with a great deal of standing water at the outset of the study. During the experiment, no rainfall occurred and observations of surface soil-water content exhibited a dry-down pattern over the period. Observations of surface soil-water content were made at sites distributed over the area. Significant variations in the level and rate of change in surface soil-water content were noted over areas dominated by different soil textures.

Passive microwave observations were made on 8 of the 9 days of the study period. The ESTAR data were processed to produce brightness-temperature maps of a 740-km$^2$ area at a 200-m resolution on each of the 8 days. Using the retrieval algorithm of soil water content, these brightness-temperature data were converted to images of soil-water content. Gray-scale images for each day are shown in Figure 2. These data exhibit significant spatial and temporal patterns. Spatial patterns are associated with soil textures, and temporal patterns are associated with drainage and evaporative processes.

Follow-on analyses revealed that these types of observations had potential in determining the

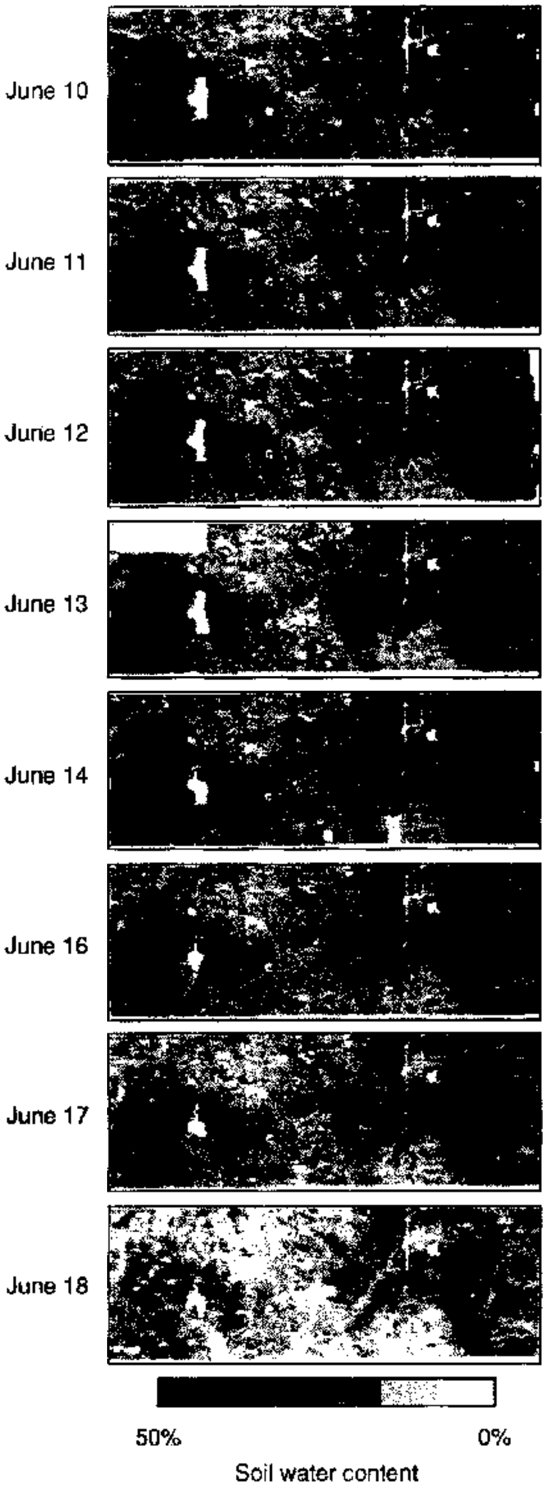


**Figure 2** Maps of soil water content for the Little Washita watershed (Oklahoma, USA) derived from passive microwave data for June 1992. Spatial resolution is 200 m.

hydraulic properties of the soils. Like soil moisture, these data can be obtained at the field point scale; the data are not available over large areas, do not take into account spatial variability, and are costly and difficult to obtain. Temporal observations of soil moisture have been used in several investigations to estimate hydraulic properties. These studies constrained the problem by minimizing the role of the surface flux, or by specifying it, in order to estimate the hydraulic properties of the integrated soil profile.

It has been hypothesized that the change in soil-surface moisture following a wetting event and 2 days later could be used to determine the average hydraulic conductivity of the soil profile. Using simulations for soils of different textures, it has been observed that a relationship exists that is strongest for sandy soils and weakens as clay content increases. In addition, it has been observed that a measurement of the 0–5 cm layer provides as much information as a 0–30 cm observation on the deep soil moisture profile. These relationships were affected by evaporation; however, the effect was more significant for soils with lower hydraulic conductivity and with a shallow depth of observation.

This concept has been applied utilizing remotely sensed data from Washita'92 to derive spatially distributed information. As noted above, investigators hypothesized that surface soil-moisture changes could be used to identify soil-texture classes. A conclusion was that even if this method were not highly accurate for each and every sensor resolution element (pixel), the results would significantly improve the definition of the spatial domain of the soil hydraulic properties. In their opinion, this result was more valuable than the accuracy of individual point estimates.

## Passive Microwave Satellite Observing Systems and Products

### Special Satellite Microwave/Imager

Currently, all passive microwave sensors on satellite platforms operate at high frequencies (>6 GHz) (Table 1). Of particular note, due to the longevity of its data record, is the Special Satellite Microwave/Imager (SSM/I) package on the Defense Meteorological Satellite Platforms. These satellites have been in operation since 1987 and provide high frequencies and two polarizations. Interpreting data from the SSM/I to extract surface information requires accounting for atmospheric effects on the measurement. When one considers the atmospheric correction, the significance of vegetation attenuation, and the shallow contributing depth of soil for these high frequencies, it becomes apparent that the data are of limited value for estimating soil-water content. Spatial resolution of the SSM/I is very coarse (Table 1).

There have been few attempts at generating standard land-surface products using SSM/I data. To date the only standard soil-moisture-related product produced from SSM/I data is the Soil Wetness Index (SWI).

**Table 1** Microwave satellite or sensor systems

| *Satellite* | *Period of coverage* | *Frequency (GHz)* | *Polarization* | *Spatial resolution of lowest frequency* | *Repeat frequency (days)* |
|---|---|---|---|---|---|
| *Passive* | | | | | |
| SSM/I | 1987–present | 19.4, 22.2, 37.0, 85.5 | H and V | 70 km | 1–2 |
| TMI | 1998–present | 10.7, 19.4, 21.3, 37.0, 85.5 | H and V | 60 km | 1 |
| AMSR-E | 2002–present | 6.9, 10.7, 18.7, 23.8, 36.5, 89.0 | H and V | 75 km | 2–3 |
| Windsat | 2003–present | 6.8, 10.7, 18.7, 23.8, 37 | H and V (U in some channels) | 50 km | 2–3 |
| SMOS | (2007) | 1.4 | H and V | 50 km | 2–3 |
| HYDROS | (2010) | 1.4 | H and V (passive) HH, VV (active) | 40 km (passive) 3 km (active) | 2–3 |
| *Active* | | | | | |
| ERS | 1991–present | 5.3 | VV | 30 m | 35 |
| Radarsat-1 | 1995–present | 5.3 | HH | 7 to 100 m | 24 |
| Envisat | 2002–present | 5.3 | HH and VV or HH and HV or VV and VH | 30 m | 35 |
| Radarsat-2 | (2005) | 5.3 | HH, VV, HV and VH | 3 to 100 m | 24 |
| ALOS | (2005) | 1.27 | HH or VV and HV or VH | 10 to 100 m | 46 |

This index is intended to provide information on significantly wet soil conditions (areal extent of flooding), which can be more reliably detected than variations at lower levels of soil moisture.

A few studies have attempted to extract actual soil moisture from SSM/I data, with some degree of success. For the limited validation data set available the approach performed relatively well. However, part of this may be attributable to the limited conditions and light vegetation conditions evaluated.

### Tropical Rainfall Measurement Mission Microwave Imager

Another current option is the Tropical Rainfall Measurement Mission (TRMM) Microwave Imager (TMI). It is a five-frequency, dual-polarized, and passive-microwave radiometer. The lowest TMI frequency is 10 GHz, about half that of the SSM/I. The TMI has a higher spatial resolution (Table 1) as compared to the SSM/I. TRMM only provides coverage of the tropics, which includes latitudes between 38°N and 38°S for the TMI instrument. However, a unique capability of the TMI is its ability to collect data daily, and in many cases more often, within certain latitude ranges. The data from the TMI have been used to generate validated soil-moisture products for 5 years over the southern USA.

### Advanced Microwave Scanning Radiometer

Several new multifrequency passive-microwave satellite systems were launched in 2002 and 2003. As opposed to the previously available systems, these offer a lower-frequency channel operating at C band (~6 GHz), which should provide a more robust soil moisture measurement. These satellites were the NASA Aqua, Japanese ADEOS-II, and the US interagency Coriolis satellite. The ADEOS-II satellite stopped operations in October of 2003 and will not be discussed here.

Aqua includes an instrument called the Advanced Microwave Scanning Radiometer (AMSR). AMSR has a 6.9 GHz (C-band) channel with a nominal 60-km spatial resolution. This instrument should be able to provide information about soil-water content in regions of low vegetation cover, less than 1 kg m$^{-2}$ vegetation water content. Aqua collects data at nominal times of 1.30 a.m. and p.m. local time.

As opposed to previous passive-microwave satellite missions, the Aqua project will include soil moisture as a product. The algorithm planned for use by NASA is the multiple-channel approach. The Japanese Aerospace Exploration Agency will also produce a soil-moisture product using one of four algorithms under consideration. Preliminary studies indicate that there is widespread radiofrequency interference (RFI) in the C-band channels. Therefore, it is likely that the most useful channels for soil moisture will be those operating at the slightly higher X band.

The Coriolis satellite includes the Windsat instrument, which is a multifrequency passive microwave radiometer system. It is similar to AMSR with some differences in frequencies and more polarization options. It is a prototype of one component of the next generation of operational polar orbiting satellites that the USA plans to implementing by the year 2010.

### Future Passive Systems

Programs are underway to develop and implement space-based systems with a 1.4 GHz (L-band) channel that would provide improved global soil-moisture information. Toward that goal the European Space Agency (ESA) is developing a sensor system called the Soil Moisture Ocean Salinity (SMOS) mission and NASA has initiated a mission called Hydros.

SMOS will utilize two-dimensional synthetic aperture radiometry at L band and provide 40-km H and V polarizations. The ability of SMOS to obtain multiple-incidence angle observations of the surface is the key element of the soil-moisture retrieval algorithm. The launch of the mission is anticipated in 2007.

Hydros will use an active and passive L-band instrument design for mapping at resolutions ranging from 3 to 40 km. Soil moisture will be estimated by Hydros using a microwave radiometer and radar data separately and in combination, taking advantage of the simultaneous, coincident, and complementary nature of the measurements. Hydros is scheduled for 2010.

## Active Microwave Satellite Observing Systems

### Current Active Systems

At present, several radar satellites are in orbit. ESA has operated a satellite synthetic aperture radar (SAR) series called ERS since 1991, which provides C-band VV data. It includes both a SAR and a scatterometer. Although numerous investigations have been conducted that attempt to utilize ERS SAR data, few results have been reported in the area of soil-moisture estimation. This is due to the limitations of using a single mid-frequency channel single-polarization SAR with an exact repeat cycle of 35 days.

The most recent satellite in the ERS series is called Envisat (launched in 2002) and it has a C-band advanced synthetic aperture radar (ASAR) with multiple polarization capabilities. It also offers the option of varying the incidence angle to allow for different viewing angles and more frequent coverage if angle is not a critical parameter in the application. Data from ASAR are just beginning to be made available to investigators.

The Canadian Space Agency operates a C-band satellite SAR called Radarsat, which offers HH polarization and has a variable viewing angle and a wide swath (large range of incidence angles). These choices offer more frequent temporal coverage of a particular region of the Earth if angle is not important.

There have been a several of attempts at using the single-channel SAR data to retrieve soil moisture. The general consensus from these studies is that there are too many physical variables that have to be known in order to derive soil moisture. These variables include soil moisture, surface roughness, and vegetation. Site-specific studies and creative selection of data sets have resulted in some degree of success.

### Future Active Systems

Japan will include an L-band SAR called PALSAR on the Advanced Land Observing System (ALOS) in 2005 or later. PALSAR will have a multipolarization mode as well as varying incidence angles. Canada is developing Radarsat-2, which will be similar to Radarsat-1 but will be fully polarimetric. The satellite will include a mode in which the spatial resolution is 3 m. The expected launch is in 2005.

As noted above, progress in SAR soil moisture mapping has been limited by the available data, single channel. The new generation of multipolarization SAR systems will provide at least one additional measurement. There is hope that more robust soil moisture retrieval methods can be developed as these data become available.

## List of Technical Nomenclature

| | |
|---|---|
| $\theta$ | Viewing angle (degrees) |
| $\sigma^0$ | Backscattering coefficient (decibels) |
| $b$ | Vegetation parameter ($m^2 kg^{-1}$) |
| $e$ | Emissivity (dimensionless) |
| $f$ | Frequency (GHz) |
| k | Dielectric constant (dimensionless) |
| $p$ | Polarization (dimensionless) |
| $r$ | Reflectivity (dimensionless) |
| $T$ | Physical temperature (degrees Kelvin) |
| $T_B$ | Brightness temperature (degrees Kelvin) |
| VWC | Soil moisture (volumetric percent) |
| $W_{veg}$ | Vegetation water content ($kg\,m^{-2}$) |

*See also:* **Evaporation of Water from Bare Soil**; **Evapotranspiration**; **Remote Sensing:** Organic Matter; **Water Cycle**

## Further Reading

Altese E, Bolognani O, Mancini M, and Troch P (1996) Retrieving soil moisture over bare soil from ERS-1 synthetic aperture radar data: sensitivity analysis based on a theoretical scattering model and field data. *Water Resources Research* 32: 653–661.

Bindlish R, Jackson TJ, Wood E *et al.* (2003) Soil moisture estimates from TRMM microwave imager observations over the southern United States. *Remote Sensing of Environment* 85: 507–515.

Dubois PC, van Zyl J, and Engman ET (1995) Measuring soil moisture with imaging radars. *IEEE Transactions on Geoscience and Remote Sensing* 33: 915–926.

Jackson TJ (1993) Measuring surface soil moisture using passive microwave remote sensing. *Hydrological Processes* 7: 139–152.

Jackson TJ and Schmugge TJ (1989) Passive microwave remote sensing system for soil moisture: some supporting research. *IEEE Transactions on Geoscience and Remote Sensing* GE-27: 225–235.

Jackson TJ and Schmugge TJ (1991) Vegetation effects on the microwave emission from soils. *Remote Sensing of Environment* 36: 203–212.

Jackson TJ, Le Vine DM, Swift CT, Schmugge TJ, and Schiebe FR (1995) Large area mapping of soil moisture using the ESTAR passive microwave radiometer in Washita '92. *Remote Sensing of Environment* 53: 27–37.

Mattikalli NM, Engman ET, Jackson TJ, and Ahuja LR (1998) Microwave remote sensing of temporal variations of brightness temperature and near-surface soil water content during a watershed-scale field experiment, and its application to the estimation of soil physical properties. *Water Resources Research* 34: 2289–2299.

Njoku EG and Kong J (1977) Theory for passive microwave remote sensing of near-surface soil moisture. *Journal of Geophysical Research* 82: 3108–3118.

Njoku EG, Jackson TJ, Lakshmi V, Chan TK, and Nghiem SV (2003) Soil moisture retrieval from AMSR-E. *IEEE Transactions on Geoscience and Remote Sensing* 41: 250–261.

Satalino G, Mattia F, Davidson MW *et al.* (2002) On current limits of soil moisture retrieval from ERS-SAR data. *IEEE Transactions on Geoscience and Remote Sensing* 40: 2438–2447.

Shutko AM (1986) *Microwave Radiometry of Water and Terrain Surfaces*. Moscow: Nauka Publishing House.

Ulaby FT, Moore RK, and Fung AK (1986) *Microwave Remote Sensing: Active and Passive*, vol. III, *From Theory to Application*. Dedham, MA: Artech House.

Wagner W, Lemoine G, and Rott H (1999) A method for estimating soil moisture from ERS scatterometer and soil data. *Remote Sensing of Environment* 70: 191–207.

Wang JR and Schmugge TJ (1980) An empirical model for the complex dielectric permittivity of soils as a function of water content. *IEEE Transactions on Geoscience and Remote Sensing* GE-18: 288–295.

Wigneron JP, Waldteufel P, Chanzy A, Calvet JC, and Kerr Y (2000) Two-dimensional microwave interferometer retrieval capabilities over land surfaces (SMOS mission). *Remote Sensing of Environment* 73: 270–282.

# RHIZOSPHERE

**A C Kennedy and L Z de Luna**, USDA Agricultural Research Service, Pullman, WA, USA

## Introduction

The rhizosphere is a narrow zone of soil surrounding and influenced by the root of vascular plants. This zone is characterized by intense biological activity owing to the release of root exudates, which stimulate or inhibit rhizosphere organisms. The interactions between the soil, the plant, and organisms that comprise the rhizosphere characterize the complexity and the dynamics of this region. Plant/microbe and plant/faunal interactions abound in the rhizosphere. These interactions range from symbiotic relationships such as $N_2$ fixation and mycorrhizal associations to pathogenic interactions. Management strategies such as bioremediation and biological control may be successful when rhizosphere ecology is considered. A greater understanding of the rhizosphere and its effects on organisms that inhabit this area will allow for manipulations that benefit plant production and the environment.

Hiltner coined the term 'rhizosphere' in 1904 to describe the zone of soil under the influence of plant roots. His observations were initially based on the interactions between symbiotic $N_2$-fixing bacteria and the legume root, but were then expanded to include all interactions. This zone of soil ranges from only a few hundred micrometers to greater than 5 mm from the root surface (**Figure 1**). Within this region increased root exudation and changes in physical and chemical soil properties may increase the activity and populations of most soil microorganisms. The rhizosphere is constantly in a state of dynamic flux. Several overlapping gradients in this zone vary with time and distance from the root, and influence rhizosphere colonization by soil organisms.

These gradients and the many interactions within them make the understanding of this zone difficult. The most dramatic gradient in the rhizosphere is formed by nutrients released by the root. Plant roots excrete various organic substances into the

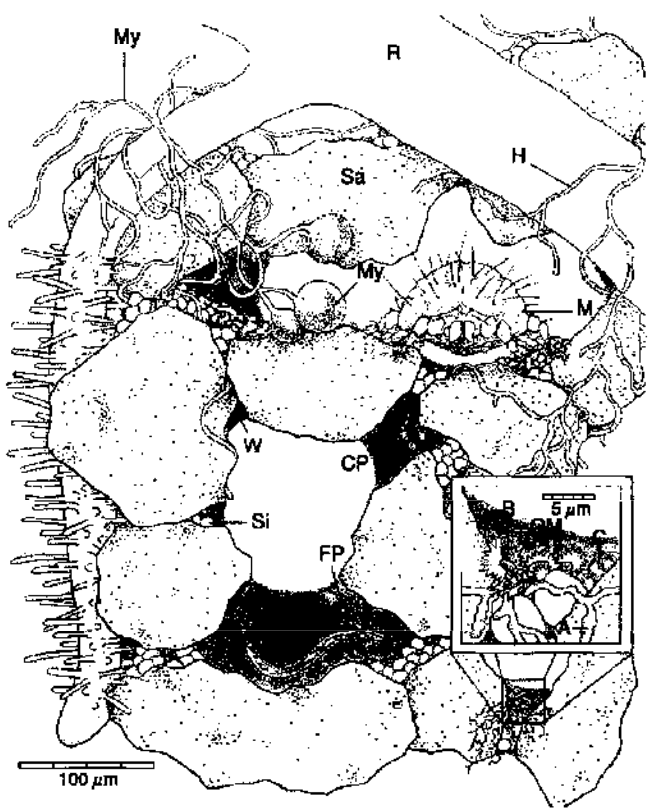


**Figure 1** The rhizosphere habitat. R, root; B, bacteria; A, actinomycetes; My, mycorrhizal spores and hyphae; H, hyphae of saprophytic fungi; N, nematode; CP, ciliate protozoa; FP, flagellate protozoa; M, mite; Sa, sand; Si, silt; C, clay; OM, organic matter; W, water. Original drawing by Kim Luoma: Reproduced from Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (1998) *Principles and Application of Soil Microbiology*. Upper Saddle River, NJ: Prentice Hall.

rhizosphere, and these substances provide a rich source of nutrients for the microbial community. The greatest exudation arises from those cells involved in cell elongation and lateral root formation. Carbon compounds decline exponentially as the distance from the root increases. Increased exudation can result from extremes in temperature, water stress, P deficiency, increased light intensity, and increased populations of microorganisms. Herbicides, pathogens, foliar treatments, and symbiotic associations may also influence exudation rate. Decreased exudation from the root can result from N deficiency and decreased light intensity. There may also be a relationship between exudate pattern and microbial populations colonizing the root surface (rhizoplane) or the cell layers of the root itself (endorhizosphere).

Plant exudates of high molecular weight, which are primarily released by root tips, form a sheath of slime on the external surface of the root called mucilage. In contrast, mucigel contains plant mucilage and other plant products, as well as bacterial cells and their products. These secretions are important because they help to form and maintain stable soil aggregates. Most soils are extremely low in readily available C, and C-containing substances leaked by the root are of major importance to microbial growth. Exudates can be sugars, amino acids, vitamins, tannins, alkaloids, phosphatides, and other substances such as growth factors, fluorescent substances, nematode cyst or egg-hatching factors, and fungal growth stimulants and inhibitors. Sugars and amino acids, the most-studied root exudates, provide readily available nutrients to the rhizosphere inhabitants and may be a factor in governing susceptibility or resistance of plants to root-infecting fungi.

Plant transpiration can induce severe changes in the water potential at the root surface, which may cause water stress for the organisms near the root as well as reducing the movement of necessary nutrients. Dry soils may inhibit the growth of the root and enhance the release of exudates. Waterlogged soils on the other hand reduce oxygen availability and may increase exudation due to root injury. These conditions may influence the quantity of exudates released and the growth of the microbial and macrofaunal populations in the rhizosphere. Root exudates and the release of $H^+$, $HCO_3^-$, or organic compounds, as well as nitrogen nutrition influence rhizosphere pH. Rhizosphere pH can fluctuate as much as one or two pH units over very short time periods. These changes in the soil pH can also affect the growth and survival of the soil biota both positively and negatively. In general, bacteria are negatively affected by low pH, while fungi are not as sensitive to extremes in pH.

Although there is an increase in microbial activity in the rhizosphere, the distribution of those organisms on the root surface is quite variable. Electron micrographs show that microorganisms colonize only 7–15% of the actual root surface. Bacteria are found on the root surface in small colonies at the junctions of epidermal cells or at tears in the root surface (Figure 2). Colonization of the root surface and the rhizosphere is dependent upon factors involving the organisms, the plant, and the environment (Table 1). Microorganisms are thought to move down the root by three different mechanisms: (1) motility of the organism (active transport); (2) water movement on the surface of the root (passive transport); and (3) movement of microorganisms on the root apex as the root cells elongate (passive transport).

## Organisms in the Rhizosphere

### Bacteria

Rhizosphere bacteria play vital roles in plant nutrition, growth promotion, and disease interactions. Several studies have indicated that bacteria are the most numerous inhabitants of the rhizosphere, although they account for only a small portion of the total

**Figure 2** Colonization of the root by bacteria. Note the clumping of colonies at intersection of cells. (Reproduced from Begonia MFT, Kremer RJ, Stanley L, and Jamshedi A (1990) Association of bacteria with velvetleaf roots. *Transactions of the Missouri Academy of Science* 24: 17–26.

**Table 1** Factors influencing rhizosphere colonization

| |
|---|
| *Microbial/macrofaunal characteristics* |
| Versatile nutrient requirements |
| Fast initial growth rate |
| Enhanced cellulase or other enzyme production |
| Antibiotic production |
| Chromophore influence on metal nutrition |
| Unique physiological attributes |
| Tolerance to antibiotics, fungicides, or other chemicals |
| *Plant characteristics* |
| Plant species/cultivar |
| Plant and root age |
| An unoccupied niche in soil, rhizosphere, or rhizoplane |
| Altered plant genetics |
| Foliar treatments |
| Soil-applied pesticides |
| Exudation patterns |
| Cropping sequence and rotation |
| *Environment* |
| pH |
| Soil type |
| Soil texture |
| Soil porosity and compaction |
| Soil moisture |
| Soil atmosphere |
| Temperature |
| $O_2$ availability |
| Fertility, organic matter, and available nutrients |
| Pesticide and herbicide application |

Adapted from Kennedy AC (1998) The rhizosphere and spermosphere. In: Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (eds) *Principles and Applications of Soil Microbiology*, pp. 389–407. Upper Saddle River, NJ: Prentice-Hall, with permission.

biomass due to their small size. Typically, a 1-g sample of rhizosphere soil contains approximately $10^8$–$10^{12}$ bacterial cells; Gram-negative bacteria predominate in the rhizosphere; of these, the most efficient bacterial root colonizers belong to the genus *Pseudomonas*. Actinomycetes are filamentous chemoorganotrophic bacteria that form asexual spores. Populations of these bacteria are usually higher in rhizosphere than in nonrhizosphere soil. There are indications that actinomycetes may comprise higher percentages of the total bacterial numbers than fluorescent pseudomonads in the root zone of some plants. Although some actinomycetes may cause plant diseases, many species have the potential to produce antibiotics that could impact plant–microbe interactions, such as nodulation in legumes or biocontrol efforts against plant pathogens.

### Fungi

Fungi are eukaryotic, nonphotosynthetic, spore-forming organisms that are primarily responsible for residue decomposition and organic matter accumulation through the mineralization of cellulose, lignin, and other organic materials. In the rhizosphere and rhizoplane, fungi may function as symbionts, pathogens, nutrient sinks, and a food source for other soil microorganisms. Fungal mycelia also bind soil particles together to form aggregates.

### Fauna

Animals that spend part or all of their lives in close association with plant roots and have significant effects on the structural and functional properties of the rhizosphere make up the rhizosphere fauna. Soil fauna are classified into three types based on body size. Single-celled animals such as protozoa belong to the microfauna, while rotifers, nematodes, and microarthropods are classified as mesofauna. Earthworms, millipedes, centipedes, and insects comprise the macrofauna.

**Microfauna (protozoa)** Protozoa are eukaryotic, unicellular organisms, most of which are microscopic in size, although some members may attain macroscopic dimensions. While the group exhibits differences in feeding behavior, all require a water film for metabolic activity. It is estimated that there are between $10^4$ and $10^5$ protozoa per gram of soil. Populations of protozoa are generally higher in rhizosphere soil (R) than bulk soil (S), with large numbers occurring in hotspots such as around dead plant roots. Protozoa feed on bacteria and fungi and thus contribute to the mineralization of N, leading to an increase in N uptake by plants. When rhizosphere protozoa prey on bacteria, up to 60% of the excess N and P is excreted in forms that can be readily utilized by plants. Plants may also benefit from the suppression of pathogenic bacteria and fungi by protozoa in the rhizosphere. Detrimental effects of protozoa include a suppression of *Rhizobium* populations in the rhizosphere, resulting in reduced root nodulation of *Phaseolus vulgaris*.

**Mesofauna (nematodes and microarthropods)** Nematodes are microscopic, unsegmented, round worms, usually 0.5–2.0 mm in length. There are five trophic groups of nematodes: bacterivores, which feed on bacteria; fungivores, which feed on fungal mycelia; predators, which feed on other nematodes and small invertebrates; omnivores, which feed on a variety of foods; and herbivores, which are plant parasites. Nematodes play a significant role in mineral cycling. Their consumption of bacteria and other food sources results in excretion of N, mainly in the form of ammonia, and enhanced mineralization of P and S in some systems. Nematodes are also very important in agriculture, because some species are parasitic on crops. Feeding by these plant-parasitic nematodes results in wilting, nutrient-deficiency symptoms, yield losses, and other problems linked to damaged root systems.

Microarthropods such as mites (Acari: Arachnida) and springtails (Collembola: Insecta), are a major fraction of the mesofauna present in soils; however, they generally do not harm plants. Although some soil mites prey on nematodes and other small arthropods, they are generally fungal feeders. A few species pierce roots and suck plant sap, ingest roots, or feed on decomposing plant matter.

## Interactions in the Rhizosphere

There are a myriad of interactions occurring in the rhizosphere (Table 2). The most studied of these are $N_2$-fixation, mycorrhizal associations, plant growth promotion, and plant growth inhibition. In recent years, bioremediation and biological control have also been investigated.

### Dinitrogen-Fixation Symbiosis

Dinitrogen-fixing bacteria can make otherwise unavailable $N_2$ available to the plant. Rhizobia and related bacteria form nodules on the roots of plants. This interaction can be mutually beneficial as the plant provides nutrients and protection for the bacteria and the bacteria provide N for the plant. Other $N_2$-fixing organisms form associations with plants. These bacteria colonize the root surface and interior cells, and intercellular spaces between the cortex and endodermis and in the xylem cells of grasses. Along with plant-available N, they may provide other plant growth-promoting compounds also.

**Table 2** Activities of organisms in the rhizosphere

| |
|---|
| *Decomposition of plant residue and organic material* |
| Humus synthesis |
| Mineralization of organic N, S, and P |
| *Increased plant nutrient availability of P, Mn, Fe, Zn, Cu* |
| Symbiotic mycorrhizal associations |
| Production of organic chelating agents |
| Oxidation–reduction reactions |
| P solubilization |
| *Biological dinitrogen fixation* |
| Free-living bacteria and blue-green algae |
| Associative microorganisms |
| Symbiotic legume and nonlegume |
| *Plant growth promotion* |
| Production of plant growth hormones |
| Protection against root pathogens and pseudopathogens |
| Enhanced nutrient-use efficiency |
| *Plant growth inhibition* |
| Pathogenicity by root pathogens |
| Growth reduction by deleterious rhizobacteria |
| *Biological control* |
| Reduction of plant diseases |
| Inhibition of soil nematodes and insects |
| Reduction in weed growth |
| *Bioremediation* |
| Degradation of industrial and agricultural chemicals |
| Volatilization of metal contaminants |
| Hyperaccumulation of metals in plant tissues |
| *Enhanced drought tolerance of plants* |
| *Improved soil aggregation* |

Adapted from Kennedy AC (1998) The rhizosphere and spermosphere. In: Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (eds) *Principles and Applications of Soil Microbiology*, pp. 389–407. Upper Saddle River, NJ: Prentice-Hall, with permission.

### Mycorrhizal Associations

Mycorrhizal fungi form a mutualistic symbiosis with plants and infect roots without causing root disease. These fungi can be found in the rhizosphere of most plants and form associations with all gymnosperms and more than 83% of dicotyledonous and 79% of monocotyledonous plants. Mycorrhizal fungi can form structures either on the outside (ectomychorrhizae) or inside (endomycorrhizae) of plant roots. The fungal hyphae allow the roots to contact a greater volume of soil. Some types of mycorrhizal fungi increase the solubilization of nutrients such as phosphorus. They assist the plant in increasing nutrient absorption, especially in stressed environments (e.g., phosphorus- and water-deficient soils), selective ion uptake, and provide protection from extremes in the environment. Plant exudation patterns may be altered after colonization by these fungi, thus affecting the rhizosphere microbial and macrofaunal communities. The fungi may also protect plant roots from invasion by pathogens. Extraradicle hyphae of endomycorrhizae secrete glomalin, a glycoprotein that aggregates soil particles, increasing water-stable aggregates and improving soil structure. This association can increase the survival and growth of a plant, especially in adverse or low-nutrient environments, and may have potential in the revegetation of disturbed sites.

### Plant Growth-Promoting Organisms

Plant growth-promoting organisms are those organisms that in some way enhance the growth of plants. Plant growth-promoting organisms enhance seed germination and plant growth and are often specific bacteria or fungi. Many different mechanisms are responsible for plant growth promotion. These organisms have been used in biological control programs to protect plants against plant pathogens, as biofertilizers to fix atmospheric $N_2$, and in phytostimulation, which directly improves plant growth through the production of plant growth regulators. Some may make nitrogen more available to plants and have the potential to reduce application of inorganic fertilizers. These organisms may produce plant growth-stimulating compounds such as gibberellic and indoleacetic acid. The reduction in pathogen load may also be a mechanism of plant growth promotion. Iron-chelating compounds called siderophores can function to make iron more available to an organism and less available to the plant pathogen, thus impairing the pathogen. Since iron is a key nutrient in metabolism and a cofactor in enzyme activity, microbial growth may be influenced by these siderophores. The microbial populations that colonize the interior of the root and form intimate associations with the root are considered endophytes. Endophytic bacteria may increase plant growth, confer disease resistance, and aid the plant in withstanding stresses such as drought. The success of endophytes in colonizing the rhizosphere may be the ability of the microorganism to compete with other bacteria on the external or internal portions of plant roots. Introduction of plant growth-promoting organisms may change the overall composition of the microbial community, especially in the rhizosphere.

### Plant Growth-Inhibiting Organisms

Organisms that inhibit plant growth colonize the rhizosphere and can alter plant development. Plant growth inhibition can occur with or without disease symptoms. The rhizosphere harbors pathogens such as fungi and bacteria that attack germinating seeds and plant roots. Pathogens may produce extracellular enzymes that liquefy cell walls, thereby allowing degradation of the cell contents. Some pathogens produce toxins that affect the permeability of the cell membrane or inhibit an enzyme, leading to interruptions in a metabolic pathway. These microbial toxins may affect many species across different plant families or they may be host-specific. Polysaccharides produced by some pathogens may block vascular tissues and cause wilting. Excessive production of auxins and cytokinins, resulting from integration of a tumor-inducing bacterial plasmid into the plant genome, leads to gall formation in the roots. Feeding by plant-parasitic nematodes exposes root cells to destructive enzymes and toxic metabolites, which ultimately result in decreased water and nutrient uptake in plants. Puncture wounds from the feeding process could also become infection courts for the entry of other microbial pathogens into the roots. Deleterious rhizobacteria (DRB) inhibit plant growth without causing obvious disease symptoms. These organisms are present on the root surface and may occur in the intercellular spaces of cortical cells. DRB do not parasitize the plant, but survive on organic compounds released by plant root cells. Their detrimental effects on plant growth are due mainly to the production of metabolites that disturb plant physiological processes, including phytotoxins, plant growth regulators, volatile substances, or antibiotics. Their colonization and inhibition are often found to be species- or cultivar-specific, and root exudates may play a part in some plant–DRB interactions.

### Bioremediation

Industrial development and intensification of agriculture in the past several decades have contributed to environmental contamination, releasing hazardous wastes into terrestrial ecosystems. Due to their toxic and carcinogenic nature, and tendency to bioaccumulate in humans and other organisms, the timely and cost-effective clean-up of contaminated areas has become a priority. Plant–microbial systems have shown potential in bioremediation, a process that utilizes the natural role of plants and microorganisms in transformation, mineralization, and complexation of organic and inorganic pollutants. In contrast to conventional soil remediation technologies, bioremediation is less costly, utilizes relatively simple technology, and is minimally disruptive to the contaminated site. Bioremediation may be facilitated by plant root exudates that show high binding affinities to certain pollutants or by rhizosphere microorganisms that mineralize contaminants into their nontoxic derivatives.

Plant roots, together with their associated microbial communities, may enhance mineralization of hazardous compounds in the soil. The 'rhizosphere effect' on bioremediation may occur in several ways. First, it is thought that carbon substrates from roots stimulate increases in xenobiotic degrader populations and activity. Second, the presence of dense populations of diverse microorganisms in the rhizosphere may enhance bacterial conjugation, which could result in the evolution of new pathways of degradation. Third, plants may provide a substrate that increases the production of the enzymes involved in mineralization of the hazardous compound. Rhizosphere organisms can influence the degradation of recalcitrant compounds in the soil (**Table 3**). Rhizosphere microorganisms may also facilitate the

**Table 3** Plant species shown to facilitate microbial degradation of hazardous chemicals in the rhizosphere

| *Plant* | | |
|---|---|---|
| *Common name* | *Species* | *Chemical* |
| African clover | *Trifolium africanum* | 2,4-D[a] |
| Alfalfa | *Medicago sativa* | PAHs[b] |
| Bahia grass | *Paspalum notatum* | TCE[c] |
| Beans | *Phaseolus vulgaris* | Diazinon[e], Parathion[f], Temik[d] |
| Cattail | *Typha* sp. | Surfactants[g] |
| Corn | *Zea mays* | Atrazine[h], Temik[d], fuel oil hydrocarbons |
| Cotton | *Gossypium hirsutum* | Temik[d] |
| Flax | *Linum usitatissimum* | 2,4-D[a] |
| Goldenrod | *Solidago* sp. | TCE[c] |
| Lespedeza | *Lespedeza cuneata* | TCE[c] |
| Loblolly pine | *Pinus taeda* | TCE[c] |
| Meadow brome | *Bromus biebersteinii* | 2-Chlorobenzoic acid |
| Peas | *Pisum sativum* | Diazinon[e] |
| Prairie grasses | *Andropogon gerardi, Sorghastrum nutans, Panicum virgatum, Elymus canadensis, Schizachyrium scoparius, Bouteloua curtipendula, Agropyron smithii, Bouteloua gracilis* | PAHs[b], TCE[c] |
| Reed | *Phragmites communis* | VOCs[i], 4-chlorophenol |
| Rice | *Oryza sativa* | Propanil[j], *p*-nitrophenol, Alachlor[k], carbofuran[l] residues |
| Ryegrass | *Lolium perenne* | Anthracene, TCE[c], PAHs[b] |
| Salt marsh plants | *Distichlis spicata, Juncus gerardi, Phragmites australis, Spartina alterniflora, Sporobolus airoides* | PAHs[b] |
| Sorghum | *Sorghum vulgare* | PAHs[b] |
| Soybean | *Glycine max* | TCE[c], surfactants[g] |
| Sugarcane | *Saccharum officinarum* | 2,4-D[a] |
| Summer cypress | *Kochia scoparia* | Atrazine[h] |
| Tall fescue | *Festuca arundinacea* | PAHs[b] |
| Tobacco | *Nicotiana tabacum* | Maleic hydrazide[m] |
| Wheat | *Triticum aestivum* | Diazinon[e], Mecoprop[n], 2,4-D[a], MCPA[o] |
| Wild rye | *Elymus dauricus* | 2-Chlorobenzoic acid |
| Zinnia | *Zinnia angustifolia* | Mefenoxam[p] |

[a] 2,4-Dichlorophenoxyacetic acid.

[b] Polycyclic aromatic hydrocarbons.

[c] Trichloroethylene.

[d] 2-Methyl-2(methylthio)propionaldehyde *O*-(methylcarbamoyl)oxime.

[e] *O*-*O*-Diethyl-*O*-(2-isopropyl-6-methyl-4-pyrimidyl) phosphorothioate.

[f] *O*-*O*-Diethyl-*O*-*p*-nitrophenyl phosphorothioate.

[g] Dodecyl linear alkylbenzene sulfonate, dodecyl linear alcohol ethoxylate, dodecyltrimethylammonium chloride.

[h] 2-Chloro-4-ethylamino-6-isopropylamino-*s*-triazine.

[i] Volatile organic compounds (benzene, biphenyl, chlorobenzene, dimethylphthalate, ethylbenzene, naphthalene, *p*-nitrotoluene, toluene, *p*-xylene, bromoform, choloroform, 1,2-dicholoroethane, tetracholoroethylene, 1,1,1-trichloroethane.

[j] 3′,4′-Dichlorpropionanilid-*N*-(3,4-dichlorphenyl) propanamid.

[k] 2-Chloro-*N*-(2,6-diethylphenyl)-*N*-(methyoxymethyl)acetamide.

[l] 2,3-Dihydro-2,2-dimethyl-7-benzofuranyl methylcarbamate.

[m] 1,2-Dihydro-3,6-pyridazinedione.

[n] 2-(2-Methyl-4-chlorophenoxy)propionic acid.

[o] 2-Methyl-4-cholorophenoxyacetic acid.

[p] [(*R*)-*N*-(2,6-dimethyl-phenyl)-(methoxyacetylamino) propionic acid methyl ester].

removal of metal contaminants such as mercury, selenium, and zinc from the soil. Bacterial isolates and consortia have been extensively studied in bioremediation, but fungi, actinomycetes, mycorrhizal associations, and even rhizosphere fauna have shown degrading abilities. Clearly, the greatest challenge in bioremediation technology is the elucidation of mechanisms involved in remediation of contaminated sites by plants and their associated organisms. This knowledge will allow for the selection and engineering of the best plant–microbe combinations and optimization of the remediation process.

### Biological Control

The rhizosphere is the location for biological control interactions that make use of pathogens, parasites, or other predators to reduce the population or activity of another organism. Three major strategies for biological control are classical, inundative, and integrated management. The classic approach involves the use of natural enemies for release, dissemination, and self-perpetuation on target pests. The inundative approach is most often used in the rhizosphere and involves the addition of a virulent strain of a biocontrol agent to suppress pests. In this case, the biocontrol agent is not self-sustaining and must be applied to the target host every season. The integrated management approach is a broad strategy that involves management practices to conserve or enhance native enemies of pests.

Biological control can be used in the management of pathogens, nematodes, and weeds. The use of biological control in disease management allows for alternatives to chemical control (Table 4). The interactions of organisms in the rhizosphere and the ecological significance of these interactions that occur in the root and seed environment assist in the success of biocontrol strategies. The reduction in plant pathogens such as *Fusarium oxysporum*, *Gaeumannomyces graminis*, *Pythium*, and *Phytophthora* species due to suppressiveness of some soils is thought to be due to physiochemical and microbial soil factors. The suppression is specific and may have various mechanisms. The mechanisms can include antibiosis, production of siderophores or volatile compounds, parasitism, competition for nutrients, and competition for ecologic niches and induced disease resistance. Crop monoculture can, with time, result in a decrease in disease due to the development of disease-suppressive soils. Other management practices that may alter the rhizosphere environment and be considered as biocontrol approaches are crop rotation, soil amendments, cover cropping, fumigation, or soil solarization.

Fungal pathogens of protozoa and nematodes can be used in biological control of these plant pests. Fungal mycelial growth into the cortical cells is thought to assist in controlling clubroot disease caused by protozoa. Plant-parasitic nematodes alter the exudate patterns of the plant as they feed. This in turn changes the microbial community of the rhizosphere at these sites, which affects the colonization of the fungal and bacterial antagonists of the nematodes. One of the more interesting and most studied fungal interactions for biological control is the use of nematode-trapping fungi to reduce nematode populations.

**Table 4** Examples of plant diseases for which potential biological control agents have been identified

| *Disease* | *Host plant(s)* | *Antagonist* |
|---|---|---|
| Damping-off (*Pythium* spp.) | Beet | *Actinoplanes* sp. |
| | Corn | *Burkholderia cepacia* PHQM 100 |
| | Cucumber | *Pseudomonas aureofaciens* 63-28 |
| | | *Serratia plymuthica*; *Pythium oligandrum*; *Trichoderma longibrachiatum* CECT 2606 |
| | Bird's foot trefoil | *Pseudomonas fluorescens* VO61 |
| | Eggplant | *Cladorrhinum foecundissimum* |
| Sheath blight (*Rhizoctonia solani*) | Rice | *Pseudomonas fluorescens* VO61 |
| Root rot (*Rhizoctonia solani*) | Wheat | *Bacillus* sp. |
| Take-all (*Gaeumannomyces graminis* var. *tritici*) | Wheat | *Pseudomonas fluorescens* Q8r1-96 |
| *Fusarium* wilt (*Fusarium* spp.) | Chickpea | *Bacillus subtilis* GB03 |
| | Cucumber | *Paenibacillus* sp. 300 |
| | Tomato | *Pseudomonas* spp.; *Fusarium oxysporum* Fo47[a]; *Fusarium solani*[a]; *Fusarium* spp.[a] |
| | Radish | *Pseudomonas putida* WCS417 |
| Root rot (*Phytophthora* sp.) | Pepper | *Trichoderma harzianum* 2413 |
| *Verticillium* wilt (*Verticillium dahliae*) | Eggplant | *Talaromyces flavus* |
| | Pepper | *Pythium oligandrum* |

[a] Nonpathogenic strain

These predatory fungal species produce many different types of trapping structures such as constricting rings and adhesive knobs or hyphal meshes. However, the factors that induce significant predaceous behavior need to be studied further.

Soil microorganisms that have potential in weed management have been increasingly studied in recent years. Fungal and bacterial pathogens of weeds that attack weed seeds or cause root damage are in various stages of development as biological control agents. These organisms may reduce weed populations through weed inhibition or mortality. Deleterious rhizobacteria that are host-specific have been isolated from many grass and broadleaf weed species. Bacteria have been shown to inhibit winter annual grass weeds and velvetleaf (*Abutilon theophrasti* Medik.). Rhizosphere fungi may also be used to control weeds. Propagules of *Fusarium oxysporum* f. sp. *erythroxyli* applied to soil results in vascular wilt symptoms and death of cocoa plants. The saprophytic, soilborne fungus *Gliocladium virens* produces viridiol, a steroidal phytotoxin that causes severe root necrosis in susceptible weed seedlings. Isolates belonging to *Fusarium* spp. have also shown promise for biological control of leafy spurge. Future development of biological control strategies require ecologic investigations of these systems and will enhance pest-management programs.

## Summary

The rhizosphere is a dynamic region governed by complex interactions between plants and the organisms that are in close association with the root. The composition and pattern of root exudates affect microbial activity and population numbers, which in turn have an impact on the nematodes and microarthropods that share this environment. Beneficial or harmful relationships exist between rhizosphere organisms and plants, which ultimately affect root function and plant growth. In addition, the rhizosphere may include organisms that do not directly benefit or harm plants but clearly influence plant growth and productivity. A better understanding of the soil–root and soil–seed interface is needed to manage microorganisms, increase plant growth, and reduce the impact of plant production and agriculture on the environment. The benefits of studying the rhizosphere include the use of plant growth-promoting organisms and the suppression of plant diseases and weeds using biocontrol agents. Rhizosphere organisms can also be used to enhance the formation of stable soil aggregates and as bioremediation agents of contaminated soils. With greater understanding of the ecology and biota in the rhizosphere, this zone of increased nutrients, biotic activity, and interactions can be manipulated to improve plant productivity and environmental quality.

*See also:* **Bacteria:** Plant Growth-Promoting; Soil; **Fauna**; **Fungi**; **Microbial Processes:** Environmental Factors; **Nematodes**; **Pollutants:** Biodegradation; **Root Exudates and Microorganisms**

## Further Reading

Anderson TA and Coats JR (1994) *Bioremediation Through Rhizosphere Technology*. ACS Symposium Series 563. Washington, DC: American Chemical Society.

Andow DA, Ragsdale DW, and Nyvall RF (1997) *Ecological Interactions and Biological Control*. Boulder, CO: Westview Press.

Crossley DA Jr, Mueller BR, and Perdue JC (1992) Biodiversity of microarthropods in agricultural soils: relations to processes. *Agriculture, Ecosystems and Environment* 40: 37–46.

Curl EA and Truelove B (1986) *The Rhizosphere*. Berlin, Germany: Springer-Verlag.

Hornby D (1990) *Biological Control of Soil-borne Plant Pathogens*. Oxford, UK: CAB International.

Keister DL and Cregan PB (1991) *The Rhizosphere and Plant Growth*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kennedy AC (1998) The rhizosphere and spermosphere. In: Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (eds) *Principles and Applications of Soil Microbiology*, pp. 389–407. Upper Saddle River, NJ: Prentice-Hall.

Lynch JM (1990) *The Rhizosphere*. Chichester, UK: John Wiley.

Paoletti MG (1999) *Invertebrate Biodiversity as Bioindicators of Sustainable Landscapes*. Amsterdam, The Netherlands: Elsevier.

Papendick RI and Campbell GS (1975) Water potential in the rhizosphere and plant and methods of measurement and experimental control. In: Bruehl GW (ed.) *Biology and Control of Soil-Borne Pathogens*, pp. 39–49. St. Paul, MN: American Phytopathological Society.

Pinton R, Varanini Z, and Nannipieri P (2001) *The Rhizosphere*. New York: Marcel Dekker.

Upadhyay RK, Mukerji KG, and Chamola BP (2000) *Biocontrol Potential and its Exploitation in Sustainable Agriculture*, vol. 1, *Crop Diseases, Weeds, and Nematodes*. New York: Kluwer Academic/Plenum Publishers.

Waisel Y, Eshel A, and Kafkafi U (1996) *Plant Roots: The Hidden Half*. New York: Marcel Dekker.

Whipps JM (2001) Microbial interactions and biocontrol in the rhizosphere. *Journal of Experimental Botany* 52: 487–511.

# RICHARDS, LORENZO A.

**W R Gardner**, Texas A & M University, USA

Most soil physicists, past or present, would agree that L.A. Richards was one of the pre-eminent soil physicists of the twentieth century. Although he was primarily an experimentalist, his name will always be associated with the equation describing the movement of water in unsaturated soil, one of the most useful equations in soil physics.

Richards was born in Fielding, Utah, on 24 April, 1904, and died in Riverside, California on 12 March, 1993. Fielding is a small agricultural community some 25 miles (40 km) west of Logan, which is the site of Utah State University. This proximity was fortunate for Ren (as he was known to his friends), because it made it possible for him to attend Utah Agricultural College (UAC) (as it was then called), as did three of his brothers, Bert L. Richards, Matthias C. Richards, and Sterling J. Richards. He married Zilla Linford, whose family was also replete with scientists. Among them were her brothers Maurice, Leon, Hooper, and Henry. Ren, Sterling, and Leon all majored in physics. Ren and his brother Sterling eventually ended up in Riverside, California, working in soil physics.

All three studied physics at the UAC under Willard Gardner, who received the eighth PhD in physics from the University of California in Berkeley. From this association came Ren's first scientific publications entitled 'The usefulness of capillary potential in soil moisture and plant investigations.' This paper summarized the current understanding of capillary potential as described by Edgar Buckingham and later expanded by Willard Gardner. It went on, however, to describe the principles and construction of the tensiometer and proposed a tension plate apparatus. The latter was intended for use on small samples of soil in the laboratory. He included in the paper water retention data as a function of capillary potential for four soils, some of the first data published. Following the work of Buckingham and Gardner, Richards also extended Darcy's law to unsaturated soils:

$$V = -K\,\mathrm{grad}\Phi \qquad [1]$$

where $V$ is the flux density of the water, $K$ is the permeability, and $\Phi$ is the capillary potential. Possibly influenced by his older brother Bert, who was a plant scientist, Ren went on to consider the availability of water to plants as it moved through the soil to plant roots. He pointed out that the term 'availability' involves two notions: namely the ability of plant roots to absorb the water with which it is in contact and the readiness or velocity with which the water moves to replace that which has been used by the plant. The first was dependent upon the capillary potential or the security with which the water was held by the soil combined with the osmotic potential or salinity. He was to return often to the problem of the measurement of this property in his later career. The second factor or rate of flow of water to the plant roots should be described by the appropriate solution of eqn [1]. Eqn [1], when combined with the equation of continuity:

$$q = -\rho\partial\theta/\partial t \qquad [2]$$

yields:

$$\begin{aligned}\partial\theta/\partial t = {} & \partial(D\partial\theta/\partial x)\partial x \\ & + \partial(D\partial\theta/\partial y)\partial y \\ & + \partial(D\partial\theta/\partial z)\partial z \\ & + \partial K/\partial z \qquad [3]\end{aligned}$$

where $\theta$ is the soil water content and $D$ is the soil water diffusivity and $\rho$ is the bulk density of the soil.

Eqn [3] is known as Richards' equation. It was not widely attributed to him until towards the end of his scientific career, but will undoubtedly always carry his name. The equation can also be written in terms of the matric or capillary potential $\psi$. There seems to be little doubt that eqn [3], or at least the capillary form of the equation, was known at Logan during the time that Ren was a graduate student there; however, it was not published until Richards included it as an important element in his PhD thesis from Cornell. It is appropriate that he be given the credit for it. His thesis, published in 1931, was titled 'Capillary conduction of liquids through porous medium.' In this Richards used two parallel porous fired ceramic plates to apply a differential pressure across a sample of soil and measured the flux between the plates. By varying the pressure across the plates he could maintain a variable degree of unsaturation in the soil. He had suggested the use of porous plates to evaluate the capillary potential of soil samples in his master's thesis. His PhD was apparently the first actual use of such plates.

After graduation from Cornell University with his PhD, Richards, like most physicists of that day, had difficulty in finding a job to his liking. Finally in 1935 he was offered a position at Iowa State University with a joint appointment in both the Physics and the Agronomy Departments. Here he began to apply his ideas to real problems and began to make his mark as an experimental physicist.

At that time the tensiometer was largely a novelty. At Logan he was introduced to the principle. The first attempts usually consisted of a glass funnel to which a porous plate was attached across the mouth of the funnel. Securing the plate to the funnel without leakage proved almost impossible. It was most fortuitous for the future of soil physics that Zilla Richards developed into a skilled ceramist. Ren and Zilla's collaboration eventually resulted in ceramic tubes which could be sealed very effectively to copper (and later acrylic) tubing and remain leakproof for days and eventually weeks. The diffusion of air through the ceramic then became the major limitation to their use, and this could be dealt with by periodic refilling with de-aired water.

Richards studied the use of tensiometers in field soils and in lysimeters and demonstrated their practicality and utility. During this time he also explored a number of other physical properties of soils, including the use of thermojunctions for the measurement of the relative humidity of soils. This particular problem was to occupy much of his time for the remainder of his career. Ren devised a method for measuring the flow of water in soils. He and his only student, M.B. Russell, studied the heat of wetting of soils and he and other colleagues pursued the use of tensiometers in the field.

In 1938 Ren was offered a position at the US Salinity Laboratory. Here he established a soil physics program which attracted visitors from all over the world and found an ideal scientific home for his researches for the remainder of his career. At this laboratory he continued to perfect the tensiometer and developed the pressure-membrane and pressure-plate apparatus. The pressure-membrane apparatus was initially conceived as a means of extracting water from soil by applying pressure to a soil sample supported on a porous membrane, with the entire apparatus contained in a brass chamber vented to the atmosphere below the membrane. However it soon became apparent that by increasing the air pressure by increments to 15 atmospheres and measuring the water remaining in the soil sample the capillary potential could be measured. As anticipated by Richards, this proved to be a significant advance.

Richards' work on soil physics was interrupted in 1943 by World War II when Ren joined the staff of the Jet Propulsion Laboratory at the California Institute of Technology. Here he worked on sea-borne rocket launchers. He worked with the diligence and insight that characterized all his research, and he was able to see this project through to completion and traveled to the Pacific Theater of Operations to observe the launchers in action. He eventually obtained several patents on his design. These he added to most of the early patents on tensiometers and pressure plates.

Upon his return to the Salinity Laboratory at the war's end, he completed his work on the pressure-membrane apparatus. After satisfying himself that the sausage casing which he initially used for the membrane was permeable to water but not to salts, he began to employ them widely. The design had one drawback. The membrane was subject to punctures in soil samples that contained small stones and eventually developed leaks. Eventually Zilla succeeded in making increasingly large clay plates, which could withstand pressure up to 15 000 cm (15 atmospheres), the presumed limit for plant growth. These plates were 30 cm in diameter, rendering them suitable for large samples or a large number of small samples. Replacement of the sausage by ceramics made the equipment usable by almost anyone. There rapidly developed such a demand for ceramic cups and plates that Richards set his sons up in a business known as the LARK Instrument Company. This stood for L.A. Richards' kids, who by participating in the manufacture of the ceramics helped pay their way through college.

In 1947 through 1949 Ren published the definitive papers on the measurement of soil moisture tension (capillary pressure). There has been little need for improvement in the equipment since that time. The demand for tensiometers and pressure plates continued to grow until Richards' assistant Percy Skaling set up his own business in order to meet this demand. Skaling was very careful to follow the design of the equipment exactly as described by Richards, except for very minor changes to facilitate manufacture. When colleagues would complain that their own equipment did not seem to work so well as Richards claimed, his response was always the same. He inquired if they had built it exactly as shown and had followed the experimental procedure exactly. They almost always confessed that they had attempted to make improvements in some way. He would tell them to go back into the laboratory and follow his design and instructions to the letter and then if it still did not work he would listen to them. They seldom came back, because he had already tried and discarded almost all variations in favor of the final design.

Ren learned much about soil water properties that is now taken for granted and is often rediscovered at regular intervals. More than any other individual L.A. Richards made soil physics an experimental science, in both field and laboratory. He made it possible to measure with considerable accuracy one of the most fundamental physical properties of soils, the relation between the soil water content and the capillary (matric) potential. He showed that this relationship was not single-valued but, rather, exhibited hysteresis. The relationship was dependent upon the previous history of wetting and drying and depended in particular upon the number of such cycles a soil sample had been through. To facilitate field measurements of the soil moisture gradient, he developed a multiple tensiometer to measure the matric potential at several depths simultaneously.

Richards always had his eye on the main problem. In a laboratory filled with the best equipment available, most of it of his own design, and with a thorough understanding of the limitations of this equipment, he set himself the task of quantifying most of the known soil physical properties. He often suggested that, at best, we could only measure some of the soil properties on some soils, some of the time. It is still true.

A good estimate of the upper limit of available soil water for plant growth was needed in order to characterize the salinity of a soil. For about 48 h after irrigation the soil continues to drain, albeit ever more slowly. He had already devised the content of 'leaching requirement' for a soil, i.e., the water content at which the drainage rate is equal to the irrigation rate, thus establishing a steady-state soil water content. When the leaching requirement is met, the accumulation rate of salts in the soil no longer continues to increase and the salt content remains constant over the long term. Many attempts were made to determine the water content at which the leaching rate would equal the application rate but, since the drainage rate was more nearly a function of the conductivity of the soil than the water content, these attempts were not very satisfactory. Richards and his colleagues carried out a series of field experiments, and finally suggested that the field capacity concept, i.e., the upper limit of available water, was at best an elusive or approximate concept and had done more harm than good. He concluded that field profiles should be characterized by a curve rather than a single point.

At the other end of the plant growth continuum, the wilting point, the soil water content had been related to the cessation of the elongation of sunflower leaves. Richards developed a freezing-point method to provide a much simpler means of estimating the wilting point of soils.

From his observations of the decreasing rate of outflow from the pressure plate apparatus, Richards concluded that it should be possible to estimate the unsaturated conductivity of the soil as a function of soil water content. A number of individuals attempted this and eventually W.R. Gardner, with Richards' help, succeeded in achieving this by means of a series of successive linear approximations.

Upon completion of the development of the tensiometer and pressure plate apparatus, Richards returned to the thermojunction or thermocouple method of measuring the vapor pressure of soil moisture. This method had the advantage over the tensiometer in that it used the vapor pressure of the soil moisture to measure the total (matric plus osmotic) potential well beyond the tensiometer range and also make measurements on plant or other wet material. This was especially desirable for work in arid regions where the tensiometer was of limited utility. It also had the advantage over the pressure plate apparatus in that it was not limited to soil samples but could be used *in situ* on plants as well as soil. Richards had become familiar with early attempts with this method at Logan, where M.D. Thomas had published work on it. While a student, Ren's brother-in-law Leon Linford had assisted Thomas with the measurements. Later at Iowa State, Ren had begun work on the method but with limited success. At Riverside he devised a constant-temperature bath with a precision approaching one-ten-thousandth of a degree in the hope of measuring the difference between a small droplet of water and temperature of the soil moisture with which it had been equilibrated to within a few hundredths of a degree. Eventually it was with the availability of high-gain electronic amplifiers that Richards was able to eliminate the need for such stable temperatures and also the need for a very precise galvanometer. He finally succeeded in achieving a goal that he had been pursuing for some 30 years. While his colleagues were quick to apply this technology to several problems, Ren was reluctant to use it until it could be applied to plant leaves *in situ* and continued the pursuit of greater sophistication.

While Richards was working on the fundamental problem of measuring the vapor pressure of soil water, he always had two or three lesser problems which he addressed while he thought about the number-one problem. He devised the saturation extract and the portable electrical resistant bridge, which gave a very good approximation of soil salinity. He invented a device for the measurement of the modulus of rupture for soil samples. This was useful in estimating the effect of sodium salts on the crusting of soils, a common problem with alkali soils.

Richards guarded his research time jealously, but at the same time possessed a strong sense of loyalty to the laboratory that afforded him virtually unlimited freedom. Therefore, he accepted the responsibility of editing *Diagnosis and Improvement of Saline and Alkali Soil* better known as Handbook 60. Released in 1954, this handbook became the 'bible' of saline and alkali soils and proved to be one of the most popular and useful US Department of Agriculture publications. It summarized the information and techniques developed during the first 15 years in the life of the Salinity Laboratory. It is a remarkable document that has stood the test of time, in that many of the concepts and equipment are still valid as a first approximation. A listing of the chapter headings gives a good idea of the scope of the Handbook: (1) Origin and nature of saline and alkali soils; (2) Determination of the properties of saline and alkali soils; (3) Improvement and management of arid and semiarid regions in relation to salinity and alkalinity; (4) Plant response and crop selection for saline and alkali soils; (5) Quality of irrigation water; (6) Methods of soil characterization; (7) Methods of plant culture and analysis; (8) Methods of analysis of irrigation waters.

Chapters 1, 2, 3, and 5, particularly, show Richards' influence. One particular example is the characterization of saline and alkali soils. The Russian school of soil science dealt with saline and alkali soils as though they always were and always would remain so. Handbook 60 assumed that soils could be 'reclaimed' if properly treated and managed. Although Richards was far from the first to espouse this idea, the Handbook gave this concept validity by showing how to achieve and sustain this goal The fact that it has not always been achieved is usually due to improper management, or some impediment to proper management, a *sine qua non.* To put it in Richards' own words, "the question for this laboratory to answer is whether or not irrigated agriculture can be made permanent."

Some 60 years later it would be difficult to improve upon the flow diagram that Richards and his collaborators devised for the diagnosis and reclamation of saline and alkali soils. This Handbook solidified Richards' position not only as the philosopher of the Laboratory but as the philosopher of soil physics. He understood what was important and how it differed from what was merely useful. While he worked from time to time on the latter, he never strayed for long from the former.

As significant as was the equation which bears his name, and the perfection of the tensiometer and invention of the pressure plate, they were eclipsed by Ren's collaboration with C.H. Wadleigh on the chapter in *Soil Water and Plant Growth.* For some time there had been an ongoing controversy between F.J. Veihmeyer at the University of California at Davis on the one side and many physicists and physiologists on the other. The debate centered on the issue of whether soil water was equally available to plants between the upper and lower limits of soil water. Veihmeyer said yes and almost everyone else said no. Wadleigh and Richards researched the American and English literature thoroughly and combined it with their own researches and concluded unequivocally that Veihmeyer was essentially wrong. The problem was in the measurement of soil water. In very sandy soils, especially in the case of fruit trees, Veihmeyer's approximation was not far wrong. In the general case, it was not good enough. Although Veihmeyer remained unconvinced, Richards' and Wadleigh's view has prevailed.

Richards has also had a profound effect upon the terminology of soil physics. In 1940 he published some comments on the permeability units for soils. In 1952 he chaired a committee of the Soil Science Society of America that recommended adoption of units of permeability and infiltration. In 1956, almost in passing, he suggested that the term 'soil suction' be abandoned in favor of the 'matric potential.' He eschewed the term 'capillary potential' because it implied that capillarity was the only physical mechanism influencing the water potential. Such was Richards' influence in the field of soil physics that the term was accepted almost by acclamation and with little public debate.

In 1951–1952 Richards was elected President of the Soil Science Society of America. In this position he succeeded in achieving a long-sought goal. He was able to persuade the Society, which heretofore had published only the papers presented at the Annual Meeting, to change from proceedings to a true journal which published peer-reviewed papers submitted by its members and meeting set standards. It was to be several years before it changed its name to the *Journal of the Soil Science Society of America*, but Richards had made this move inevitable. In 1965, Richards was elected President of the American Society of Agronomy. In this role he provided vigorous leadership to the society and organized and codified the business of the Society with a numerical scheme somewhat analogous to the system used by libraries. This system was sufficiently flexible and easily used so that it was adopted immediately by the Soil Science and the Crop Science Societies and continues in use long after only a few members know where it came from. Well beyond his official retirement Ren continued to work on an automated irrigation system and numerous related patentable devices.

*See also:* **Tensiometry**; **Water Content and Potential, Measurement**

## Further Reading

Blair GY, Richards LA, and Campbell RB (1950) Rate of elongation of sunflower plants and freezing point of soil moisture in relation to permanent wilt. *Soil Science* 70: 431–439.

Buckingham E (1907) *The Mechanics of Soil Moisture.* US Department of Agriculture Division of Soils Bulletin 10.

Gardner W (1920) The capillary potential and its relation to soil moisture constants. *Soil Science* 10: 357–359.

Gardner WR (1956) Calculation of capillary conductivity from pressure plate outflow data. *Soil Science Society of America Proceedings* 29: 317–320.

Gardner WR and Ehlig CF (1962) Impedance to water movement in soil and plant. *Science* 138: 532–533.

Gardner WR and Ehlig CF (1965) Physical aspects of internal water relations of plant leaves. *Plant Physiology* 40: 705–710.

Ivie JO and Richards LA (1957) A meter for recording slow liquid flow. *Review of Scientific Instruments* 8: 86–89.

Klute A and Richards LA (1962) Effect of temperature on relative vapor pressure in soil: apparatus and preliminary measurements. *Soil Science* 93: 391–396.

Neal OR, Richards LA, and Russell MB (1938) Observations on moisture conditions in lysimeters. *Soil Science Society of America Proceedings* 2: 35–44.

Richards LA (1931) Capillary conduction of liquids through porous medium. *Physics* 1: 315–333.

Richards LA (1938) Wet thermojunctions for measuring relative humidity. *Iowa Academy of Science Proceedings* 45: 175–178.

Richards LA (1940) Concerning permeability units for soils. *Soil Science Society of America Proceedings* 5: 49–53.

Richards LA (1941) A pressure membrane apparatus for soil solution. *Soil Science* 51: 377–386.

Richards LA (1942) Soil moisture tensiometer materials and construction. *Soil Science* 53: 241–248.

Richards LA (1948) Porous plate apparatus for measuring moisture rentionion and transmission by soil. *Soil Science* 66: 105–110.

Richards LA (1949) Methods of measuring soil moisture tension. *Soil Science* 68: 95–112.

Richards LA (1953) Modulus of rupture as an index of surface crusting of soils. *Soil Science Society of America Proceedings* 17: 321–323.

Richards LA (1954) Multiple tensiometer for determining the vertical component of the hydraulic gradient in soil. *Soil Science Society of America Proceedings* 18: 7–9.

Richards LA (1955) Water content changes following the wetting of bare soil in the field. *Soil Science Society of Florida Proceedings* 15: 142–148.

Richards LA (1955) Portable conductivity bridge and cells for salinity appraisal. *Soil Science* 80: 55–59.

Richards LA (1956) A soil salinity sensor of improved design. *Soil Science Society of America Proceedings* 30: 333–337.

Richards LA and Neal OR (1937) Some field observations with tensiometers. *Soil Science Society of America Proceedings* 1: 71–91.

Richards LA and Ogata G (1961) A thermocouple for vapor pressure measurements in biological materials and soils systems at high humidity. *Science* 128: 1089–1090.

Richards LA and Ogata G (1961) Psychrometric measurements of soil samples equilibrated on pressure membranes. *Soil Science Society of America Proceedings* 25: 456–459.

Richards LA and Wadleigh CH (1952) Soil water and plant growth. In: Shaw BT (ed.) *Monograph*, vol. 2 *American Society of Agronomy: Soil Physical Conditions and Plant Growth*, pp. 75–251. New York: Academic Press.

Richards LA, Russell MB, and Neal OR (1938) Further developments on apparatus for field moisture studies. *Soil Science Society of America Proceedings* 2: 55–64.

Richards LA, Neal OR, and Russell MB (1939) Observations on moisture conditions in lysimeters. II. *Soil Science Society of America Proceedings* 4: 55–59.

Richards LA, Committee on Terminology (1952) Report of subcommittee on permeability and infiltration. *Soil Science Society of America Proceedings* 16: 85–88.

Russell MB and Richards LA (1938) Heat of wetting of soils. *Iowa Academy of Science Proceedings* 45: 179–185.

United States Salinity Laboratory Staff (1954) In: Richards LA (ed.) *Diagnosis and Improvement of Saline and Alkali Soils.* USDA Handbook 60.

Veihmeyer FJ and Hendrickson AH (1933) *Some Plant and Soil-Moisture Relations.* American Soil Survey Association Bulletin 15.

# ROOT ARCHITECTURE AND GROWTH

**L E Jackson**, University of California–Davis, Davis, CA, USA

## Introduction

Root architecture and root growth are important for exploration of soil for water and nutrients, as the acquisition of these resources drives plant growth. There are a myriad of root forms, ranging from woody roots to root hairs, and many of their various functions still remain unknown. Several important concepts, however, govern the development of roots and their direct and indirect effects on plant growth: (1) roots are highly plastic, i.e., architecture and growth are highly responsive to environmental conditions and therefore can optimize resource acquisition; (2) tradeoffs exist in root form and function. The costs of producing roots can yield benefits in terms of the acquisition of soil water and nutrients, and ultimately plant growth and fitness; (3) enormous variation exists among plant species in their root responses to abiotic stress; (4) genetic analysis has revealed that root architectural traits are often multigene or quantitative traits; and (5) roots affect attributes of plant communities and ecosystems, due to their effects on resource acquisition, growth, productivity, water budgets, and nutrient cycling.

## Description of Roots and Root Architecture

Root systems are composed of the primary root that originates as part of the developing embryo in the seed, postembryonic, shoot-borne roots, and lateral roots that emerge from all root types. In the monocotyledonous plants such as grasses, seminal root axes (also called primary axes) grow out of the developing embryo, while nodal axes (also called secondary, adventitious, crown, or brace roots) subsequently arise from the shoot and gradually replace the seminal axes. In contrast, for dicotyledonous plants, the entire root system originates as a single root in the embryo, from which laterals arise, and which can later become a taproot, and/or thicken and become woody owing to cambial activity. Root hairs are extensions of epidermal cells and vary in length from 0.2 to 2 mm, depending on the species. They are continually being produced near the elongating root tip and are short-lived. Rhizomes, which occur in both types of plants, are underground, elongate stems that can be distinguished from roots by the presence of leaves, leaf scars, or buds.

Root anatomy changes along the growing root. The root tip consists of four overlapping zones (**Figure 1**): the root cap that protects the meristem, the meristem itself, a region of elongation, and a region of maturation, where root hairs emerge and vascular tissues form. A cross section of the mature root also has several distinctive zones. The epidermis is the outer layer of the root, with the protective exodermis immediately underneath it. These layers surround the cortex, which stores nutrients. The endodermis is the innermost layer of the cortex. The Casparian strip is a waxy substance that is present in endodermal cell walls. It forces water and nutrients to pass through the protoplasts of the endodermal cells, presumably to increase the selective uptake of nutrients and thus regulate ion absorption and accumulation. Within the vascular cylinder of the root, the xylem consists of

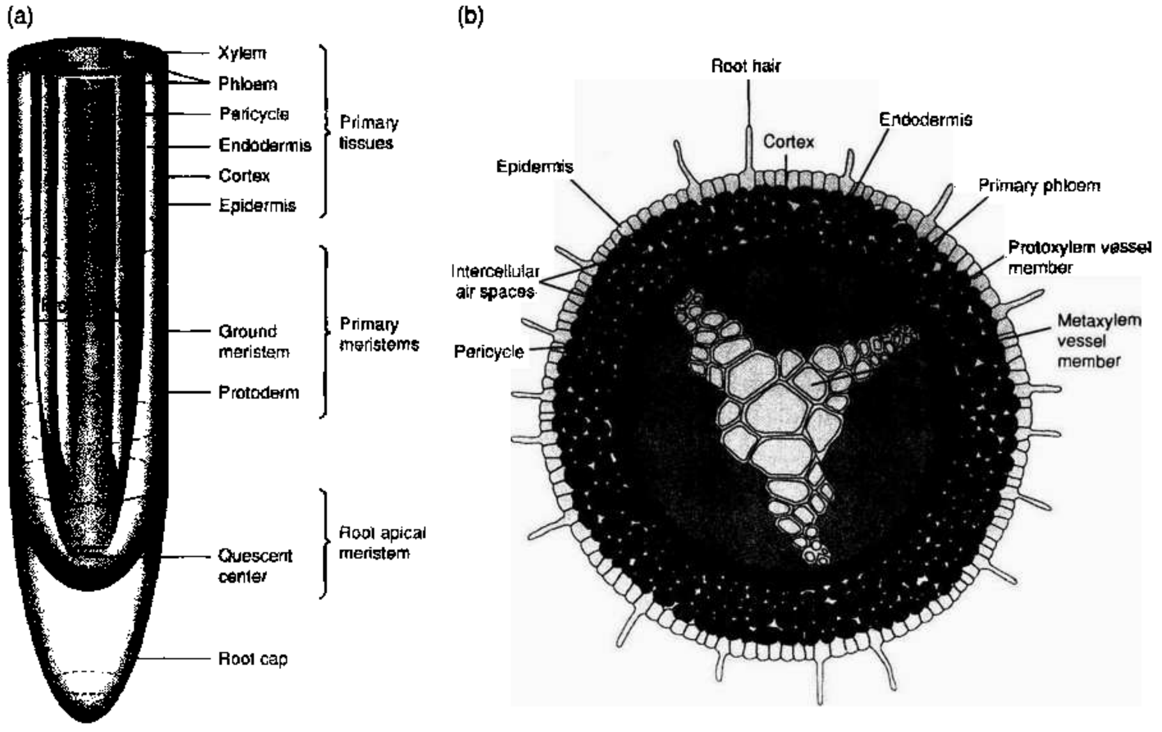


**Figure 1** Longitudinal view of a developing root (left) and cross section of a maturing root (right), showing the different tissues, meristems, and differentiation of xylem and phloem elements. The outermost layer is the epidermis and root hairs, and the protective exodermis. In the cortex, there are intercellular air spaces between the cells. The endodermis and pericycle perform specialized functions related to uptake and differentiation, respectively, and surround the vascular cylinder, which contains the xylem and phloem. (Reproduced with permission from Rost TL, Barbour MG, Stocking CR, and Murphy TM (1998) *Plant Biology*. Belmont, CA: Wadsworth Publishing.)

hollow tracheids, which are important for water transport, and phloem elements, which are food-conducting tissue. Within the xylem, protoxylem is the first formed element of the primary xylem. Metaxylem is the part of the primary xylem that differentiates after the protoxylem and before the secondary xylem if any of the latter is formed in a given taxon. Phloem differentiates in the areas between the columns of xylem cells. The pericyle is the outermost boundary of the vascular cylinder. Its long-lived cells are the sites where woody secondary growth originates, as does the bark of woody roots; also lateral roots are initiated here and then grow outward through the cortex and epidermis.

Formation of lateral roots, branching to form clusters of fine roots, differentiation of root hairs, mechanics of soil penetration, and changing root distribution with depth are among the root architectural characteristics that affect nutrient and water acquisition from the soil. In addition, other processes such as nutrient absorption kinetics, root exudates, and microbial activity in the rhizosphere affect nutrient availability and therefore plant uptake of nutrients from the soil. (*See* **Nitrogen in Soils: Plant Uptake**.)

Use of topological models for the description of root architecture has proven to be a useful way to quantify the branching structure of root systems (**Figure 2**). Topological models consider the number of links and their arrangement: interior links are internodes between two branching points, and external links originate at a branching point and end in a meristem. The number, length, and spatial distribution of internal and external links can be markedly different between even closely related species and can change with developmental age and environmental conditions. In wheat, root axes are produced synchronously with development of specific leaves, and, in other species, vascular connections direct the nutrient and carbon flow between particular shoots and roots. Thus, root topology can be related to shoot development and physiology. Combining topological analysis with measures of root length and allocation, in relation to shoot biomass, provides a robust way of describing the architecture, growth, and response to the environment of a root system (**Figure 3**).

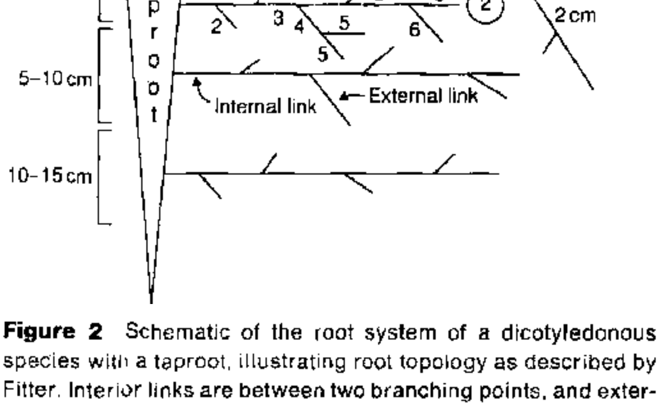


**Figure 2** Schematic of the root system of a dicotyledonous species with a taproot, illustrating root topology as described by Fitter. Interior links are between two branching points, and external links originate at a branching point and end in a meristem. The magnitude of the root system or lateral axis is the number of external links (7 for lateral 2), the altitude is the number of links in the longest unique path from the base link to an external link (6 for lateral 2). The numbering on lateral 2 shows all pathlengths. Measurements of length and biomass of the root axes enhance the descriptive information obtained from the topological indices, which theoretically are related to nutrient uptake and growth strategy.

Roots branch in response to endogenous and exogenous factors. The plant hormone auxin has a key role in the initiation of lateral root primordia. It can accumulate at the root apex, which usually exerts some degree of apical dominance, in the developing lateral root primordial, or both. Cytokinins, another type of plant hormone, inhibit lateral root formation. Low nutrient levels, especially of nitrogen, can act as a stress signal that decreases cytokinin content, and this is associated with higher root-biomass allocation. The ratio of auxin and cytokinin concentrations is thought to be important for the regulation of developmental steps for the complete formation of lateral roots. Ethylene, a plant hormone that is involved in root response to stress, can either enhance or repress root growth, depending on its concentration. Plant hormones clearly have important effects on root branching, but the complexity of their interactions is still a major unknown in plant physiology.

Lateral root formation and the relative growth rate of roots are stimulated by nitrate, ammonium, and phosphate, such that highly branched roots and high root-length density are found in enriched soil patches. For immobile ions, root proliferation can increase nutrient uptake from such patches. In contrast, a mobile ion such as nitrate can be readily captured by a root system with low root-length density. It has been hypothesized, however, that localized root proliferation in response to nitrate may be a mechanism by which competing plants can quickly usurp a limiting resource. Water uptake is also enhanced by increased root surface area, even though its transport may be limited by the hydraulic conductivity of the xylem elements in some circumstances.

## Tradeoffs in Root Form

Construction and maintenance costs of roots must be considered in the assessment of root architectural patterns. An optimization approach suggests that

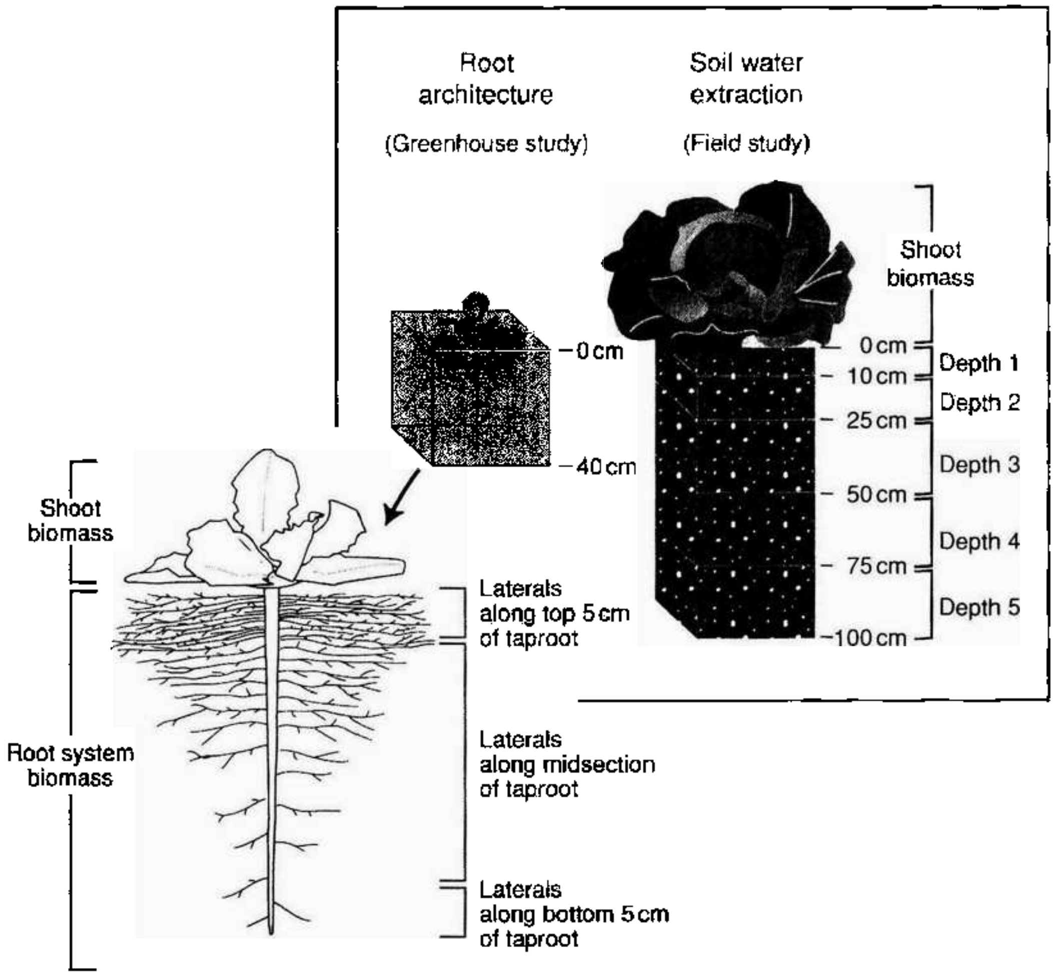


**Figure 3** A strategy for examining relationships between root structure and function. In a greenhouse study in pots, root architectural traits were measured, while in a field study, water extraction was measured at different depth increments. When closely related lines of lettuce were tested with both methods, genotypes with higher numbers of laterals near the tip of the taproot or with longer taproots tended to extract more deep water from soil (traits were scaled for biomass), and this resulted in identification of quantitative trait loci (QTL) that could be genetically mapped. (Reproduced with permission from Johnson WC, Jackson LE, Ochoa O, van Wijk R, Peleman J, St. Clair DA, and Michelmore RW (2000) Lettuce, a shallow-rooted crop, and *Lactuca serriola*, its wild progenitor, differ at QTL determining root architecture and deep water extraction. *Theoretical and Applied Genetics* 101: 1066–1073.)

total plant growth will be largest when the root system maximizes nutrient and water acquisition per unit resource supplied by the shoot. Allocation of large amounts of carbohydrates and nutrients to roots may detract from shoot growth, leaf area, and photosynthetic carbon gain, thereby reducing total plant growth, especially if the increased root allocation does not achieve increased rates of nutrient uptake. Under low nutrient availability, root diameter often decreases, and thinner roots are thought to increase the capacity for exploration of soil per unit carbon expenditure than would thicker roots. Roots produced in enriched soil patches can be short-lived, eliminating their maintenance costs when soil depletion causes nutrient acquisition rates to decrease. In contrast, carbon allocation to long-lived storage roots and woody roots, which have high construction costs, limits shoot growth and thus leaf area and the potential for photosynthetic carbon gain; yet long-term beneficial effects for nutrient and water uptake may outweigh these initially negative effects on total growth.

## Root Traits for Stress Tolerance

Plants cope with environmental stresses such as drought, soil compaction, salinity, nutrient limitation and imbalances, acid soil, heavy metal toxicity, and waterlogging. Each stress is mediated in part through the root system. Some plant species show highly specific adaptations, involving in many cases root architecture and root growth, to these abiotic stresses, along with certain physiological mechanisms.

Drought tolerance is partially mediated by root characteristics. Some plants have deep root systems that access water that would otherwise be lost to drainage. Another way for roots to exploit deep moisture is hydraulic lift, whereby deep moisture is transferred at night to upper soil layers with higher root-length density for extraction during the following day. Low root hydraulic conductance is another trait that enables drought avoidance by decreasing the rate at which water is transported to and used by the shoot in dry environments with deeply stored soil water, e.g., Mediterranean-type environments. By restricting water use during the early part of the growing season via low transport rates, more water is available during the latter period of the year when rainfall is absent or minimal. Delayed maturation of the metaxylem vessels in developing roots is one mechanism by which low root hydraulic conductance is achieved. Another example of root-mediated drought tolerance is found in the developmental plasticity of cactus roots, where roots in moist microsites have less lignification and suberization, and higher hydraulic conductivity than roots in dry microsites, thereby apparently maximizing the potential for water uptake from wet patches and minimizing water loss from roots in dry patches.

As soil dries, hydraulic conductivity between the root and soil decreases. Shrinkage of roots also decreases the contact between root and soil. Low transpiration rates ensue, further causing the movement of liquid water by mass flow in the soil to decline. Roots grow through and away from dry sites and toward wetter patches in soil. The root cap is probably the site where this hydrosensing occurs.

Roots in contact with dry soil can be the source of inhibitors of shoot growth and stomatal opening, which in turn limits water loss from the plant. Hydraulic signals that elicit physiological responses to drought move from the roots to the shoots in the transpiration stream, and the hormone abscisic acid (ABA) and root-derived cytokinins appear to play a role in this signaling process. ABA modulates the expression of many genes and physiological processes that protect cells from excessive water loss (*See* **Water-Use Efficiency**), and it also appears to be involved in increasing root growth and consequently root-to-shoot ratio, which would increase water uptake from deeper, moister layers of the soil profile and reduce dehydration when upper layers of the soil profile dry out. Experiments have compared plants with root systems split between well-watered and dry zones of soil, i.e., partial root-zone drying, and a well-watered control. With partial root-zone drying, there can be an increase in water-use efficiency (WUE), along with a decrease in stomatal conductance and total vegetative growth, but little effect on fruit production, presumably because the developing fruit creates a strong sink for carbohydrate partitioning. The mechanisms by which root signals are transported and operate on various plant organs still remains largely unknown, but, even so, partial root-drying may be a potential method to limit water use and still maintain high yields in the commercial production of fruit-bearing crops such as grapes.

In environments where drought is normally a temporary event, maintaining viable roots under drought stress may be advantageous, especially for perennials that benefit from the higher production and maintenance costs of long-lived roots over a period of years. In species from deserts and Mediterranean-type environments, however, where plants intermittently experience severe and long-term soil drought, especially near the soil surface, fine roots are often shed when soil dries, and new root production can occur rapidly after rewetting by rainfall. Many crop species also shed roots in dry surface soil.

Compaction, like soil drying, increases the mechanical impedance to root penetration of soil. In soils with high bulk density, roots become clumped into small pores, since larger pores are scarce, and this limits the soil volume for nutrient and water uptake. While growing through soil, roots must overcome radial, axial, and frictional forces, and there is typically a negative correlation between root elongation rate and penetrometer resistance. The osmotic potential of root cells increases in impeded root tips, possibly as a mechanism to increase cell turgor and soil penetrability. Root impedance by mechanical forces in soil also triggers root signals that reduce leaf expansion and transpiration rates. Reduced leaf-expansive growth and shoot water-loss induced by mechanical stress, even under conditions of high nutrient and water supply, may be responses that prepare the plant for the high probability that these resources may soon become scarce as a result of reduced root growth.

Salinity inhibits root-elongation growth and development in nonhalophytes, and this, in turn, indirectly limits the ability of roots to explore soil for nutrients and water, which decreases the growth rate of the plant. Direct effects of salinity on roots include the osmotic stress from low soil-water potential that can induce both the biosynthesis of costly organic osmolytes to regulate osmotic adjustment in the root and the accumulation of specific ions in the root tissue which can lead to induced toxicities and deficiencies. Salt tolerance is achieved by maintaining a ratio of low levels of $Na^+$ and high levels of $K^+$ in the cytoplasm. Elevated $Ca^{2+}$ concentrations in the soil medium can increase the root elongation rate in some plant taxa in saline soils, possibly because there is less displacement of membrane-associated $Ca^{2+}$ by $Na^+$ than in less salt-tolerant genotypes.

Nutrient limitations are well known to be mediated by continuous root growth into unexplored soil, as well as by high total root surface area, both of which increase the capacity of the root system to take up nutrients. Nutrient bioavailability is mediated by root interception, mass flow, and diffusion (**Figure 4**). As roots grow, they displace soil and intercept available nutrients, for example, adsorbed on to soil particles. This process of root interception, however, is thought to provide only a small proportion of the plant's total nutrient uptake. Mass flow is the transfer of water and dissolved nutrients to the root surface and is driven by transpiration. Rates of diffusion, i.e., transport in soil along the concentration gradient, depend on the mobility and effective diffusion coefficients of specific ions, soil replenishment, and plant uptake. When transpiration rates are high, large quantities of water and dissolved solutes travel to the root via mass flow, but, when soils dry, diffusion of nutrients becomes the main way by which roots access nutrients.

Both old and newer regions of the root are important for nutrient absorption. Even though older parts of the root system are more suberized and may be surrounded by depletion zones, they have been shown to be capable of high potentials for nutrient uptake in laboratory experiments. Since they constitute a large proportion of the root surface area of the plant, they are probably important for a considerable proportion of total nutrient uptake. The highest rates of nutrient uptake per unit root length, however, are found in the region just behind the root apex, where root hairs are most abundant.

Nutrient-poor soils elicit several changes in root morphology, as well as physiology, such as increased capacity for nutrient uptake (*See* **Nitrogen in Soils: Plant Uptake**) but the degree of these responses can be very different among plant taxa. Higher root-to-shoot ratios and lower total root biomass typically occur with nutrient deprivation.

Phosphate (P), an immobile nutrient, does not move readily to the root by either mass flow or diffusion. P-deficient soils tend to cause large changes in a number of root characteristics. Under low-phosphate conditions, roots become thinner and the elongation of laterals increases, especially those at higher orders, i.e., near the meristematic tips of each lateral root axis, and leaf expansion and surface area decline. Also, more root hairs are produced and they have faster elongation rates. Root hairs have their greatest effect on absorption of ions that diffuse slowly in soil. Expansion of root surface area via increased production of root hairs is much less costly than a greater investment in roots.

Acid soils (*See* **Acid Rain and Soil Acidification; Acidity**) decrease root growth, partly due to direct effects of low pH, and also due to the higher availability and toxicity of heavy metals, such as aluminum. Aluminum toxicity decreases both root elongation and cell division, and root cells become shorter and wider, giving roots a stubby appearance. The root tips are the primary sites of aluminum toxicity. In many tropical soils that experience low pH and aluminum toxicity, P is either deficient or is readily fixed to soil particles. Thus, the low root-length density of plants in response to aluminum toxicity is an important limitation to P uptake and total plant growth. Other heavy metals, such as cadmium, copper, and zinc, also decrease root elongation, and root dry mass to a lesser extent, but the mechanisms for these responses are not clearly understood.

In waterlogged or flooded soils, air is displaced from soil pore spaces, and various degrees of oxygen

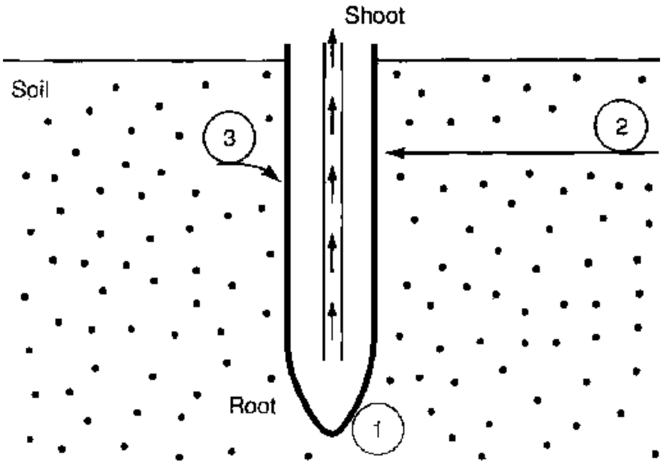


**Figure 4** Schematic of the movement of minerals in soil to the root surface: 1. With root interception, the soil volume is displaced by the root volume; 2. With mass flow, the bulk soil solution is transported along the water potential gradient, which is driven by transpiration; 3. With diffusion, nutrient ions are transported along the concentration gradient. (Reproduced with permission from Marschner H (1995) *Mineral Nutrition of Higher Plants*. San Diego, CA: Academic Press.)

depletion typically occur in heterogeneous microsites in the soil. Flooding of a previously well aerated plant can cause immediate cessation of root growth, root respiration, and root death, partly due to the production of toxic substances such as nitrite or acidification of the root cells due to lactic acid produced during anaerobic fermentation. In some plant species after waterlogging, adventitious roots with well-developed aerenchyma, large air spaces in the root cortex, can emerge from the base of the stem, or existing roots can develop aerenchyma (Figure 5). The formation of aerenchyma under hypoxic conditions is associated with the production and accumulation of the plant hormone ethylene. Aerenchyma increases the root porosity and the long-distance transport of oxygen to apical zones of the root system. Pressurized gas transport, driven by temperature gradients between aboveground aerated parts of the plant and submerged root systems, also exists in some flooding-tolerant species.

Mycorrhizal symbioses help overcome some types of stress. Mycorrhizal colonization increases the surface area of the root system by extending hyphae beyond the root surface and consequently increases nutrient uptake, especially of ions with low mobility in soil, e.g., phosphorus, zinc, and copper. Mycorrhizae increase the distance of nutrient exploration away from the root and are important in sites with patchy distribution of nutrients. Water uptake is thought to be increased for some plant species. Arbuscular mycorrhizae (AM) are endophytes with intracellular colonization by Zygomycete fungi, whereas ectomycorrhizae are either Ascomycete or Basidiomycete fungi that form a mantle or sheath enclosing the rootlet. Dependence on AM tends to be highest in plants with coarse, fibrous root systems with few root hairs. Many herbaceous species with more fibrous root systems and high specific root length, i.e., thinner roots, are facultative mycotrophs that become less mycorrhizal with increased phosphorus

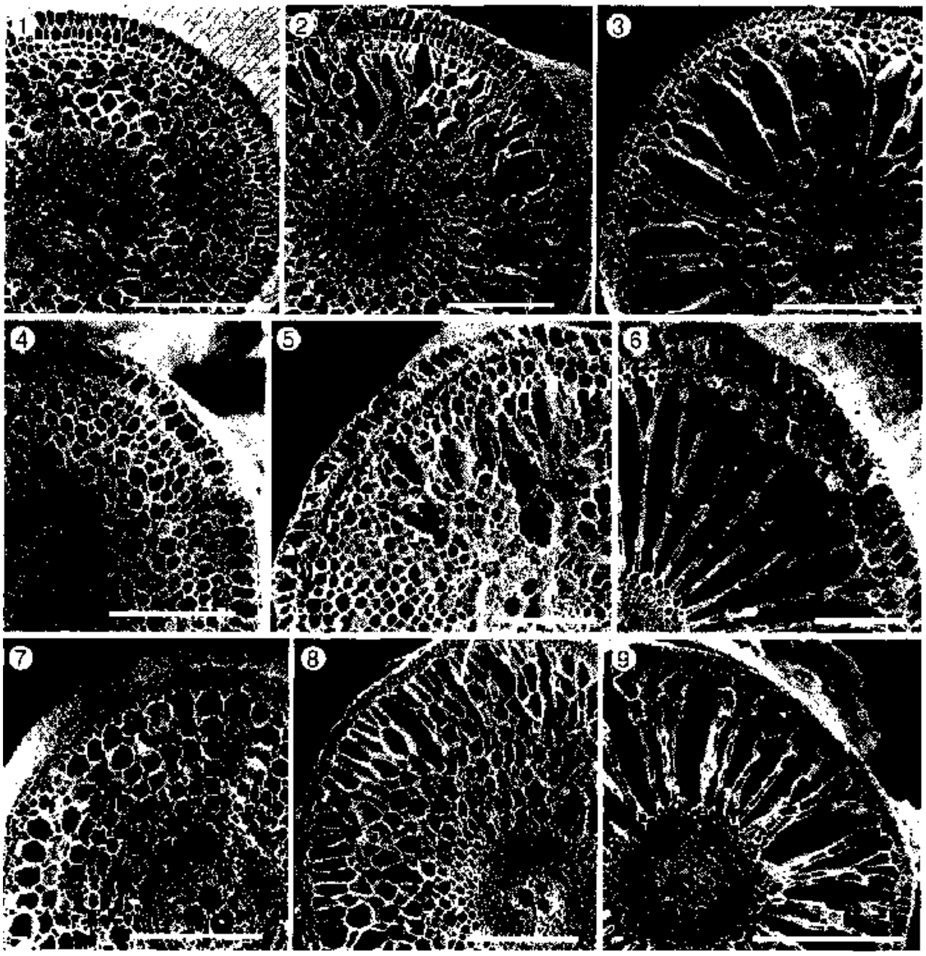


**Figure 5** Scanning electron micrographs of transverse sections of roots of wetland plant species at differing distances from the root tip. Micrographs 1–3 show *Sagittaria lancifolia* at 5, 23, and 45 mm from the root tip, respectively. Micrographs 4–6 show *Thalia geniculata* roots at 2, 8, and 32 mm, respectively. Micrographs 7–9 show *Pontederia cordata* 6, 31, and 58 mm, respectively. The development of cortical air spaces (aerenchyma) in roots is an adaptive response to flooding. (Reproduced with permission from Longstreth DJ and Borkhsenious ON (2000) Root cell ultrastructure in developing aerenchyma tissue of three wetland species. *Annals of Botany* 86: 641–646.)

availability. Several studies indicate that root branching and lateral root formation increase with AM colonization. In some plant species, a decrease in specific root length (centimeters per gram of dry weight root) occurs in AM versus noncolonized roots, i.e., mycorrhizal roots are thicker. For ectomycorrhizal roots, the bulky tissue of the mycorrhizal mantle surrounding the roots serves a role in nutrient storage and in root-fungus contact, but does not greatly increase the surface area for nutrient uptake. Instead the extraradical mycelium that extends from the mantle and their rhizomorphs, conglomerations of hyphae with distinct structural shapes, perform this function. Ectomycorrhizal colonization can increase the number of root tips on laterals. Both AM and ectomycorrhizal colonization frequently result in allocation of a lower proportion of total plant weight to root dry weight.

## Genetic Analysis of Root Traits

As with many aspects of plant growth and development, root architecture and morphological traits are usually under the control of a complex gene system in which several genes, rather than a single gene, operate together to create a given phenotype. Multigene control of root traits and root responses make the study of root genetics difficult. Also, roots are hard to study due to the large numbers of plants needed for genetic studies, especially in soil in field experiments, where their phenotypic evaluation is most valid.

For rice, a large effort has been invested in understanding root responses to drought. Genetic variation among rice cultivars exists for root thickness, total root length, and root penetration ability, i.e., for proliferation in a soil hardpan. Using quantitative trait locus (QTL) analysis that relates phenotypic variation to specific genetic regions in a genetically segregating population, these traits have been genetically mapped, and several chromosomal regions have been found to affect each of the traits. Like root traits, drought avoidance traits in rice, e.g., relative water content and the extent of leaf rolling that limits transpirational surface area, also give many QTLs that differ between sites and between years at the same site. Yet there is little co-location on the genetic map of drought avoidance and root morphological QTLs, implying that root traits do not seem to be contributing consistently to drought avoidance in rice. Whether this is true or an artifact due to screening techniques remains to be seen. A more straightforward outcome has been achieved for lettuce. In lines derived from a cross between wild and cultivated lettuce, QTLs for deep water extraction and deep rooting traits co-localize (Figure 3), and genetic markers for these QTL regions are now being introgressed into cultivated lettuce, with the hope of creating a deeper-rooted crop that requires less irrigation input.

Single-gene traits for root morphological traits have been identified in some plants, but are best studied in the tiny crucifer *Arabidopsis thaliana*, a current model species used by plant geneticists. Mutants for traits such as failure to initiate lateral roots, auxin-related lateral root growth, and variable lateral root initiation in response to the soil environment exist in this species. Transgenic *Arabidopsis* plants have also been used to demonstrate the function of a specific gene controlling the proliferation of lateral roots in nitrate-rich patches. In normal, nontransgenic plants, the expression of this gene is induced rapidly when nitrate is supplied, but, in transgenic plants in which the gene is repressed, roots no longer proliferate in response to nitrate. The gene may therefore be a key factor in determining the developmental plasticity of roots in response to an environmental cue.

The strong effect of the abiotic and biotic environment in the soil on root development and branching patterns is obviously important for understanding the inheritance of root traits and their adaptation to specific habitats, and for crop improvement for more sustainable use of water and nutrients. For many crops, extensive genotypic variation exists for root-system traits among cultivars, but the exact genetic basis of these traits is difficult to ascertain and therefore hard to implement in crop-breeding programs. Molecular techniques including QTL analysis and marker-assisted selection will help to overcome these problems.

## Root Architectural Effects on Plant Communities and Ecosystem Processes

Plant species and ecotypes are adapted to the environments in which they grow. Plant communities of nutrient-poor sites are typically composed of plants with slow growth rates, high nutrient concentrations in tissue, and high root-to-shoot ratios. Even with a high nutrient supply, they have low rates of nutrient uptake and growth. In contrast, natural vegetation on nutrient-rich sites, e.g., many disturbed sites, often consists of species with the capacity for high rates of nutrient uptake and growth, rapid response to increased availability of soil nutrients, and low root-to-shoot ratio. These traits ultimately contribute to higher fitness even though the specific respiratory costs for nutrient uptake may be high. Despite these generalities, different ecophysiological strategies can usually be found within the same plant community. For example, in the vegetation of the arid Great Basin in the USA, sagebrush and one prevalent perennial grass showed no root response to enriched fertile microsites, while another perennial grass species exploited P via rapid root proliferation in the patches.

Spatial and temporal patterns of changes in root growth, architecture, and distribution affect nutrient and water availability in soil, as well as retention and loss of these resources to the atmosphere, runoff, and leaching. These relationships are best understood for crop systems. Like most annuals, crop plants proliferate roots in the top 0.5 m of soil until subject to drought, at which time roots grow increasingly deeper and rely on deeply stored water. Many annual species reduce or terminate root growth at anthesis, and reproductive growth and seed production often rely heavily on remobilization of nutrients and carbon from roots and other vegetative parts. To examine these effects on water and nutrient budgets, models increasingly rely on information on root growth, activity, and phenology, rather than using solely a 'root-sink' term to explain unaccounted-for fluxes in the soil. These include accounting for matric and osmotic potential effects on water and passive nutrient uptake, root-age effects on root water and nutrient uptake activity, as well as the influence of nutrient deficiency and ion toxicity on root growth.

Root distribution often reflects changes in soil strength and drainage channels in the soil profile. For *Eucalyptus* trees, fine roots concentrate in the surface soil and in zones of preferential infiltration, causing rapid drying and nutrient depletion in these areas. During the wet season, some roots grow into compacted zones, utilizing seasonally available water when moisture reduces soil strength in these areas. Thus, both spatial and temporal aspects of root growth help to reduce nutrient loss caused by deep drainage.

Competition between plants in the same plant community affects root distribution, water uptake, and WUE. The water source used by the roots of different species can be traced with deuterium ($^2$H) and oxygen-18 ($^{18}O_2$), which are stable isotopes that vary among rainwater, soil water, and groundwater. Carbon isotope composition ($\delta^{13}$C) is related to plant intrinsic WUE, which is the ratio between $CO_2$ assimilation and leaf water vapor conductance. Studies that use isotopic techniques along with measurements of root biomass and leaf gas exchange combine an assessment of root distribution, and short- and long-term physiological activity. In tree–grass mixtures, for example, grasses exert a negative effect on tree root growth, but usually not vice versa. Drought, however, typically has a more negative effect on grass than on trees, which have more conservative leaf gas exchange characteristics and deeper roots, which utilize deeply stored water. Sorting out these complex, competitive relationships has been possible with a combination of techniques, including stable isotopes. An example is tracking the natural abundance of the stable isotope $^2$H in soil water, which varies with depth in dry environments, and in plant xylem water, which gives a weighted average of all soil water acquired by functional roots (Figure 6).

Successful plant invasion is due in part to root traits, along with a suite of related ecophysiological characteristics. For salt cedar (*Tamarix* sp.), a nonnative phreatophytic species in riparian areas of

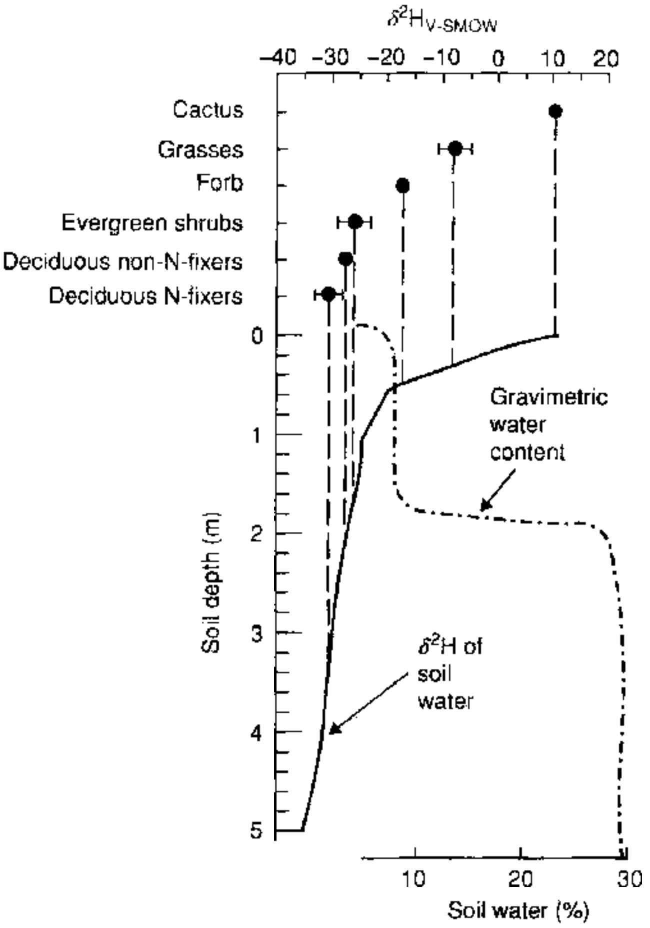


**Figure 6** Hydrogen isotope content, $\delta^2$H, in soil water and plant xylem water in a subtropical savanna in Texas, USA. *Prosopis glandulosa*, honey mesquite, is the dominant woody plant. In the soil, $\delta^2$H decreases exponentially with depth, and gravimetric moisture is higher below approximately 2 m depth. Shallow-rooted plant functional groups (grasses, forbs, and cacti) obtain their water from the upper 1 m of the profile, based on their $\delta^2$H. More negative $\delta^2$H values of plant and soil water indicate that tree and shrub roots access water in the zone of high gravimetric content, while roots of grasses, cacti, and forbs do not. Hydrogen isotope composition, $\delta^2$H or $^2$H/$^1$H, is expressed relative to a conventional standard, V-SMOW (Vienna Standard Mean Ocean Water). (Reproduced with permission from Boutton TW, Archer SR, and Midwood AJ (1999) Stable isotopes in ecosystem science: structure, function and dynamics of a subtropical savanna. *Rapid Communications in Mass Spectroscopy* 12: 1263–1277.)

**Table 1** Global land area ($10^6$ km$^2$), total root biomass ($10^9$ Mg), total fine-root biomass ($10^9$ Mg), live fine-root biomass ($10^9$ Mg), live fine-root length, and live fine-root surface area for different biomes, as determined by calculations and conversion factors

| *Biome* | *Land area ($10^6$ km$^2$)* | *Total root biomass ($10^9$ Mg)* | *Total fine root biomass ($10^9$ Mg)* | *Live fine root biomass ($10^9$ Mg)* | *Global fine root length ($10^{15}$ km, live)* | *Global fine root surface area ($10^7$ km$^2$, live)* |
|---|---|---|---|---|---|---|
| Tropical rainforest | 17.0 | 83 | 9.7 | 5.7 | 0.069 | 12.6 |
| Tropical seasonal forest | 7.5 | 31 | 4.3 | 2.1 | 0.026 | 4.7 |
| Temperate evergreen forest | 5.0 | 22 | 4.1 | 2.5 | 0.031 | 5.5 |
| Temperate deciduous forest | 7.0 | 29 | 5.6 | 3.1 | 0.038 | 6.9 |
| Boreal forest | 12.0 | 35 | 7.2 | 2.8 | 0.031 | 5.5 |
| Woodland and shrubland | 8.5 | 41 | 4.4 | 2.4 | 0.099 | 9.9 |
| Savanna | 15.0 | 21 | 14.9 | 7.7 | 0.91 | 63.6 |
| Temperate grassland | 9.0 | 14 | 13.6 | 8.5 | 1.01 | 71.2 |
| Tundra/alpine | 8.0 | 10 | 7.7 | 2.7 | 0.059 | 4.2 |
| Desert | 18.0 | 6.6 | 4.9 | 2.3 | 0.072 | 9.9 |
| Cultivated | 14.0 | 2.1 | 2.1 | 1.1 | 0.13 | 5.0 |
| Totals | 121.0 | 294.7 | 78.5 | 40.9 | 2.475 | 199.2 |

Reproduced with permission from Jackson RB, Mooney HA, and Schulze E-D (1997) A global budget for fine root biomass, surface area, and nutrient contents. *Proceedings of the National Academy of Sciences of the USA* 94: 7362–7366.

North American deserts, greater root allocation and deeper roots allow higher transpiration and photosynthesis rates under dry, hot conditions than can be achieved by native riparian species. These root traits, along with prolific seed production and vegetative reproduction, help explain the widespread invasive success of salt cedar in this environment.

At a regional scale, root distribution can affect hydrology. An example is the replacement of deep-rooted perennial trees and shrubs, up to 20 m deep, by shallow-rooted pasture and crop plants in the seasonally dry Murray–Darling Basin in Australia. Historically, the native vegetation would have relied on deep soil recharge and groundwater, while the shallow-rooted agroecosystems today occupy only the top meter of soil and utilize far less water than the forest and woodland communities, whose evergreen canopies had higher evapotranspiration rates. The current outcome of the vegetation transition is groundwater recharge and rising water tables, and since the groundwater is now usually saline, secondary salinization.

Improved understanding of the factors controlling root construction, biomass allocation, and distribution will improve ecological, hydrological, and global climate models. Roots are central to many ecosystem processes, including water absorption, which affects groundwater and atmospheric water fluxes, and indirectly influences water availability for soil microbial processes that control carbon dioxide, $NO_x$, and methane fluxes, as well as nutrient concentrations and fates, which also play a role in these processes. Fine-root biomass and surface area, and distribution through the profile are particularly important for these processes due to their relatively short lifespan, low C:N ratio, which enhances turnover rates, and their impact on nutrient absorption and thus primary production. A recent analysis indicates that substantial differences exist between biomes in total fine root biomass and depth distribution (Table 1), and that approximately 33% of global net primary production is used for fine-root production. These findings indicate the need for more physiological and ecological information on the importance of roots in terms of total plant productivity, carbon turnover, and long-term storage of carbon in perennial roots in different ecosystems.

*See also:* **Nitrogen in Soils:** Plant Uptake; **Water-Use Efficiency**

## Further Reading

Araus JL, Slafer GA, Reynolds MP, and Royo C (2002) Plant breeding and drought in $C_3$ cereals: what should we breed for? *Annals of Botany* 89: 925–940.

Boutton TW, Archer SR, and Midwood AJ (1999) Stable isotopes in ecosystem science: structure, function and dynamics of a subtropical savanna. *Rapid Communications in Mass Spectrometry* 12: 1263–1277.

Drew MC, Jackson MB, and Giffard S (1979) Ethylene-promoted adventitious rooting and development of cortical air spaces (aerenchyma) in roots may be adaptive responses to flooding in *Zea mays* L. *Planta* 147: 83–89.

Eissenstat DM (1997) Tradeoffs in root form and function. In: Jackson LE (ed.) *Ecology in Agriculture*, pp. 173–199. San Diego, CA: Academic Press.

Esau K (1977) *Anatomy of Seed Plants*, 2nd edn. New York: John Wiley.

Fitter AH (1987) Functional significance of root morphology and root system architecture. In: Fitter AH (ed.) *Ecological Interactions in Soil*, pp. 87–106. Oxford, UK: Blackwell Scientific Publications.

Jackson RB, Mooney HA, and Schulze E-D (1997) A global budget for fine root biomass, surface area, and nutrient contents. *Proceedings of the National Academy of Sciences of the USA* 94: 7362–7366.

Johnson WC, Jackson LE, Ochoa O, van Wijk R, Peleman J, St. Clair DA, and Michelmore RW (2000) Lettuce, a shallow rooted crop, and *Lactuca serriola*, its wild progenitor, differ at QTL determining root architecture and deep water extraction. *Theoretical and Applied Genetics* 101: 1066–1073.

Laclau J-P, Arnaud M, Bouillet J-P, and Ranger J (2001) Spatial distribution of *Eucalyptus* roots in a deep sandy soil in the Congo: relationships with the ability of the stand to take up water and nutrients. *Tree Physiology* 21: 129–136.

Lambers H, Chapin FS III, and Pons TL (1998) *Plant Physiological Ecology*. New York, NY: Springer-Verlag.

Marschner H (1995) *Mineral Nutrition of Higher Plants*. San Diego, CA: Academic Press.

O'Toole JC and Bland WL (1987) Genotypic variation in crop plant root systems. *Advances in Agronomy* 41: 91–145.

Price AH, Cairns JE, Horton P, Jones HG, and Griffiths H (2002) Linking drought-resistance mechanisms to drought avoidance in upland rice using a QTL approach: progress and new opportunities to integrate stomatal and mesophyll responses. *Journal of Experimental Botany* 53: 989–1004.

Rost TL, Barbour MG, Stocking CR, and Murphy TM (1998) *Plant Biology*. Belmont, CA: Wadsworth Publishing.

Smith SE and Read DJ (1997) *Mycorrhizal Symbiosis*. San Diego, CA: Academic Press.

Waisel Y, Eshel A, and Kafkafi U (2002) *Plant Roots: The Hidden Half*. New York: Marcel Dekker.

# ROOT EXUDATES AND MICROORGANISMS

**B-J Koo and D C Adriano**, University of Georgia, Aiken, SC, USA
**N S Bolan**, Massey University, Palmerston North, New Zealand
**C D Barton**, University of Kentucky, Lexington, KY, USA

## Introduction

Plant roots influence the physical, chemical, and biological conditions of the soil in the rhizosphere. The biogeochemical reactions induced by microorganisms at the soil–root interface (i.e., rhizosphere) play an important role in the bioavailability of nutrients and metals to plants. This microenvironment is characterized by distinct physical, chemical, and biological conditions compared with the bulk soil, largely created by the plant roots and its microbial associations. Such associations can include nonsymbiotic and symbiotic organisms such as bacteria and mycorrhizal fungi. The microbial populations are an essential part of the rhizosphere and affect the rhizosphere soil by their various activities such as water and nutrient uptake, exudation, and biological transformations.

Organic acids, sugars, amino acids, lipids, coumarins, flavonoids, proteins, enzymes, aliphatics, and aromatics are examples of the primary substances found within this microzone. Among them, the organic acids have received considerable attention owing to their role in providing substrates for microbial metabolism and for serving as intermediates for biogeochemical reactions in soil.

Nutrients and metals are typically present in the soil solution at low concentrations and tend to form sparingly soluble minerals (except nitrogen, sulfur, and boron), or may be adsorbed to a solid phase through ion exchange, hydrogen bonding, or complexation. The extent to which they are transferred from the soil to the biota (i.e., microbes or plants) is dependent on the biogeochemical interactions and/or processes among the soil, plant roots, and microorganisms in the rhizosphere. At this interface, the presence of root exudates may influence chemical reaction kinetics within the soil environment and subsequently affect biological activities. As such, understanding the role of the rhizosphere on biogeochemical processes within the soil is essential for developing bioremediation technologies of inorganic and organic contaminants.

## Rhizosphere Microzone

The rhizosphere, first described in 1904 by Lorentz Hiltner, has been the focus of intensive research for many years because of its importance in plant nutrition and pathogenesis. More recently, research on the rhizosphere has been directed toward its influence on controlling the persistence, mobility, and bioavailability of contaminants in soils.

The rhizosphere generally refers to the portion of soil found adjacent to the roots of living plants. The rhizosphere is subject to the influence of chemicals excreted by roots of living plants and the microbial community in this microzone. Its domain varies for different plant species and for age and morphology of roots. Depending on plant species, the width of the rhizosphere zone has been shown to extend from 2 to 80 mm away from the root surface. The soil in the rhizosphere supports a typically diverse and densely populated microbial community and is subject to chemical transformations caused by the presence of root exudates and metabolites of microbial degradation.

The rhizosphere microzone is distinguished from the bulk soil zone, more commonly known as the edaphosphere, by enhanced microbial activity and increased concentration of root exudates. Nevertheless, it is rather difficult to separate this zone physically from the root surface or rhizoplane. The rhizosphere effect is expressed quantitatively as the ratio of the number or activity of microorganisms or level of root exudates in rhizosphere soil ($R$) to that in the edaphosphere soil ($E$), i.e., the $R/E$ ratio. The $R/E$ ratio for microorganisms and root exudates is often found to range, respectively, from 2 to 20 and from 5 to 100, indicating enhanced microbial activity in the rhizosphere.

Root exudates are known as one of the most important factors affecting microbial growth in the rhizosphere. For example, cell numbers are several orders higher in the root zone than in the background soil lacking plants. The microbial community is more diverse, active, and synergistic than in nonrhizosphere soil. Root exudates can also selectively influence the growth of bacteria and fungi that colonize the rhizosphere by serving as selective growth substrates for soil microorganisms. For example, *Pseudomonad* species are particularly stimulated by enhanced carboxylic acid production from root exudation. Rhizosphere microbial communities can also vary in structure and species composition in different root locations or in response to other environmental factors, including soil type, plant species, or nutritional status.

## Microorganisms and Structure of Microbial Communities in the Rhizosphere

Rhizosphere communities are comprised of many cultivable and noncultivable microorganisms. Mycorrhizal fungi that live in symbiosis with higher plants are essential living components of the rhizosphere. Bacteria are also prominent components and may be active participants in the community or simply adventitious members. Therefore, the rhizosphere is the ideal system to study when seeking an understanding of the critical functional interdependence and metabolic capabilities of complex microbial communities containing both prokaryotic and eukaryotic members.

The rhizosphere of land plants hosts a large, diverse community of microorganisms that interact with each other and with the plant's roots. The mycorrhizal fungi are very important, ubiquitous groups within this community. The vegetative mycelium of mycorrhizal fungi and the root tips of more than 90% of land plants, including angiosperms, gymnosperms, and pteridophytes, form a mutually supportive symbiosis. This novel composite organ, the mycorrhiza, is the site of nutrient and carbon (C) transfer between the symbionts. The various mycorrhizal associations allow plants to colonize and grow efficiently in suboptimal environments and, in most terrestrial ecosystems mycorrhizae, not roots, are the chief organs of nutrient uptake by land plants. The arbuscular endomycorrhizae (AM) and the ectomycorrhizae (EM) are the most important mycorrhizal types.

The AM symbiosis is probably the most ancient of all such root-microbe symbioses. Approximately 80% of the higher green plants, including numerous crop species, are symbiotic with AM glomalean fungi, a relatively small group of filamentous fungi with aseptate hyphae and multiple nuclei within distinct cell wall boundaries. In soil, this small monophyletic group of aseptate filamentous fungi forms a very fragile mycelium that can produce large asexual spores filled with numerous nuclei. The extraradical hyphae of most AM fungi improve the uptake of immobile resources, notably inorganic phosphate, since they can acquire these beyond the depletion zone surrounding a root. They can also interfere with pathogens, increasing micronutrient uptake and altering drought-resistance. In some circumstances, AM fungi benefit plants by increasing the uptake of the more mobile inorganic- or even organic-nitrogen (N) compounds.

Only approximately 3% of perennial woody plants develop ectomycorrhizas. Ectomycorrhizal trees, *Betulaceae*, *Fagaceae*, *Salicaceae*, and *Pinaceae*, dominate forest ecosystems in boreal, temperate, and Mediterranean regions. The mycobionts span all of the phyla of true fungi, the *Zygomycota*, *Ascomycota*, and *Basidiomycota*, with at least 15 families within the *Basidiomycota* and *Ascomycota*. The mycelia of EM fungi are usually much more structurally differentiated than those of AM fungi. Many ectomycorrhizal fungi seem to be long-lived and form large genets in stable, established forest ecosystems. The huge diversity of fungal partners optimizes the uptake and mobilization of mineral resources distributed in discrete substrates in soil, litter, and within other soil organisms. EM fungi have better access than plant roots or AM fungi to organic phosphorus (P) and N compounds and seem well adapted to colonize the heterogeneous organic soil and litter layers of forests.

Both the AM and EM associations form an extraradical- and an intraradical-hyphal network that are functionally different entities. The extraradical hyphae colonize and exploit the soil volume to scavenge efficiently essential micro- and macronutrients. The intraradical hyphal network provides essential nutrient resources (e.g., phosphate and amino acids) to the plant. These nutrient contributions are reciprocated by the host's provision in its roots of a stable carbohydrate-rich niche for the fungal partner, making the relationship one of mutual symbiosis.

In the rhizosphere of plants, mycorrhizal fungi are always accompanied by another important group of microorganisms, the bacteria that also prosper in the organic-rich environment (mostly sugars, amino acids, and organic acids) released from the roots and mycorrhizal fungi. Among them, there is a large category of growth-promoting rhizobacteria that influence plant growth directly or indirectly by releasing a variety of compounds, from mineral nutrients to phytohormones and antimicrobial compounds. The establishment of mycorrhizae on roots is affected by the microbial populations of the rhizosphere, and especially by some bacteria, which can have either a positive or a negative effect on mycorrhiza formation. Plant-associated rhizosphere bacteria also have an important role in establishing and improving plant growth on different soil types, since they affect the availability of essential elements and provide plants with extra N.

## Composition of Root Exudates

Root exudates refer to a suite of substances in the rhizosphere that are secreted by the roots of living plants and microbially modified products of these substances. They consist of low-molecular-weight organic compounds that are freely and passively released root-cell material and mucilage associated with roots. Root exudates are a complex mixture of soluble organic substances, which may contain sugars, amino acids, organic acids, enzymes, and other substances (Table 1). The rhizosphere also contains cell lysates, which are released when root cortex cells are ruptured by mechanical damage, microbial grazing, fungal infection, or the emergence of lateral roots that break through the root epidermis. In addition to these substances, rhizodeposition includes the secretion of enzymes, such as acid phosphatase secreted in response to P deficiency.

Although the identified components differ in molecular weight, amount released, and biochemical role, the dominant exudate components have consistently been shown to be organic acids, sugars, and amino acids. Typically, the total organic acid concentration in roots is approximately 10–20 $\mathrm{mmol\,l^{-1}}$ (i.e., 1–4% of total dry weight) for corn (*Zea mays* L.), with the other main organic solutes, namely amino acids (10–20 $\mathrm{mmol\,l^{-1}}$) and sugars (90 $\mathrm{mmol\,l^{-1}}$), present in root cells. Higher concentrations are generally found in rhizosphere soil than in bulk soil. The carboxylic acids that play a significant role in complexing metal ions in soils are not utilized as rapidly as carbohydrates by the soil microorganisms. It appears that microorganisms preferentially utilize the carbohydrates in root exudates. Consequently, organic and amino acids may be the prevalent chemical substances in root exudates.

Typical organic acids found in root exudates are malic, malonic, acetic, citric, fumaric, succinic, lactic, tartaric, oxalic, and others. Organic acids of root exudates seem to have a major role in solubilization of mineral nutrients. Malic and citric acids are major root exudate components for some plant species under aseptic conditions. For example, in corn grown under

**Table 1** Composition of root exudates

| *Compositions* | *Substances identified* |
|---|---|
| Organic acid | Acetic, butyric, citric, glutaric, lactic, maleic, malic, malonic, oxalic, propionic, pyruvic, succinic, tartaric, valeric |
| Amino acid and amide | α-Alanine, β-alanine, arginine, asparagine, aspartic acid, cystine/cysteine, glutamine, glycine, histidine, lysine, methionine, phenylalanine, proline, serine/homoserine |
| Enzyme | Amylase, invertase, phosphatase, protease, polygalacturonase |
| Growth factor | *p*-Amino benzoic acid, auxins, biotin, choline, inositol, *n*-methyl nicotinic acid, niacin, pantothenate, pyridoxine, thiamine |
| Phenolic acid and coumarin | Caffeic acid, cinnamic acid, coumarin, ferulic acid, salycilic acid, syringic acid, vanillic acid |
| Sugar | Arabinose, fructose, fucose, galactose, glucose, maltose, oligosaccharide, raffinose, rhamnose, ribose, sucrose, xylose |
| Others | Nucleotide, flavonone, fatty acids, proteins, sterols, lipids, aliphatics, aromatics, carbohydrates |

sterile conditions, exudation of malic and citric acids increases significantly under nutrient deficiency.

Compared with the organic acids, amino acids appear to play a less significant role in complexing or mobilizing plant nutrients. The proteinaceous amino acids are not readily detected in root exudate solutions. The nonproteinaceous amino acids (phytosiderophores) in the rhizosphere have been shown to enhance the mobility of plant micronutrients in soils.

## Collection and Quantitation of Root Exudates

Upon reaching a predetermined growth stage, root exudates must be separated from the growing medium. If plants are grown in solution culture, aliquots of the nutrient solution can be purified and concentrated for the assay. The plants can also be transferred from a nutrient solution to another container and, upon a predetermined exposure time, root exudates are collected in water. When plants are grown in solid medium such as soil, root exudates are collected by first removing the plant from the soil then placing it in aerated and sterilized, double-distilled water for 2–6 hours. Subsequently, the exudates are separated and concentrated for analysis. Typically, solutions are filtered using 0.2- or 0.45-$\mu$m membrane filters that are then concentrated by rotary evaporation or by lyophilization. The collected material is stored at temperatures ranging from −20 to 4°C until analysis then diluted to a smaller volume for analysis.

Root exudates might be adsorbed on XAD-4, a porous polymer resin, which behaves in a way similar to C-18 high-pressure liquid chromatographic material when used in the reversed-phase mode. Root exudates trapped on the resin are extracted with either distilled water or methyl alcohol for lipids. The system has several advantages over previously used extraction procedures and is easily adaptable to large-scale recovery and extremely low-level concentrations of substances, as is very often the case with germination stimulants and other biologically active exudates.

Only a few attempts have been made to analyze comprehensively the chemical composition of root exudates. This is due to the necessity of employing a wide array of analytical methods for characterizing the different exudate components. Prior knowledge of the presence of substances is essential for selecting the methodologies for proper separation, identification, and quantification of substances from the mixtures. In addition, concentrations of substances in root exudates may be low, further complicating their detection and quantification.

The chemical behaviors of low-molecular-weight organic acids commonly found in root exudates are similar, making their separation in a chromatographic column a challenging analytical problem. A gas chromatographic (GC) approach based on differences in the chemical structure of organic acids has also been utilized. GC with mass spectrometry (GC-MS) analysis has been used to quantify the changes of quantities of root exudate in relation to colonization of roots by mycorrhizal fungi in which the concentrated root exudate solution is injected directly into the instrument. To enhance the resolution, root exudates can be separated into cationic, anionic, and neutral fractions prior to analysis by GC-MS using ion exchange resins. Compounds are then identified by comparing the retention time and mass spectra to respective reference compounds. Organic acids in the root exudates have also been determined by ion chromatography (IC) operating at room temperature. The separation is carried out on an IonPac AS 11 or AS 6 column (250 × 4 mm) protected by an IonPac AG 11 guard column (50 × 4 mm).

## Factors Affecting Root Exudates

The quantity and composition of root exudates are affected by environmental factors (Table 2). Acidification, chelation, precipitation, and redox reactions in the soil can change the nature and composition of root exudates. For example, nutritional stress

**Table 2** Influence of environmental factors on root exudation

| *Factors* | *Effect* |
|---|---|
| Age of the root | Roots in the initial stages secrete exudates more frequently |
| Light intensity | High light intensity increases exudation |
| Microorganism | Presence of microorganisms increases exudation |
| $NH_4^+/NO_3^-$ | Roots under $NH_4^+$ have lower organic acid concentrations |
| Nutrient availability | Nutrient deficiencies increase exudation |
| Oxygen status | Compositions are different under aerobic and anaerobic conditions |
| Plant species | Different plants and cultivars have different exudate compositions |
| Salinity | More organic and amino acids are produced under low salinity |
| Soil moisture | Relieving drought stress increases exudation |
| Soil pH | Acidification changes composition of exudates |
| Soil temperature | High temperature stimulates exudation |
| Soil texture | Sandy substrate produces greater amounts of exudates |
| Stress condition | Stress changes the composition of exudates |

experienced by plants, such as deficiencies of P, zinc (Zn), or iron (Fe), may increase root exudation. Plant roots alter the redox potential in the rhizosphere soil directly by excreting $CO_2$, and indirectly through the supply of readily available carbon sources for enhanced microbial respiration. In general, the redox potential is lower in the rhizosphere than in the bulk soil. The decrease in redox potential is likely to have important consequences in the redox reaction of contaminants such as arsenic (As), selenium (Se), Fe, manganese (Mn), and chromium (Cr).

Temperature and photoperiod have profound effects on the quality and quantity of root exudates because they affect photosynthesis, translocation, and respiration in plants. Increase in exudation at high temperature has been reported for many crops. The rate of exudation of certain organic acids by plants, while in vegetative and reproductive stages, may increase with the elevation of temperature and the extension of photoperiod, the mean rate being two or more times higher than the minimum exudation at low temperature with shorter photoperiod.

The imbalance in plant absorption of cation and anion affects the concentration of organic acids in the rhizosphere. In situations where roots take up cations in excess, particularly $K^+$, negative charges that are provided by organic acids, such as malic, malonic, oxalic, and citric acids, are required to maintain an electrostatic balance. Roots grown in $NH_4^+$-fertilized fields have lower concentrations of the organic acids than those grown in $NO_3^-$-fertilized fields.

## Role of Root Exudates on Plant Nutrient and Metal Dynamics

The enhanced activity of microorganisms, including mycorrhizal fungi, in the rhizosphere is important in relation to the bioavailability and mobility of nutrient ions, metals, and organic contaminants. For example, it has often been shown that the dissolution and release of water-insoluble fertilizer materials such as elemental sulfur (S) and apatite phosphate rocks are enhanced in the presence of plants; this has been related to the increased activity of S-oxidizing microorganisms and the release of organic acids. The enhanced microbial activity is likely to reduce the mobility of nonreactive ions such as nitrate and sulfate through microbial immobilization.

Interactions of Fe, Mn, copper (Cu), and Zn, and the effect of organic acids on P dissolution in soils have been researched. In addition to their role in plant nutrition, root exudates may affect vital soil processes, such as the soil organic matter transformation, cycling of nutrients in soils, and development of stable soil structure. Root exudates (e.g., polysaccharide gum) enhance stabilization of soil structure in the rhizosphere by increasing the strength of bonds between soil particles.

Organic acids such as oxalic, lactic, and succinic exuded by roots are effective in releasing metals from sparingly soluble mineral forms by forming stable chelate complexes with the metal ions. The extent of complexation depends on the organic acids involved (number of carboxylic groups), the concentration and type of metal, and the pH of the soil. Organic acids with only one carboxyl group, such as acetic, formic, and lactic acids, have very little metal-complexing ability. Based on the stability constants summarized in **Table 3**, malic and oxalic acids have high affinities for metals. These are the most frequently found organic acids in soils. However, this does not always entail efficient complexation at all soil pH levels. For example, using Geochem-PC, a descendant of the multipurpose chemical speciation program and the PC-compatible version of Geochem, the complexation of Fe by malic and oxalic acids is shown to be highly dependent upon soil pH, with little or no complexation at high soil pH.

Biogeochemical reactions occurring in the rhizosphere play an important role in the bioavailability of metals. These reactions can be heavily influenced, notably by the release of protons and organic acids from the roots. The acidifying protons can affect the chemical speciation and mobility of micronutrient metals and heavy metals/metalloids, possibly favoring the formation of ionic species in the rhizosphere. Additionally, organic acids such as those mentioned earlier had the ability to complex metals, with the resultant metalorgano complexes possessing greater mobility. Thus the combined effects of the protons and organic acids may result in chemical speciation and much-enhanced mobility of the metals that could facilitate their uptake by the root or leaching.

Phosphate dissolution rate can be greatly accelerated in soil in the presence of organic acids of root exudates, depending on soil type and concentration of organic acids. There are several possible mechanisms by which P release can occur. The first mechanism involves competitive sorption reactions (direct ligand exchange) in which exuded organic acids compete with P for common sorption sites on soil particles, thus releasing P from the solid phases. For example, citrate may be able to desorb phosphate from sesquioxide surfaces by anion exchange. The second mechanism could involve the modification of the surface characteristics in which organic ligand sorption modifies the surface charge of P sorbents. Anionic organic acids may be sorbed by minerals in the rhizosphere, which may lead to an increase in the net negative charge of clays. Along with this,

**Table 3** Stability constants (*K*) of organic acid–metal complexes

| | | *Log* K[b] | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Metal ion* | *M:L:H ratio*[a] | *Acetic* | *Butyric* | *Glutaric* | *Lactic* | *Maleic* | *Oxalic* | *Propionic* | *Pyruvic* | *Succinic* | *Tartaric* | *Valeric* |
| $H^+$ | 1:1:0 | 4.76 | 4.82 | 5.43 | 3.86 | 6.33 | 4.27 | 4.87 | 2.55 | 5.64 | 4.37 | 4.78 |
| | 1:2:0 | - | - | - | - | - | - | - | - | - | - | - |
| $Cd^{2+}$ | 1:1:0 | 1.93 | 1.25[f] | 1.60[c,f] | 1.70 | 2.40[l] | 3.89[d,e] | 1.19[g] | - | 2.72 | - | - |
| | 1:2:0 | 3.15 | 1.98[f] | - | 2.10[g] | 3.60[j] | - | 1.86[g] | - | - | - | - |
| $Cr^{2+}$ | 1:1:0 | 1.25[j] | - | - | - | - | 3.85[j] | - | - | - | - | - |
| | 1:2:0 | 2.15[j] | - | - | - | - | 6.81[j] | - | - | - | - | - |
| $Cu^{2+}$ | 1:1:0 | 2.22 | 2.14 | 3.16 | 3.02 | 3.40[l] | 6.23 | 2.22 | 2.20 | 3.28 | 2.90[g] | 2.08 |
| | 1:2:0 | 3.63 | 2.98[f] | - | 4.84 | 4.90[k] | 10.3 | 3.00[f] | 4.90 | - | - | - |
| $Ni^{2+}$ | 1:1:0 | 1.43 | 2.14 | 1.60[l] | 2.22 | 2.00[l] | 5.16 | 0.78[g] | 1.12[h] | 1.60[l] | 2.06[g] | - |
| | 1:2:0 | - | 2.98[f] | - | 2.76[g] | - | - | 1.26[f] | 0.46[h] | - | - | - |
| $Pb^{2+}$ | 1:1:0 | 2.68 | 2.13[f] | 2.48[c,f] | 2.78 | 2.75[g] | 4.91 | 2.08[f] | 2.04[f] | 5.64 | 2.60[g] | - |
| | 1:2:0 | 4.08 | 3.74[f] | 3.45[c,f] | 2.88[g] | 4.03[g] | 6.76 | 4.15 | 3.40[f] | 3.73[g] | 4.00[g] | - |
| $Zn^{2+}$ | 1:1:0 | 1.57 | 0.98[f] | 2.45 | 2.22 | 2.00[l] | 4.87 | 0.85[g] | 1.26[h] | 2.52 | 3.82 | - |
| | 1:2:0 | 1.90[l] | 1.65[f] | 1.74[g] | 3.75 | - | 7.65 | 1.23[g] | 1.98[h] | 2.00[g] | 5.00 | - |

[a] M:H:L ratio indicates the stoichiometry of the complex; M, metal; L, ligand; H, proton.

[b] Temperature (°C) at which *K* was determined at 25°C unless indicated otherwise; and ionic strength (moles per liter) at which *K* was determined at zero unless indicated.

[c] 30°C.

[d] 18°C.

[e] 3.0.

[f] 2.0.

[g] 1.0.

[h] 0.5.

[i] 0.3.

[j] 0.2.

[k] 0.16.

[l] 0.1.

acidification due to proton exudation accelerates dissolution reaction affecting clay minerals. The third mechanism involves the complexation of metal ions in the solid, which constitute the exchange matrix holding the P. The action of organic chelates in solubilizing phosphates and phosphate minerals is attributed to the formation of complexes with calcium (Ca), Fe, or aluminum (Al), thereby releasing phosphate in water-soluble form. Solubilization of P compounds by organic acids is accomplished by the formation of complexes of organic acid ligands with P-associated metal ions such as Fe, Al, and Ca. The substitution of organic ligands for orthophosphates frees up P for plant uptake. This process depends on the number and position of carboxyl (-COOH) and phenolic (-OH) functional groups in the organic acids.

The pH of the rhizosphere is important in determining contaminant and nutrient mobilization and uptake. Root-induced pH change in the rhizosphere is a known phenomenon and has been shown to have an effect on the bioavailability or solubility of nutrients such as P, Fe, Mn, Zn, and Cu by rhizosphere acidification. Rhizosphere pH can be more than 2 units lower than the bulk soil pH. Several hypotheses have been postulated to explain the differential ability of plant species to affect rhizosphere pH. These include differences in root exudates, especially low-molecular-weight organic acids and the evolution of respiratory $CO_2$ from differential uptake in cations/anions.

## Root Exudates and Biodegradation of Organic Contaminants

Rhizosphere degradation of organic compounds is often a key factor in successfully applying a phytoremediation-based strategy. Secretion of plant exudates not only stimulates overall microbial activity, but some induce specific bacterial- and fungal-degradation pathways. For instance, phenolic root exudates trigger cometabolism of chlorinated organic solvents, including trichloroethylene (TCE), by rhizosphere bacteria. The combined activity of mycorrhizal fungi and bacteria might help to overcome environmental stress imposed on plants by the presence of organic xenobiotics: consortia of mycorrhizal fungi and pseudomonads with toluate-degrading capacity improve the survival of mycorrhizal symbionts and, at the same time, promote toluate degradation in the rhizosphere. Some hydrophobic organic xenobiotics such as polyaromatic hydrocarbons (PAH) are not taken up efficiently by plants. Phytoremediation of PAHs, therefore, strongly depends on rhizosphere processes such as enhanced degradation by bacteria or fungi.

Root exudates affect the bioavailability of nutrients by directly influencing the chemistry of the rhizosphere soils and indirectly enhancing microbial activity in the rhizosphere soils. Microorganisms utilize sugars, organic acids, amino acids, vitamins, and high-molecular-weight compounds exuded by roots as sources of C and substrates for growth. Microbial metabolites may in turn increase root exudation through stimulation of root activity and growth, thus increasing the surface area over which the exudation occurs. Similarly, Fe deficiency leads to the development of shifted bacterial communities that differ qualitatively and quantitatively in species composition.

Microbial degradation is the major process by which organic contaminant residues are removed from the soil. The rate of degradation of organic contaminants has often been found to be faster in the presence of growing plants. **Figure 1** indicates the distribution of the water-soluble C, the microbial activity, and the 2,4-dichlorophenoxyacetic acid (2,4-D) residues in soil samples taken at different distances from the rhizoplane in a microcosm root container. There is an increase in the amount of soluble C in soil samples taken from close to the root surface. The increase in soluble carbon is related to the rhizodeposition of root exudates, which include low-molecular-weight organic acids, carbohydrates, and amino acids. There is a corresponding increase in microbial activity as measured by the amount of oxygen consumed by the microorganisms in this soil section. This enhanced microbial activity is related to an increase in the supply of organic substrate (i.e., root exudates) as a source of C and energy for microbial growth. There is a decrease in the amount of 2,4-D remaining in the soil, indicating that microbial degradation of pesticides occurs during plant growth. The amount of residual 2,4-D remaining in the pesticide cores is less in the soil section close to the root surface than in the other section. Since grass roots are unlikely to absorb considerable amounts of 2,4-D residues, by inference the low amount of 2,4-D residues close to the root core is attributed to the root-induced degradation of the pesticide. The enhanced degradation of pesticide in the soil section close to the root surface is related to the rhizosphere-induced cometabolism of the pesticide that requires the presence of a growth substrate other than the compound to be mineralized. Plant roots excrete a wide range of root exudates, and the availability of these exudates is considered to be the major cause for the existence of an enriched microbial population in the rhizosphere, resulting in accelerated degradation of the pesticide.

It is likely that an increase in the activity of pesticide-degrading microorganisms in the rhizosphere of certain plants is a mechanism by which plants are protected from the toxic effects of pesticides. Selective enrichment of microorganisms is likely to have a significant impact on the rhizoremediation (or

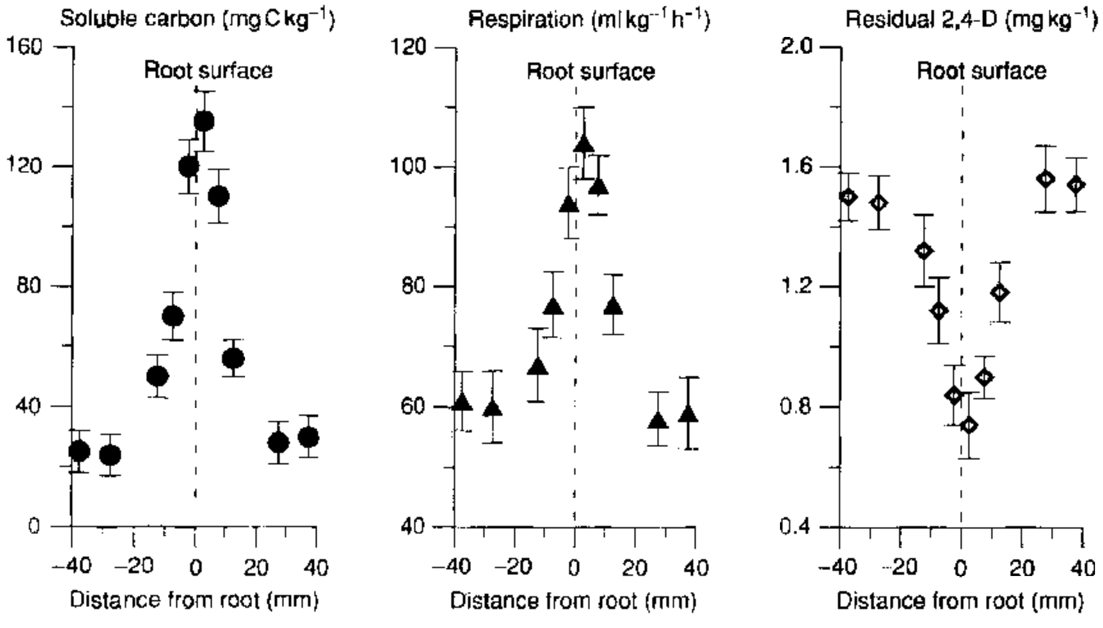


**Figure 1** Distribution of soluble carbon, biological activity, and the residual 2,4-dichlorophenoxyacetic acid (2,4-D) in thin sections of soils in the pesticide cores.

more commonly, phytoremediation) of recalcitrant organic contaminants in soils. It has become apparent that the role of the rhizosphere in the natural attenuation of pollutants in soils is being recognized increasingly.

## Summary

Plants and their associated microorganisms from complex communities, called rhizosphere communities, have diverse functions and capabilities that are important in many processes related to C management and recycling, biochemical cycling of nitrogen and P, biomass production and conversion to biofuel, and phytoremediation of heavy metals, radionuclides, and organic xenobiotics.

The relationships and interactions between and among microorganisms, plants, and water-borne chemicals are still not well-understood. However, many processes, such as the successful growth of plants, the optimization of biomass production, and carbon sequestration require that we acquire such understanding at the level of diversity, structure, functional interdependence, and capabilities of rhizosphere communities.

## Acknowledgments

The authors wish to acknowledge the support provided by Savannah River Ecology Laboratory, University of Georgia, through US Department of Energy contract number DE-FC-09-96SR18546.

*See also:* **Bacteria:** Plant Growth-Promoting; **Mycorrhizal Fungi**; **Root Architecture and Growth**

## Further Reading

Adriano DC (2001) *Trace Elements in Terrestrial Environments: Biogeochemistry, Bioavailability, and Risks of Metals*, 2nd edn. New York: Springer.

Curl EA and Truelove B (1986) *The Rhizosphere*. Berlin, Germany: Springer-Verlag.

Gobran GR, Wenzel WW, and Lombi E (2001) *Trace Elements in the Rhizosphere*. Boca Raton, FL: CRC Press.

Harley JL and Russell RS (1979) *The Soil–Root Interface*. London, UK: Academic Press.

Koo B-J (2001) *Assessing Bioavailability of Metals in Biosolid-Treated Soils: Root Exudates and Their Effects on Solubility of Metals*. PhD Dissertation. Riverside, CA: University of California Press.

Marschner H (1995) *Mineral Nutrition of Higher Plants*, 2nd edn. London, UK: Academic Press.

# SALINATION PROCESSES

**I Shainberg and G J Levy**, The Volcani Center, Bet Dagan, Israel

## Origin and Distribution of Saline Soils

Excessive levels of salts occur in large areas of soil around the world and profoundly affect land use. Two major types of soil have been defined within this group, saline and sodic soils. The former refers to a soil that contains excess soluble salts that impair its productivity. The excess salts are comprised largely of the chloride and sulfate anions and sodium, magnesium, and calcium cations in the soil solution. Sodic soils contain excess sodium in the exchange complex that impairs their productivity. Sodic soils may or may not contain excess soluble salts and may contain free sodium carbonate.

Estimates of the areas of the globe covered by salt-affected soils are presented in **Table 1**. The salt content of soils above which plant growth is affected depends upon several factors, among which are the texture of the soil, the composition of the salt, and the species of the plant. The electrical conductivity (EC) of the saturation extract has been adopted by the US Salinity Laboratory as the preferred scale for general use in estimating soil salinity; and saline soils are defined as soils with EC exceeding $4\,\mathrm{dS\,m^{-1}}$. Based on this classification, the total area of saline soils on the globe is 351.5 millions hectares, with most of the saline soils located in Asia ($194.9 \times 10^6$ ha) (**Table 1**). The decision regarding what level of exchangeable sodium in the soil constitutes an excessive degree of saturation is complicated by the fact that there is no sharp change in the properties of the soil as the degree of saturation with exchangeable sodium is increased, and that the effect of sodicity depends on the EC of the soil solution, the soil texture, the soil mineralogy, organic matter content, sesquioxide content, and the soil conditions (e.g., the wetting rate of the soil, the antecedent moisture content, and the aging effect). Thus, in the USA, and subsequently throughout much of the rest of the world, a value of exchangeable sodium percentage (ESP) greater than 15 is the criterion for separating sodic soils, whereas in Australia a value of ESP greater than 6 is used (*See* Sodic Soils). This explains the disproportionately large areas of sodic soils in Australia (**Table 1**). Whatever the definition of sodic soils, it is evident from **Table 1** that the area of sodic soils in the world is approximately twice that of saline soils.

The original and direct source of the salts in soils are the primary minerals found in soils and in exposed rocks of the Earth's crust. During the process of chemical weathering, which involves hydrolysis, hydration, solution, oxidation, and carbonation, the salts are gradually released and made soluble. Although weathering of primary minerals is the indirect source of nearly all soluble salts, there are few instances where sufficient salts have accumulated from this source alone to form a saline soil. Saline soils usually occur in areas that receive salts from other locations, and water is the primary carrier. The ocean is the source of salts in low-lying soils along the margin of sea coasts. Sometimes salt is moved inland through the transportation of spray by winds. More commonly, however, the direct source of salts is surface water and groundwater. Water acts as a source of salts when used for irrigation and when groundwater rises close to the soil surface.

Saline soils occur for the most part in regions of arid or semiarid climate. In arid regions, leaching and transportation of soluble salts to the ocean are not as complete as in humid regions. Restricted drainage is a factor contributing to the salinization of soils and may involve the presence of a high groundwater table. Where there is a high water table, upward movement of saline groundwater and evaporation of surface water result in the formation of saline soil.

## Properties of Saline and Sodic Soils

Properties of saline and sodic soils are determined by the composition of the soil solution and that of the solid phase. The composition of the soil solution is determined by the following characteristics:

1. Total concentration of soluble salt, or 'salinity';
2. Concentration of sodium relative to other cations, or 'sodicity';
3. Anionic composition of the water, especially that of bicarbonate and carbonate;
4. Concentration of elements toxic to plants (e.g., boron, B).

### Total Salt Concentration

The effect of salt on crop growth is largely of an osmotic nature. Osmotic pressure is a colligative property and is related to the total salt concentration rather than the concentration of individual ionic species. Total salt concentration is determined by total dissolved solids (TDS) in milligrams per liter of salts, by the ionic concentration in millimoles of charge per liter and by the electrical conductivity (EC) in decisiemens per meter. The relationship between salt concentration, either millimoles of charge per liter or milligrams per liter, and EC of various salt solutions is presented in **Figure 1a and 1b**, respectively. For a mixture of salts, the relationship between TDS and EC is given by:

$$\text{TDS (ppm)} = 640 \times \text{EC (dS/m)} \qquad [1]$$

**Table 1** Global distribution of salt-affected soils

| | *Area (millions ha)* | | |
|---|---|---|---|
| *Continent* | *Saline* | *Sodic (alkali)* | *Total* |
| North America | 6.2 | 9.6 | 15.8 |
| Central America | 2.0 | – | 2.0 |
| South America | 69.4 | 59.6 | 129.0 |
| Africa | 53.5 | 27.0 | 80.5 |
| South Asia | 83.3 | 1.8 | 85.1 |
| North & Central Asia | 91.6 | 120.1 | 211.7 |
| Southeast Asia | 20.0 | – | 20.0 |
| Australasia | 17.4 | 340.0 | 357.4 |
| Europe | 7.8 | 22.9 | 30.7 |
| Total | 351.5 | 581.0 | 932.2 |

Reproduced from Sumner ME, Rengasamy P, and Naidu R (1998) Sodic soils: a reappraisal. In: Sumner ME and Naidu R (eds) *Sodic Soils*, pp. 3–17. New York: Oxford University Press.

### Sodicity and Sodium Adsorption Ratio

When soils are leached with salt solutions until equilibrium is attained, a linear relation is obtained between the ratio of the concentrations of exchangeable sodium and divalent cations (ESR, exchangeable sodium ratio) and the ratio of the concentration (in millimoles of charge per liter) of the sodium cation to the square root of the concentration of the divalent cations in the solution (SAR) (**Figure 2**).

$$\text{SAR} = [\text{Na}]/([\text{Ca} + \text{Mg}]/2)^{1/2} \qquad [2]$$

where [Na] and [Ca + Mg] are the concentration of the soluble ions in millimoles of charge per liter. Analysis of a large number of soil samples from the western USA has led to the empirical equation:

$$\text{ESR} = \text{Na}_{\text{exch}}/[\text{Ca} + \text{Mg}]_{\text{exch}} = -0.0126 + 0.01475 \qquad [3]$$

where $\text{Na}_{\text{exch}}$ and $[\text{Ca} + \text{Mg}]_{\text{exch}}$ are the concentrations of the exchangeable ions, presented in Figure 2.

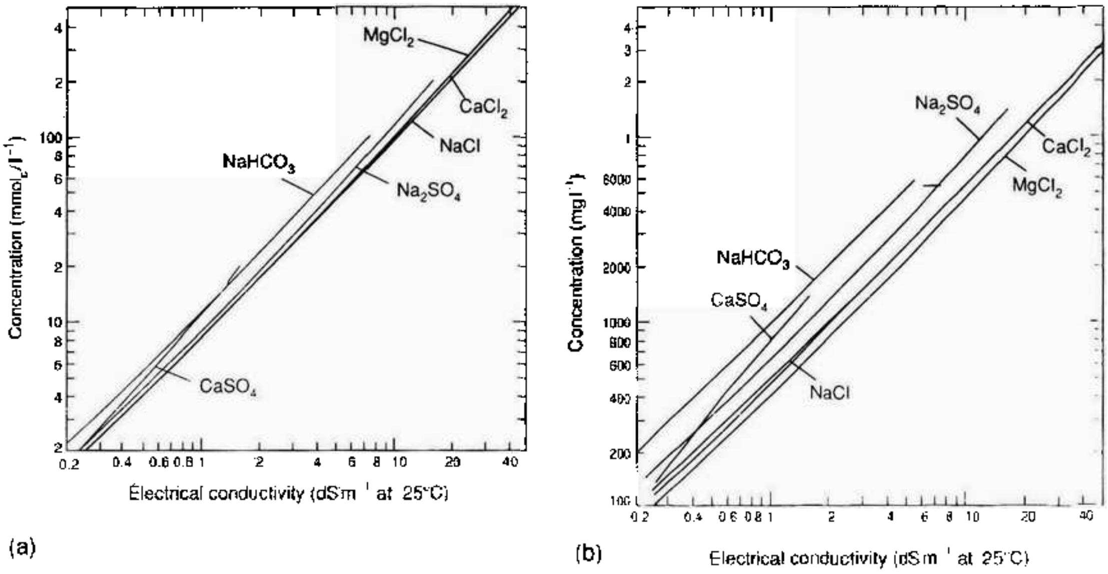


**Figure 1** Relationship between salt concentration and electrical conductivity of various salt solutions: salt concentration in (a) millimoles of charge per liter and (b) milligrams per liter. Reproduced with permission from Shainberg I and Osler JD (1978) Quality of Irrigation Water. International Irrigation Information Center. Volcani Center, Israel.

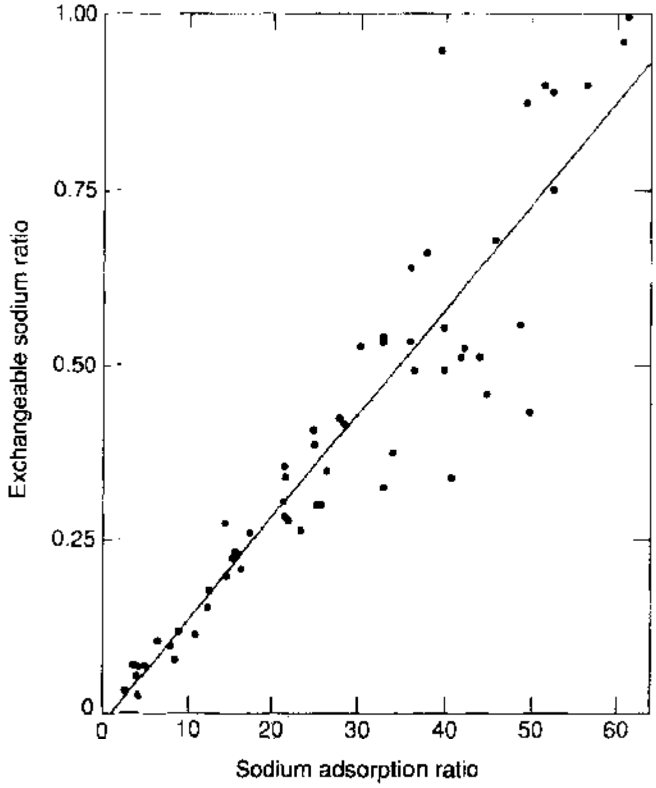


**Figure 2** Exchangeable sodium ratio (ESR) of soil samples from the western USA as related to the sodium adsorption ratio (SAR) of soil extracts. Reproduced with permission from Shainberg I and Oster JD (1978) Quality of Irrigation Water. International Irrigation Information Center, Volcani Center, Israel.

A nomogram relating the ESP of soils to the SAR of the soil solution is presented in Figure 3.

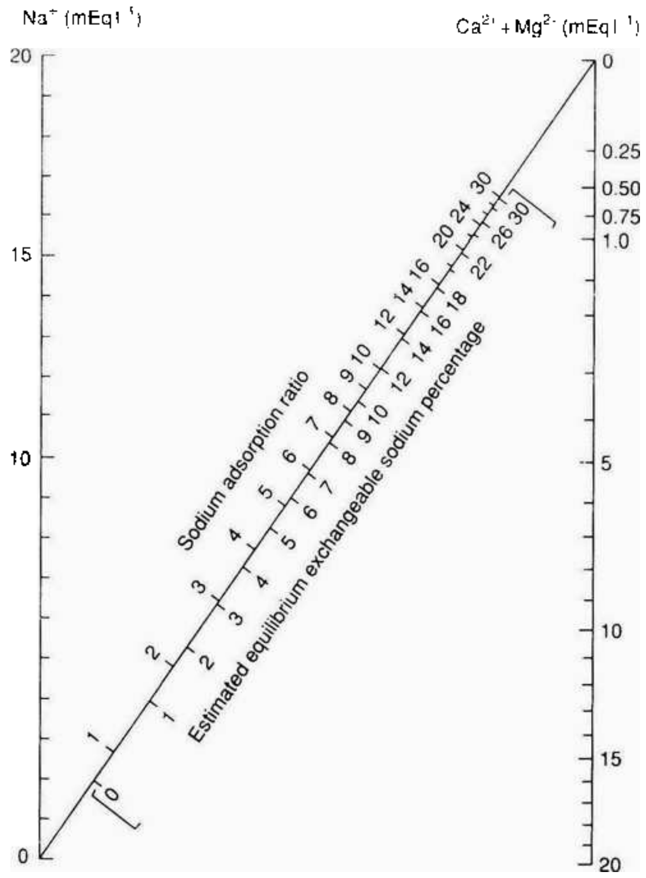


**Figure 3** Nomogram for determining the SAR value of irrigation water and for estimating the corresponding ESP value of a soil equilibrated with this water. Courtesy of US Department of Agriculture, Washington.

### Anionic Composition

A major factor affecting the final SAR value of soil water is the change in Ca and Mg concentration due to precipitation or dissolution of alkaline earth carbonates. In irrigation waters containing high concentrations of bicarbonate ions, there is a tendency for calcium and, to a lesser extent, magnesium to precipitate in the form of carbonate as the soil solution becomes more concentrated, thus leading to an increase in the SAR of the soil solution and consequently to an increase in the ESP of the soil. Bicarbonate concentration in underground water may be high because of the high partial pressure of $CO_2$. On exposure to the atmosphere, the excess $CO_2$ is released, while calcium and magnesium carbonates precipitate. Eaton assumed that all calcium and magnesium would precipitate as carbonates and proposed the concept of residual sodium carbonate (RSC) for evaluating high-carbonate water:

$$RSC = (CO_3 + HCO_3) - (Ca + Mg) \qquad [4]$$

where the concentration of the ions is expressed in millimoles of charge per liter. From the results of some limited leaching experiments, Eaton concluded that water with RSC greater than 2.5 is not suitable for irrigation. He defined the range of RSC between 1.25 and 2.5 as marginal, and less than 1.25 as safe. Although the RSC is still commonly reported in irrigation water analyses, its use is not justified. A better estimate of the precipitation tendency of carbonates is provided by the so-called saturation index. However, its application to soils leached with bicarbonate water is complicated and limited in its accuracy.

### Concentration of Elements Toxic to Plants

Some soluble trace elements may have an inhibitory effect on plant growth. Elements in this category include B, lithium, selenium, and some heavy metals. In terms of saline soils, B is probably the most important. B is essential for plant growth but is exceedingly toxic at concentrations only slightly greater than optimum. B deficiency is found primarily in humid regions or in sandy soils, while toxicity occurs most frequently in arid and semiarid regions. Crops vary in their response to B. The threshold concentration for the most sensitive crop is 0.5 mg l$^{-1}$ (lemon) and for the most tolerant is 6–10 mg l$^{-1}$ (cotton). Since an appreciable proportion of the B added to soils in irrigation water is sorbed by the soil, and plants respond to the B in soil solution, then conditions affecting equilibrium between sorbed and soluble B are highly relevant to considerations of B toxicity. Owing to adsorption, irrigation water containing marginal levels of B may not be toxic immediately; but after prolonged irrigation with such water, soluble B levels equal or exceed those of irrigation water.

## Response of Crops to Salinity

Since the publication of Handbook 60 by the US Salinity Laboratory Staff, the salt tolerance of crops has been reevaluated as a result of new data obtained under conditions of variable soil salinity. It is now thought that most plants are able to withstand higher salinities in the drainage water than previously reported.

Salinity affects plant growth by three mechanisms:

1. Osmotic effects limit the ability of plants to absorb water from the soil solution. The osmotic effect is demonstrated when plant growth is reduced in a similar way by iso-osmotic solutions, and when osmotic and soil-water matric potential have a similar and additive effect on plant growth;
2. Specific ion toxicity results from excessive concentrations of Na or Cl ions, an effect most commonly found in woody species;
3. Changes in soil physical and chemical properties can affect plant production.

Crop yield is not reduced until a threshold salinity level is exceeded. Beyond the threshold level, the yield decreases linearly with rising salinity (Figure 4). The salinity values at zero yields provide an estimate of maximum salinity that plants can tolerate. Plants can be divided into four groups (sensitive to tolerant) according to their response to salinity. Of the field crops, barley, cotton, and sugar beet belong to the tolerant group, wheat and soybean belong to the moderately tolerant group, peanut and rice belong to the moderately sensitive group, and beans and cowpeas belong to the salt-sensitive group.

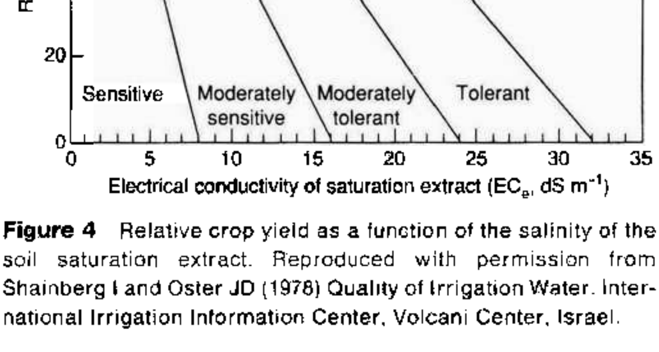


**Figure 4** Relative crop yield as a function of the salinity of the soil saturation extract. Reproduced with permission from Shainberg I and Oster JD (1978) Quality of Irrigation Water. International Irrigation Information Center, Volcani Center, Israel.

Plant response to salinity depends also on soil, climate, and plant factors. Soil-water content and frequency of irrigation affect the tolerance of crops to soil salinity. Shortening the irrigation intervals minimizes the deleterious effect of salinity. The salt tolerance of many crops is enhanced when daily drip irrigation is used. Climate can also modify plant response to salinity. Salt tolerance is often reduced under hot, dry conditions; also, crops appear to be more salt-tolerant in areas with air pollution that limits plant growth.

Plant factors such as stage of growth, variety, and rootstock affect plant response to salinity. Rice, barley, wheat, and corn are most sensitive to salinity during the early seedling stage. Beet and sunflower are more sensitive to salinity during germination than in later stages of growth. Salinity effects on the vegetative growth of many plants (e.g., cotton, wheat) are greater than on seed or fiber production. Normally, salinity suppresses top growth more than root growth. The type of rootstock of woody plants may also affect the specific tolerance of fruit crops to salinity.

## Response of Soils to Salinity and Sodicity

The physical and mechanical properties of soils (e.g., clay swelling and dispersion, aggregate stability, hydraulic conductivity, runoff, and erosion) are affected by soil salinity and sodicity. The main factor that determines the extent of the deleterious effect of Na on soil properties is the salt concentration in the soil solution. With low soil-solution salinity, as is the case in humid regions or during the rainy season in semiarid regions, even a low sodicity (ESP $<5$) has a marked and irreversible negative effect on the physical properties of most soils. Conversely, when the salt concentration reaches even a moderate level (10–20 $mmol_c/l$), high sodicity levels ($10<$ ESP $<30$) cause only small to moderate changes in the physical and hydraulic properties of soils. This is demonstrated in Israel, where irrigation water of 4.6 dS $m^{-1}$ and SAR of 26 are used for irrigation of cotton. During the irrigation season (summer), no deterioration of soil physical properties is observed in spite of the high soil sodicity, because the high salt concentration prevents the adverse effects of sodicity. On the other hand, during the subsequent rainy season, the electrolyte concentration in the soil solution at the soil surface becomes low, rendering the soil susceptible to crusting, low permeability, and increased runoff and erosion.

The response of soils to sodicity and salinity depends on the soil properties (e.g., texture, clay mineralogy, lime and sesquioxides content). Smectitic soils are more susceptible to sodicity than kaolinitic soils. Calcareous soils are less susceptible to sodicity than noncalcareous soils. Soils rich in sesquioxides are less affected by sodicity than soils with low sesquioxide content. However, the response of soils to sodicity depends also on time-dependent properties such as wetting rate of the dry soil, antecedent moisture content, and aging (time since cultivation or wetting). Fast wetting of dry soils disintegrates the aggregates and increases the susceptibility of the soils to sodicity. The effect of wetting rate depends on the antecedent moisture content of the soils. Less slaking of soil aggregates upon fast wetting occurs when the antecedent moisture content of the soil is high and the moist soil is less susceptible to the deleterious effect of sodicity. Similarly, upon aging of a moist soil, the stability of the soils' aggregates increases (aging effect) and their susceptibility to sodicity decreases.

## Reclamation of Salt-Affected Soils

In ideal soils under piston flow conditions, the EC of soil water and applied water would be equal after the passage of one pore volume of rainfall or irrigation water. However, soils are not ideal for two reasons: (1) water flows faster through soil cracks and large pores between soil aggregates than through smaller pores within aggregates. This difference in localized water flow velocities is commonly referred to as 'bypass flow,' or 'preferential flow'; (2) salt diffusion and chemical reactions are not sufficiently rapid to allow salt concentration to be the same among all pores of differing sizes, particularly during periods of rapid water movement.

Salt transport models have provided considerable insight into these processes. However, suitable mathematical models are not available to describe the multiple concurrent processes of water flow, including spatially variable hydraulic properties and all the chemical reactions – exchange, dissolution, precipitation, and $CO_2$ liquid–gas equilibriums – involved in salt transport on a field scale. Consequently, reclamation guidelines are largely based on experimental relations obtained from field and lysimeter reclamation experiments.

The method of water application and soil type are the primary variables affecting the amount of water required to reclaim saline soils. For saline soils, 60% or more of the salts initially present is leached by a depth of water equal to the depth of soil (**Figure 5**) under continuously ponded conditions. Sandy loam soils with low saturated water contents have higher leaching efficiency than clay loam or clay soils (**Figure 5**). Soil pores in sandy soils are more uniform in diameter than in clay loam or clay soils. The finer-textured soils can have large cracks with large pores when dry, and fine pores within aggregates when wet. Such bypass channels reduce leaching efficiency.

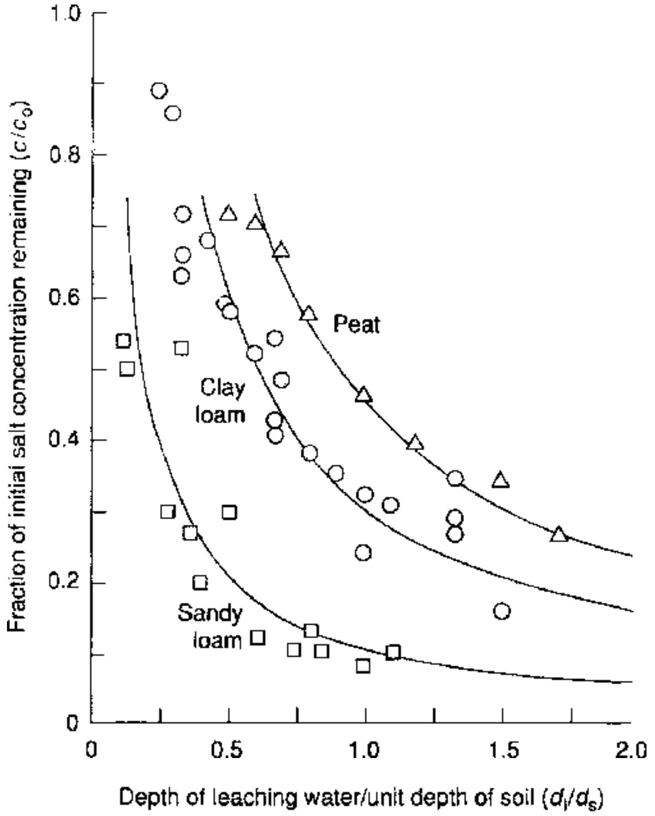


**Figure 5** Depth of leaching water per unit depth of soil required to reclaim a saline soil by continuous ponding. (Adapted from Oster JD, Shainberg I, and Abrol IP (1996) Reclamation of salt affected soils. In: Agassi M (ed.) *Soil Erosion Conservation and Rehabilitation*, pp. 315–342. New York: Marcel Dekker.)

The water requirement for leaching can be reduced by intermittent application of ponded water, particularly for fine-textured soils (**Figure 6**). Under intermittent ponding, unsaturated conditions prevail in the soil profile and the inefficient bypass flow in large cracks and pores is prevented. Similarly, with sprinkler irrigation, reclamation occurs under unsaturated conditions and the efficiency of leaching increases.

## Reclamation of Sodic Soils

The reclamation of sodic soils depends on the supply of Ca salts to replace exchangeable sodium and in maintaining adequate soil salinity to counteract the

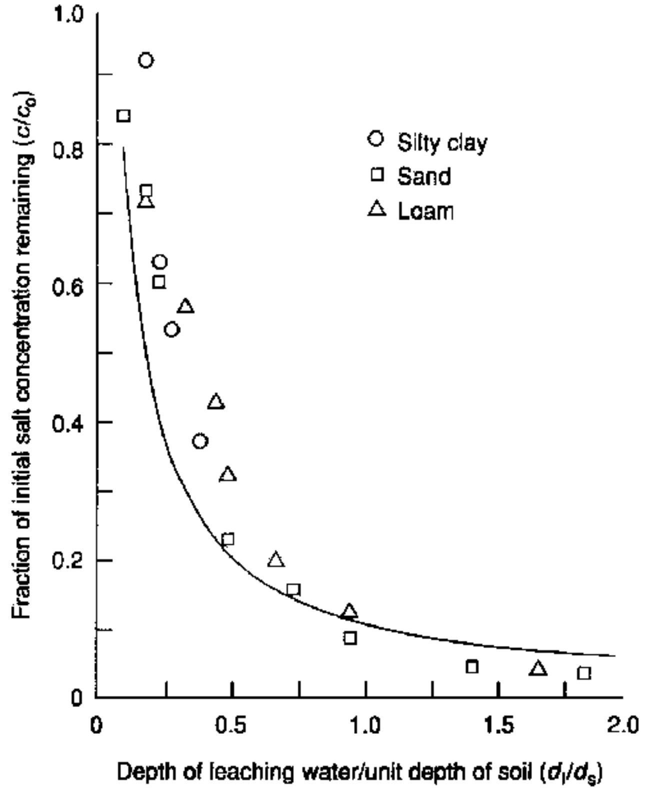


**Figure 6** Depth of leaching water per unit depth of soil required to reclaim a saline soil by ponding water intermittently. (Adapted from Oster JD, Shainberg I, and Abrol IP (1996) Reclamation of salt affected soils. In: Agassi M (ed.) *Soil Erosion Conservation and Rehabilitation*, pp. 315–342. New York: Marcel Dekker.)

adverse effect of sodicity on the hydraulic properties of the soils.

Gypsum, either incorporated into the soil or left on the soil surface, is the calcium source most commonly used to reclaim sodic soils and to improve soil water infiltration that has been decreased by low salinities. Sources include mineral deposits and phosphogypsum, a by-product of the phosphate fertilizer industry. When gypsum is incorporated into the soil, the reduction in ESP upon irrigation and leaching is primarily limited to the soil depth interval where gypsum is present. This is a consequence of the greater selectivity of exchange sites for calcium than for sodium. The gypsum required for reclamation can be estimated from the amount of exchangeable sodium in the soil layer to be reclaimed – the gypsum requirement concept.

Since the soil surface is most susceptible to sodicity, reclamation of sodic soils by spreading phosphogypsum at the soil surface is being practiced in Israel. The amounts of gypsum that are applied depend on the depth of rain, assuming that gypsum concentration in the infiltrating rain reaches 50% of gypsum saturation. Thus, in a region with winter rain of 300 mm, the amount of gypsum spread at the soil surface is 3 Mg ha$^{-1}$, and, in a region with 500 mm of rain, the amount of gypsum applied is 5 Mg ha$^{-1}$. In the commercial fields in Israel that have been irrigated with saline and sodic water as described above, phosphogypsum is spread annually on the soil surface, following tillage in the autumn, at a rate of 3–5 Mg ha$^{-1}$. This prevents seal formation and maintains high infiltration rates which, in turn, provide sufficient infiltration of rainfall to leach salts from the root zone. Dissolution of the gypsum provides enough electrolytes to prevent seal formation and maintain high infiltration rates in spite of the fact that the soil surface is not completely reclaimed. This management in the rainy season, coupled with adequate irrigation with the saline-sodic water to meet crop needs during the summer months, results in cotton yields that are similar to those obtained when nonsaline water is used for irrigation.

## Summary

Under arid and semiarid conditions, and in regions of poor natural drainage, there is a real hazard of salt accumulation in soils. The processes by which soluble salts accumulate in soils include irrigation with water containing salts, upward movement of moderate-to-saline water from high water tables, and weathering of primary and secondary minerals in the soils. Accumulation of soluble salts in the soil solution imposes stress on growing crops that can lead to decreased yields and, in severe cases, complete crop failure. Use of poor-quality water is increasing due to growing municipal demands for the available supplies of good-quality water and the need to dispose of municipal wastewater and agricultural drainage water. As a result, a better understanding of soil reclamation is becoming an increasingly important component of water and soil management to ensure the long-term sustainability of irrigated agriculture.

*See also:* **Sodic Soils**

## Further Reading

Bower CA, Ogata G, and Tucker JM (1968) Sodium hazard of irrigation water as influenced by leaching fraction and precipitation of carbonate. *Soil Science* 106: 24–34.

Bresler E, Mcneal BL, and Carter DL (1982) *Saline and Sodic Soils*. Berlin, Germany: Springer-Verlag.

Eaton FM (1950) Significance of carbonates in irrigation water. *Soil Science* 69: 123–133.

Keren R (1999) Salinity. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 1–25. Boca Raton, FL: CRC Press.

Levy GJ, Shainberg I, and Miller WP (1998) Physical properties of sodic soils. In: Sumner ME and Naidu R (eds) *Sodic Soils*, pp. 77–94. Oxford, UK: Oxford University Press.

Mass EV (1990) Crop salt tolerance. In: Tanji KK (ed.) *Agriculture Salinity Assessment and Management*. pp. 262–304. New York: American Society of Civil Engineers.

Oster JD, Shainberg I, and Abrol IP (1996) Reclamation of salt affected soils. In: Agassi M (ed.) *Soil Erosion Conservation and Rehabilitation*, pp. 315–342. New York: Marcel Dekker.

Shainberg I and Letey J (1984) Response of soils to saline and sodic conditions. *Hilgardia* 52: 1–57.

Shainberg I, Levy GJ, Goldstein D, Mamedov A, and Letey J (2001) Prewetting rate and sodicity effects on the hydraulic conductivity of soils. *Australian Journal of Soil Research* 39: 1279–1291.

Suarez DL, Wood JD, and Abrahim I (1992) Reevaluation of calcite supersaturation in soils. *Soil Science Society of America Journal* 56: 1776–1784.

Sumner ME, Rengasamy P, and Naidu R (1998) Sodic soils: a reappraisal. In: Sumner ME and Naidu R (eds) *Sodic Soils*, pp. 3–17. New York: Oxford University Press.

US Salinity Laboratory Staff (1954) *Diagnosis and Improvement of Saline and Alkali Soils*. Agricultural Handbook No. 60, USDA. Washington, DC: US Government Printing Office.

# SALINITY

Contents

## Management

**D Hillel**, Columbia University, New York, NY, USA

### Natural and Induced Soil Salinity

The term 'salinity' refers to the presence in soil and water of electrolytic mineral solutes in concentrations that are harmful to many agricultural crops. Except along seashores, saline soils seldom occur in humid regions, thanks to the net downward percolation of fresh water though the soil profile, brought about by the excess of rainfall compared with evapotranspiration. In arid regions, on the other hand, there may be periods of no net downward percolation and hence no effective leaching, so salts can accumulate in the soil. Hence the combined effect of meager rainfall, high evaporation, the presence of salt-bearing sediments, and – in many places, particularly river valleys and other low-lying areas – the occurrence of shallow, brackish groundwater which gives rise to saline soils.

Less obvious than the appearance of naturally saline soils, but perhaps more insidious, is the inadvertent induced salination of originally productive soils, caused by human intervention. Irrigation waters generally contain appreciable quantities of salts. (For example: Even with relatively good-quality irrigation water containing no more than 0.3 kg salts m$^{-3}$, applying 10 000 mm water per season would add as much as 3 000 kg salts ha$^{-1}$ per year!) Crop plants normally extract water from the soil while leaving most of the salt behind. Unless leached away (continuously or intermittently), such salts sooner or later begin to hinder crop growth. Where internal drainage is impeded, attempts to leach the soil can do more harm than good, by raising the water table, waterlogging the soil, and causing capillary rise to the surface, where evaporation infuses the topsoil with cumulative quantities of salt.

Overall salinity is generally expressed in terms of the total dissolved solutes (TDS) in milligrams per liter of solution (approximately equivalent to parts per million, ppm). Salinity may also be characterized by measuring the electrical conductivity (EC) of the solution, generally expressible in terms of decisiemens per meter (dS m$^{-1}$).

Quantitative criteria for diagnosing soil salinity were originally formulated by the US Salinity Laboratory at Riverside, California (in its *Handbook 60*, first issued in 1954), in terms of the EC of the soil's saturation extract (i.e., the solution extracted from a soil sample that had been presaturated with water). Those criteria are still in wide use today. Accordingly, a saline soil has been defined as having an EC of more than 4 dS m$^{-1}$. This value generally corresponds to approximately 40 mmol salt l$^{-1}$. In the case of a sodium chloride solution, this equals about 2.4 g l$^{-1}$.

The clay fraction in saline soils is generally well flocculated. As the salts are leached, however, the

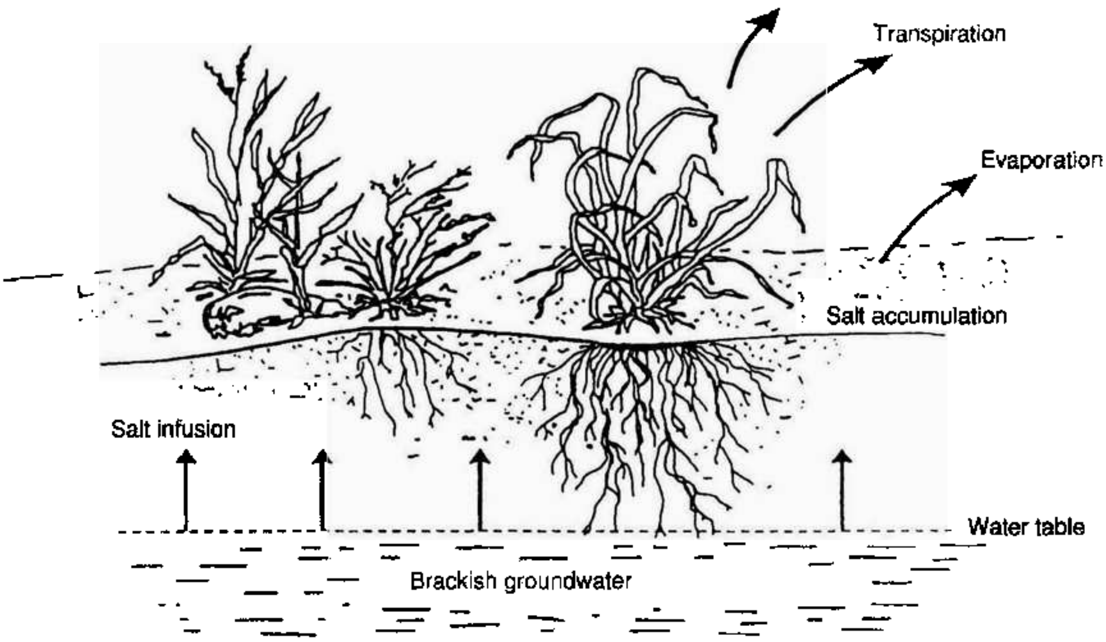


**Figure 1** The process of waterlogging and salination. Reproduced with permission from Hillel D (1998) Environmental Soil Physics, Elsevier.

flocs may tend to disperse and the soil aggregates to break down (or slake). This especially occurs where an appreciable concentration of sodium is adsorbed on to the clay particles. The tendency for flocs to disperse and for aggregates to slake and collapse results in the deterioration of soil structure by the clogging of large pores in the soil, and consequently in the reduction of soil permeability. This leads to the associated phenomenon of soil sodicity, also known as alkalinity (**Figure 1**).

## Irrigation Water Quality

Solutes can be added to the soil solution in rainfall (especially along seacoasts, where sea spray mixes with the rain) or in irrigation water infiltrated from above, in groundwater rising by capillarity from below, in the dissolution of salts initially present in solid form within the soil and subsoil, and in the artificial application of agricultural chemicals (e.g., fertilizers, pesticides, and soil amendments). Methods for removal of solutes from the soil include uptake by plants (generally of minor importance), downward transport by percolation and drainage (leaching), erosion of the soil surface by overland flow and by wind, precipitation or adsorption on to the solid phase and conversion to insoluble or unavailable forms, and – for some substances – volatilization of gaseous compounds. To avoid the accumulation of salts to toxic levels, their inputs to the soil must not exceed the rate of their removal from the soil or their conversion to unavailable forms within it. The control of soil salinity must therefore include measures to control both the inputs and the outputs of salts.

Among the various inputs of salts listed above, irrigation water often plays a major role. The quality of irrigation water may affect soil salinity and sodicity, cation exchange, soil acidity or alkalinity, nutrient availability, clay dispersion and flocculation, as well as soil structure, infiltration, and aeration. Clearly, therefore, the composition of irrigation water is an important determinant of soil productivity, crop growth, and agricultural drainage quality. The hazard posed by irrigation water containing salts of varying composition and concentration depends on soil conditions, climatic conditions, crop species and variety, and the amount and frequency of the irrigation applied. In general, irrigation water of EC less than $0.7\,\mathrm{dS\,m^{-1}}$ poses little or no danger to most crops, whereas EC values more than $3.0\,\mathrm{dS\,m^{-1}}$ may restrict the growth of most crops. The major components of salinity are the cations Ca, Mg, and Na, and the anions Cl, $SO_4$, and $HCO_3$. The potassium, nitrate, and phosphate ions, though important nutritionally, are usually minor components of soil salinity. In addition, certain constituents (such as boron) may have an important effect on crop growth even though their concentrations are usually too low to have any substantial effect on the soil's total salinity.

Irrigation waters of different sources, locations, and seasons vary greatly in quality (**Table 1**). Some irrigation waters contain as little as $50\,\mathrm{g\,salts\,m^{-3}}$, and others as much as $3000\,\mathrm{g\ salts\,m^{-3}}$. Since the volume of water applied in irrigation to a crop during its growing season commonly varies between 5000 and $20\,000\,\mathrm{m^{-3}\,ha^{-1}}$, the salt input to a crop may range between 250 and $60\,000\,\mathrm{kg\ ha^{-1}}$.

**Table 1** Classification of water quality according to total salt concentration

| *Designation* | *Total dissolved salts (ppm)* | *EC (dS m$^{-1}$)* | *Category* |
|---|---|---|---|
| Fresh water | <500 | <0.6 | Drinking and irrigation |
| Slightly brackish | 500–1 000 | 0.6–1.5 | Irrigation |
| Brackish | 1 000–2 000 | 1.5–3 | Irrigation with caution |
| Moderately saline | 2 000–5 000 | 3–8 | Primary drainage |
| Saline | 5 000–10 000 | 8–15 | Secondary drainage and saline groundwater |
| Highly saline | 10 000–35 000 | 15–45 | Very saline groundwater |
| Brine | >35 000 | >45 | Seawater |

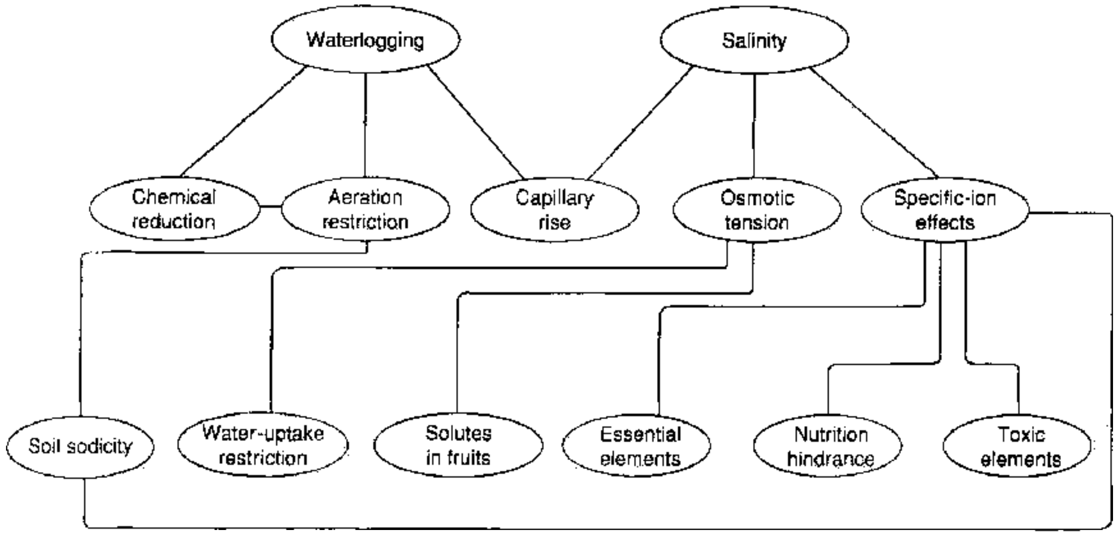


**Figure 2** Effects of salinity and sodicity on plants.

Another important criterion of irrigation water quality is the sodium adsorption ratio (SAR). High alkalinity of irrigation water, manifested when the pH value is more than 8.5, generally indicates the predominant presence of sodium ions (often associated with the bicarbonate anion) and poses a danger of soil sodification. Freshly pumped groundwater may have a high SAR even if the pH is less than 8.5, owing to the presence of dissolved $CO_2$ (samples of such water should be aerated to allow the carbon dioxide to effervesce prior to measurement of the pH).

With high-SAR water, irrigation by sprinkling will increase the soil's tendency to form a surface seal (crust) under the impact of the drops striking the bare soil. Flood irrigation may also cause the breakdown of soil aggregates by air slaking. On the other hand, application of water by drip irrigation at spaced points on the surface or below it, may lessen the physical disruption of soil structure that would otherwise take place under the influence of high-SAR water. High-pH water may cause nutritional as well as structural problems. The addition of gypsum to the water or to the soil surface may help in both respects, by promoting the flocculation of clay and the nutritional balance of the soil solution.

In certain circumstances and with appropriate precautions, brackish water may safely be used for the irrigation of salt-tolerant crops. This includes the secondary use, or reuse, of drainage waters. Especially suitable for this purpose are deep sandy soils that are prevalent in some arid regions, where internal drainage is unrestricted and there is little risk of either groundwater rise or of soil salination and sodification.

The use of brackish water for sprinkling irrigation, however, may cause foliar injury. The degree of injury depends on the following factors: concentrations of ions in the water, sensitivity of the crop at various growth stages, water stress of the plants prior to irrigation, and frequency of irrigation. The potential damage also depends on the prevailing environmental conditions, including the temperature and relative humidity of the atmosphere, which affect the rate of evaporation. Sprinkling at night, when the atmospheric temperature and evaporativity are relatively low, can reduce foliar absorption and injury.

The greater the rate of evapotranspiration and the longer the interval between irrigations, the more concentrated the residual solution in the root zone and the more severe the salt stress imposed on the plants (Figure 2). Hence, when irrigation water is brackish, a common management strategy is to increase the frequency of irrigation so as to maintain a high content of water and to prevent the rise of salt concentration in the root zone. Thanks to its ready adaptability to high-frequency irrigation, drip irrigation beneath the canopy appears to be the most appropriate method to use with brackish water, especially as it avoids direct foliar exposure to saline water.

Brackish drainage water may also be used in a system of agroforestry. Salt-tolerant trees are capable of thriving when irrigated with brackish water, and also of lowering the water table by the extraction and transpiration of water from deeper layers in the soil, thus reducing the volume and expense of needed drainage in some areas. Among the trees suitable for this type of agroforestry and biodrainage are certain species of eucalyptus, acacia, casuarina, poplar, mesquite, and tamarisk. The harvested wood may be used for fuel, pulp, or construction.

The use of drainage wastewater entails certain distinct hazards. In some places, the concentration of nitrates may become excessive, and contribute to groundwater and surface-water contamination. Heavy metals and various other toxic materials can pose a problem. They tend to accumulate in the soil and thence enter the biological food chain. Especially hazardous are industrial waste products that may be carcinogenic as well as toxic.

## Groundwater Drainage

The presence of a high water table can be either a blessing or a curse. The blessing occurs when, in periods of low rainfall or deficiency of water for irrigation, upward capillary flow from the water table augments the water supply to the roots. The curse occurs as the rising water brings up salts from below and thereby salinizes the root zone. In the field, upward capillary rise and downward percolation may take place alternately during the year. Percolation occurs typically during the rainy and early irrigation seasons, when the water requirements of the crop are relatively low and the water supply from above is high. On the other hand, upward flow takes place later in the irrigation season, when the water requirements are high and both rainfall and irrigation are restricted. Over the long term, a net downward flow of salt-bearing water through the root zone is essential to sustainable productivity.

The term 'drainage' can be used in a general sense to denote outflow of water from soil. More specifically, it can serve to describe the artificial removal of excess water, or the set of management practices designed to prevent the occurrence of excess water. The removal of free water tending to accumulate over the soil surface by appropriately shaping the land is termed 'surface drainage' and is outside the scope of this article. Finally, 'groundwater drainage' refers to the outflow or artificial removal of excess water from within the soil, generally lowering the water table or preventing its rise.

In waterlogged soils, gas exchange with the atmosphere is restricted to the surface zone, while, within the profile, oxygen may be almost totally absent and carbon dioxide may accumulate. Under anaerobic conditions, various substances are reduced chemically from their normally oxidized state. Toxic concentrations of ferrous, sulfide, and manganous ions can develop. These, in combination with products of the anaerobic decomposition of organic matter (e.g., methane) can greatly inhibit plant growth. At the same time, nitrification is prevented, and various plant and root diseases (especially fungal) are more prevalent. High-moisture conditions at or near the soil surface make the soil susceptible to compaction and puddling by animal and machinery traffic. Necessary operations (e.g., tillage, planting, spraying, and harvesting) are thwarted by poor trafficability (i.e., the ability of the ground to support vehicular traffic). Furthermore, the surface zone of a wet soil does not warm up readily at springtime, owing to greater thermal inertia and downward conduction, and to loss of latent heat by the higher evaporation rate. Consequently, germination and early seedling growth are retarded. All these phenomena are in addition to the tendency of waterlogged soils, especially in arid areas, to become saline.

The artificial drainage of groundwater is generally carried out by means of drains, which may be ditches, pipes, or 'mole channels,' into which groundwater flows as a result of the hydraulic gradients existing in the soil. The drains themselves are made to direct the water, by gravity or by pumping, to the drainage outlet, which may be a stream, a lake, an evaporation pond, or the sea. In some places, drainage water may be recycled or reused for agricultural, industrial, or residential purposes, as well as for agroforestry or the irrigation of ornamental plants. Because drainage water may contain potentially harmful concentrations of salts, fertilizer nutrients, pesticide residues, and various other toxic chemicals, as well as biological pathogens, it is not enough to provide means to 'get rid' of it; attention must be devoted to the quality of the water to be disposed of and to the long-term consequences of its disposal.

The disposal of salt-bearing effluent may pose a danger to rivers and to groundwater. If the drainage water is returned to a river or an aquifer serving as a water source, or if the drainage is to be reused directly,

**Table 2** Prevalent depths and spacing of drainage pipes in different soil types

| *Soil type* | *Hydraulic conductivity (m day$^{-1}$)* | *Spacing of drains (m)* | *Depth of drains (m)* |
|---|---|---|---|
| Clay | 1.5 | 10–20 | 1–1.5 |
| Clay loam | 1.5–5 | 15–25 | 1–1.5 |
| Loam | 5–20 | 20–35 | 1–1.5 |
| Fine, sandy loam | 20–65 | 30–40 | 1–1.5 |
| Sandy loam | 65–125 | 30–70 | 1–2 |
| Peat | 125–250 | 30–100 | 1–2 |

Reproduced with permission from Hillel D (1998) Environmental Soil Physics. Elsevier.

its load of corrosive salts and other pollutants may affect downstream agriculture, households, water utilities, and industry. Where persistent pesticides are applied to irrigated land, their residues can cause further damage to biotic communities in riverine, estuarine, and lacustrine habitats, as well as to public health. Construction of evaporation ponds for disposal of drainage water generally requires the allocation of approximately 10% of the land area. Care must also be taken to avoid leaks and consequent salination of the underlying groundwater. The loss of potentially valuable land, the costs of construction and conveyance, and the environmental impacts may make such ponds impractical in some cases.

Various equations, empirically or theoretically based, have been proposed for determining the desirable depths and spacings of drain pipes or ditches in different soil and groundwater conditions. In the Netherlands, the country with the most experience in drainage, common criteria for drainage are to provide for the removal of approximately 7 mm per day, and to prevent a water table from rising above a depth of 0.5 m from the soil surface. In more arid regions, because of the greater evaporation rate and groundwater salinity, the water table must generally be kept much deeper. In the Imperial Valley of California, for instance, the drain depth ranges from approximately 1.5 to 3 m, and the desired water table depth midway between drains is at least 1.2 m. For medium- and fine-textured soils, the depth should be even greater where the salinity risk is high. Since there is a practical limit to the depth of drain placement, it is the density of drain spacing that must be increased under such circumstances (**Table 2**).

## Leaching Processes

To prevent salts from accumulating in the root zone during repeated cycles of irrigation and evapotranspiration, the obvious remedy is to apply water in an amount greater than evapotranspiration, so as deliberately to cause a fraction of the applied water to flow through the root zone and flush away the excess salts. However, unless the water table is very deep or lateral groundwater drainage is sufficiently rapid, the extra irrigation can cause a progressive rise of the water table. Therefore, the amount of water applied must be optimized to allow leaching without water-table rise. The concept of 'leaching requirement' was first developed by the US Salinity Laboratory in Riverside, California. It has been defined as the fraction of the irrigation water that must be percolated out of the bottom of the root zone in order to prevent average soil salinity from rising above some specifiable level.

According to the standards developed there, the maximum concentration of the soil solution in the root zone, expressed in terms of EC, should be kept below 4 dS m$^{-1}$ for sensitive crops. Salt-tolerant crops such as beets, alfalfa, and cotton may give satisfactory yields at values up to 8 dS m$^{-1}$. The problem encountered in any attempt to apply such a simplistic criterion is that in the field (unlike the case of plants grown in solution culture or in small containers) the concentration of the soil solution varies greatly in space and time. In addition, the sensitivity of any crop to salinity depends on its stage of growth and on such variables as ambient temperature, atmospheric humidity, soil matric suction, and nutrient availability.

The leaching requirement depends on the salt concentration and composition of the irrigation water, on the amount of water extracted from the soil by the crop, and on the salt tolerance of the crop, which determines the maximum allowable concentration of the soil solution in the root zone. Assuming steady-state conditions of throughflow (thus disregarding short-term fluctuations in soil-moisture content, flux, and salinity), and furthermore assuming no appreciable dissolution or precipitation of salts in the soil and no significant removal of salts by the crop or capillary rise of salt-bearing water from below, the following simple equation is obtained:

$$V_d/V_i = c_i/c_d \qquad [1]$$

in which $V_d$ and $V_i$ are the volumes of drainage and of irrigation, respectively, and $c_d$ and $c_i$ are the corresponding concentrations of salt. Water volumes are normally expressed per unit area of land as equivalent depths of water, and salt concentrations are generally measured and reported in terms of EC. Because the volume of water drained is the difference between the volumes of irrigation and evapotranspiration (i.e., $V_d = V_i - V_{et}$), we can transform the last equation as follows:

$$V_i = [c_d/(c_d - c_i)]V_{et} \qquad [2]$$

This is equivalent to the formulation given in the US Salinity Laboratory's Department of Agriculture (USDA) Handbook No. 60:

$$d_i = [E_d/(E_d - E_i)]/d_{et} \qquad [3]$$

where $d_i$ is the depth of irrigation, $d_{et}$ the equivalent depth of 'consumptive use' by the crop (evapotranspiration), and $E_d$ and $E_i$ are the electrical conductivities of the drainage and irrigation waters, respectively.

The leaching requirement equation implies that, by varying the fraction of applied water percolated through the root zone, it is possible to control the concentration of salt in the drainage water and hence to maintain the concentration of the soil solution in the main part of the root zone at some intermediate level between $c_i$ and $c_d$. However, the leaching requirement concept disregards the distribution of salts within the root zone itself, as it is affected by the frequency and spatial variability of irrigation, as well as by its quantity and water quality. In particular, the variation of root-zone salinity is affected by the pattern and degree of soil moisture depletion between irrigations. The less frequent the irrigation regime, the greater the buildup of salt between successive irrigations. In some cases, the commonly recommended leaching fraction may not be sufficient to prevent the reduction of yield below its potential, especially if the climatically imposed evaporation rate is high and the irrigation water is brackish.

With modern methods of high-frequency irrigation, it is possible to maintain the soil solution in the surface zone at a concentration nearly equal to that of the irrigation water. This zone can be deepened by increasing the volume of water applied. Beyond this zone, the salt concentration of the soil solution increases with depth to a salinity level depending on the leaching fraction. High-frequency irrigation not only lowers the concentration of the soil solution in the upper zone (where most roots proliferate), but also tends to minimize the matric suction of soil moisture.

Extensive research has shown that leaching soils at a water content below saturation (e.g., under low-intensity sprinkling or intermittent ponding) can produce more efficient leaching that can be achieved by the once-standard method of continuous flooding. In a soil with macropores – cracks, wormholes, or decayed root channels – much of the water under ponding moves rapidly down those large passageways, bypassing the greater volume of the soil containing the salt, so it is largely ineffective in leaching the micropores of the soil matrix. In contrast, under low-intensity sprinkling, the soil never becomes saturated, so a greater portion of the applied water moves through the soil matrix, thus producing more efficient leaching per unit volume of water infiltrated. However, the processes of infiltration and unsaturated flow under low-intensity sprinkling are inherently slower and require more time than saturated infiltration under ponding.

Nonuniformity of irrigation as well as of soil is a complicating factor. If the leaching requirement is not met throughout the field, soil salinity will prevail in spots, wherever leaching is insufficient. Whether to apply copious amounts of water to the entire field so as to ensure that the leaching requirement is met everywhere, or to accept some reduction in yield in parts of the field, must be determined from an economic analysis of costs and benefits. Such an analysis should take into consideration the danger that saline spots might recur (and perhaps even grow in extent and severity) from year to year. One answer is to apply extra water preferentially to the spots that need it most, but such a strategy requires a flexible irrigation system that would allow controlled variability of water delivery. Although such a specialized irrigation system is likely to be expensive to install and operate, it may well be worthwhile in the long run.

## Soil Amendments

The leaching process is enhanced if the applied water contains a sufficient concentration of electrolytes to reduce swelling and dispersion of clay in the soil. Where leaching occurs with water of very low salinity (e.g., rainwater), soil permeability can be increased by the surface application of a slowly soluble electrolytic salt – preferably a substance containing a divalent cation such as calcium, to prevent the sodium ion from dispersing the clay. Such materials, commonly known as soil amendments, can replace exchangeable sodium with flocculation-promoting divalent ions.

The most commonly considered soil amendment for the purpose of improving the structure and the permeability of soils, especially those prone to becoming sodic, are dihydrate calcium sulfate (gypsum) and dihydrate calcium chloride. Gypsum is generally the preferred soil amendment, thanks to its ready availability in many places and to its relatively low cost. It may be derived from mining or be available as a by-product of the phosphate-fertilizer industry. The solubility of pure gypsum is approximately 2.15–2.63 kg m$^{-3}$, depending somewhat on temperature. Applied gypsum dissolves in the soil solution until its solubility limit is reached or until its supply is exhausted. The rate of dissolution of applied gypsum depends on its source and degree of granulation. Industrial gypsum generally dissolves more rapidly than mined gypsum, which generally contains impurities. The amount of gypsum needed to replace the exchangeable sodium depends on the initial exchangeable sodium percentage, the soil's total cation exchange capacity, its bulk density, and the depth of the soil to be treated effectively. If elemental sulfur is added to a sodic soil in lieu of gypsum, it must be oxidized *in situ* to become effective. Upon oxidation, it forms sulfuric acid, which then reacts with lime in the soil to produce gypsum.

## Early-Warning Systems

In many cases, irrigation systems are organized and irrigation is begun long before drainage is installed. Indeed, an irrigation project can often function unimpeded for years, even decades, without artificial drainage. In some cases, the land is so well drained naturally that irrigation can thus be continued for a very long time. However, far more typically, the processes of groundwater rise and salt accumulation proceed inexorably, so that sooner or later (especially in ill-drained river valleys, where most irrigation development takes place) the provision of artificial drainage becomes essential.

Granted that a drainage system must be planned in advance at the very outset of an irrigation project, the crucial question is when to begin installing and operating it. If installed too early, the drainage system may lie unused for some time and therefore be both unnecessary and uneconomical, and it may deteriorate in the interim before it comes into use. On the other hand, if installed after waterlogging and salination have advanced, it may be too late to restore productivity economically. These considerations emphasize the importance of having an early-warning system to indicate, before the problem becomes acute, that soil salination is incipient and that the need for drainage is imminent. Soil salinity is normally monitored by a combination of soil sampling, soil-solution sampling by vacuum extraction, and *in situ* salt sensors. Mobile devices with combined electromagnetic induction and four-electrode soil conductivity sensing systems are now under development for monitoring and mapping the distribution of soil salinity over an entire field, but such instruments have not yet entered into general use.

The detection and diagnosis of salinity are difficult in the early stages of its occurrence. Visual inspection of the soil surface may be misleading. For example, the white precipitate formed on the surface of furrow-irrigated or drip-irrigated soil may be due to relatively harmless calcite or gypsum. Visual inspection of crops provides obvious clues to salt stress only after the condition is well advanced. Crop plants suffering salt stress eventually exhibit stunted growth, smaller leaves than normal, and discoloration. Such symptoms first occur in spots, rather than uniformly over the field. Since factors other than salinity (e.g., water stress, nutrient deficiencies, or misapplied pesticides) may produce similar symptoms, visually observed indications of apparent salinity should be checked by chemical analyses of soil, plant, and water samples. One good way to detect the early appearance of salinity is to place plants that are known to be particularly salt-sensitive at regularly spaced intervals throughout the irrigated area. Such interspersed detector plants may reveal early symptoms of physiological stress and thus provide a timely warning of problems that are likely to be exacerbated over time.

An important indicator of the salinity hazard is the depth of the water table. There is cause for concern whenever it approaches within 1.5–2 m of the soil surface. Hence the water table should be monitored regularly by means of observation wells and piezometers (**Figures 3** and **4**). Both are vertical tubes that

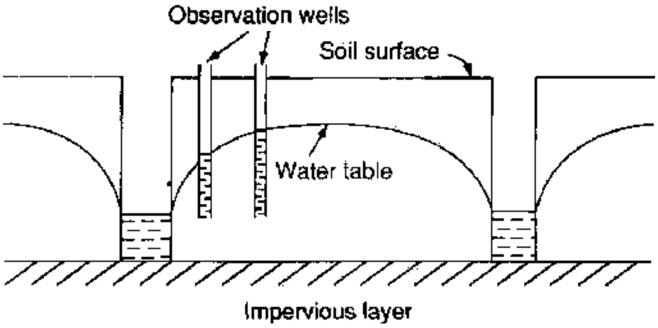


**Figure 3** Observation wells to determine elevation of the water table. Reproduced with permission from Hillel D (1998) Environmental Soil Physics, Elsevier.

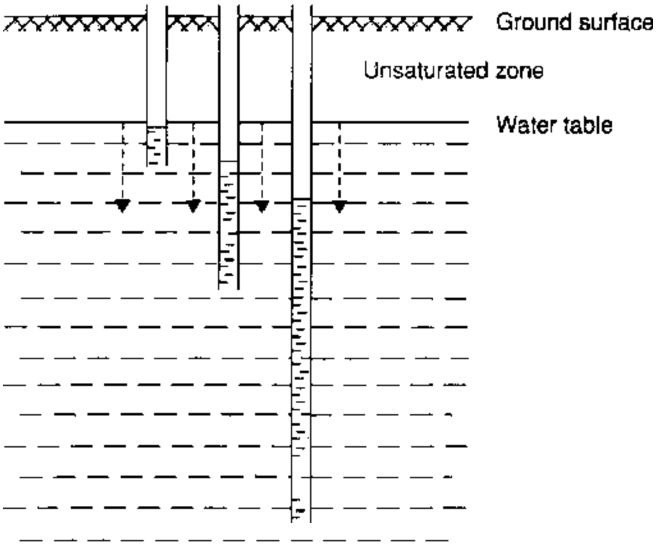


**Figure 4** Piezometers to determine vertical pressure gradients below the water table. Reproduced with permission from Hillel D (1998) Environmental Soil Physics, Elsevier.

are inserted into the soil to a depth far below the water table. The difference is that an observation well is perforated to permit free inflow of groundwater along the length of the tube below the water table; in contrast, a piezometer's only opening is at the bottom. Hence a piezometer indicates the hydraulic head (pressure) of the water at the bottom of the tube, rather than the position of the water table. Several piezometers, inserted side by side to different depths, can indicate the vertical gradient of the hydraulic head below the water table. The direction and magnitude of that gradient are indicative of the tendency of the groundwater to rise or fall.

*See also:* **Solute Transport**

## Further Reading

Bresler E, McNeal BL, and Carter DL (1982) *Saline and Sodic Soils: Principles–Dynamics–Modeling.* Berlin, Germany: Springer-Verlag.

Dinar A and Zilberman D (eds) (1991) *The Economics and Management of Water and Drainage in Agriculture.* Boston, MA: Kluwer Academic.

Hanks RJ (1984) Predictions of crop yield and water consumption under saline conditions. In: *Soil Salinity Under Irrigation.* Berlin, Germany: Springer-Verlag.

Hillel D (2000) *Salinity Management for Sustainable Irrigation: Integrating Science, Environment, and Economics.* Washington, DC: The World Bank.

Hoffman GJ (1990) Leaching fraction and root zone salinity control. In: *Agricultural Salinity Assessment and Management.* New York: American Society of Civil Engineers.

Keren R and Miyamoto S (1990) Reclamation of saline, sodic, and boron-affected soils. In: *Soil Salinity Under Irrigation.* Berlin, Germany: Springer-Verlag.

Lauchli, A and Epstein E (1990). Plant responses to saline and sodic conditions. In: *Agricultural Salinity Assessment and Management,* pp. 113–137. New York: American Society of Civil Engineers.

Massoud FI (1981) *Salt-Affected Soils at a Global Scale and Concepts for Control.* Rome, Italy: FAO Land and Water Development Division.

Oster JD and Frenkel H (1980) The chemistry of the reclamation of sodic soils with gypsum and lime. *Soil Science Society of America Journal* 44: 41–45.

Rhoades JD, Chanduvi F, and Lesch SM (1999) *Soil Salinity Assessment Methods and Interpretation of Electrical conductivity Measurements.* FAO Irrigation and Drainage Paper 57. Rome, Italy: FAO Land and Water Development Division.

Richards LA (ed.) (1954) *Diagnosis and Improvement of Saline and Alkali Soils.* USDA Handbook No. 60. Washington, DC: US Government Printing Office.

Shainberg I and Shalhevet J (eds) (1984) *Soil Salinity Under Irrigation.* Berlin, Germany: Springer-Verlag.

Tanji KK (ed.) (1990) *Agricultural Salinity Assessment and Management.* New York: American Society of Civil Engineers.

Yaron D (ed.) (1981) *Salinity in Irrigation and Water Resources.* New York: Marcel Dekker.

# Physical Effects

**D Russo**, The Volcani Center, Bet Dagan, Israel

## Introduction

Environmental control and agricultural management problems involve the simultaneous transport of water and mixed salts solutions with ions that interact with the soil matrix. The soil solution–soil matrix interactions include phenomena such as cation exchange, anion exclusion, precipitation and dissolution, swelling and reorganization of the soil colloid structures, and, concurrently, rearrangement of the soil pore-size distribution; the latter can affect the soil hydraulic conductivity and retentivity. These interactions, therefore, may considerably affect water flow and solute transport.

The magnitude of the soil solution–soil matrix interactions depends to a large extent upon the types and amounts of inorganic and organic soil colloids. Among soil mineral colloids, the most reactive constituents are the smectite minerals (e.g., montmorillonite, beidellite, monronite), characterized by dioctahedral structure. The presence of these minerals imparts a considerable cation exchange capacity and/or specific surface area to soils from arid and semiarid regions. Furthermore, the smectite minerals are capable of considerable expansion at moderate-to-high exchangeable sodium levels in the presence of relatively low salt soil-solution concentration. Hence, they can impart rather substantial salinity-dependent water retention and conductivity changes to soils, and, concurrently, may considerably affect water flow and solute transport.

Here the emphasis is on the effect of salinity on soil physical properties relevant to water flow and solute transport, in particular, soil water retentivity and conductivity.

## Water Retention and Swelling in Salt-Affected Soils

The amount of water retained by the soil depends largely upon the mineral and organic colloid contents of the soil, although the structural arrangement of soil pores may also be of considerable importance in this respect. In particular, the presence of substantial quantities of soil organic matter and/or smectite minerals results in a considerable water-retention capacity. Soils with high smectite content may also swell considerably more in the presence of high sodium and/or low salt concentrations. Soils high in organic colloids are much more stable and do not tend to swell excessively under adverse chemical conditions. They may disperse, however, at high pH levels or in the presence of low salt concentrations.

The effects of concentration and composition of the soil solution on water retention by a soil containing 20% clay (dominated by montmorillonite) are depicted in **Figure 1**. The functional relationships (under wetting conditions) are given in terms of the capillary pressure head, *h* (considered here as a positive quantity), volumetric soil-water content, $\theta$, and equilibrium solution concentration, $C = C_{\mathrm{Na}} + C_{\mathrm{Ca}}$, where $C$ is concentration in milliequivalents per liter, and composition (expressed in terms of sodium adsorption ratio: $\mathrm{SAR} = C_{\mathrm{Na}}/\surd(C_{\mathrm{Ca}}/2)$. The points in **Figure 1** represent actual measurements for various values of SAR and C; the solid lines represent best visual fits between measured points. Note that the ordinate scale of **Figure 1** is shifted upward by an order of magnitude and the data are accordingly translated along this axis, to distinguish between different sets of concentrations. In general, each of the water-retention relationships in **Figure 1** can be represented by a single continuous curve. When solution concentration exceeds $50\,\mathrm{mEq\,l^{-1}}$ a single-valued function is obtained for any value of SAR. The same is true for $\mathrm{SAR} = 0$ and $2\,\mathrm{mEq\,l^{-1}} < \mathrm{C} < 5\,\mathrm{mEq\,l^{-1}}$, as can be seen from the lowest curve of **Figure 1**. Within the range $2\,\mathrm{mEq\,l^{-1}} < \mathrm{C} < 10\ \mathrm{mEq\,l^{-1}}$ and $0 < \mathrm{SAR} < 50$, water content $\theta$ depends not only on the pressure head, *h*, but also on both concentration and composition of the soil solution. For a given *h*, $\theta$ increases as C decreases and as SAR increases.

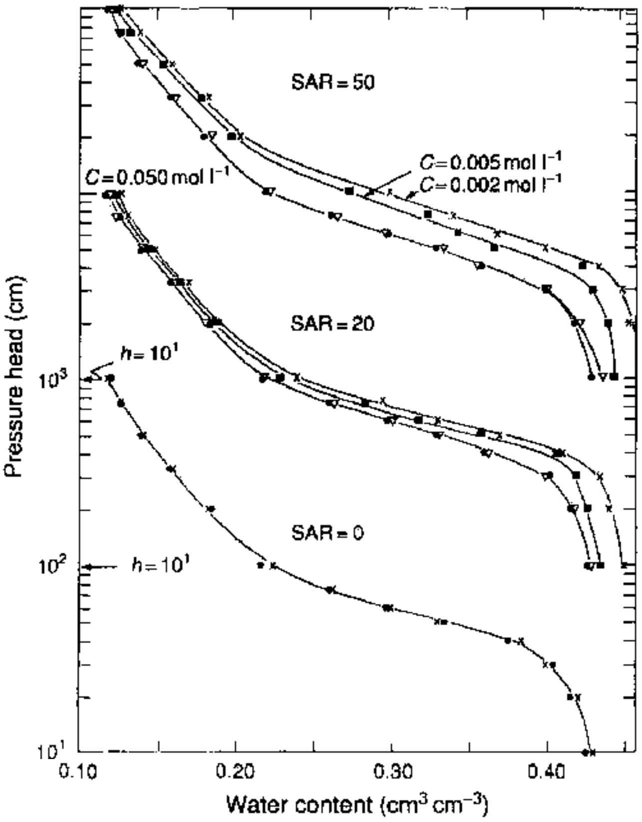


**Figure 1** Capillary pressure head, *h*, as a function of water content, $\theta$, for selected values of the concentration, *C*, and the sodium adsorption ratio, SAR, of the equilibrium solution. Arrows indicate that the point $h = 10$ cm is shifted and the data are translated along the *h*-axis. (Reproduced with permission from Russo D and Bresler E (1977) Effect of mixed Na-Ca solutions on the hydraulic properties of unsaturated soils. *Soil Science Society of America Journal* 41: 714–717.)

The experimental findings presented in **Figure 1** can be qualitatively explained in terms of electrostatic properties and the porous nature of the soil. It is assumed that the soil pore space can be divided into structural and textural porosity. The latter is regarded as an intrinsic property of the soil, arising from the size and arrangement of basic particles within the soil mass. In these pores the $h(\theta)$ function can be determined quantitatively from the pressure that develops between the clay platelets, within the clay mass between the pores as the soil solution composition and concentration vary.

The classic view of the osmotic pressure of soil colloids is that cations in the diffuse layer between clay particles reduce the free energy of the water relative to that in the external equilibrium solution. This causes water molecules to diffuse into the space between the clay particles, since the cations are electrostatically constrained from leaving the interlayer region. Hence, the particles move apart to accommodate the extra water. If, on the other hand, the clay particles are confined, hydraulic pressure develops between them. For an Na/Ca-Cl soil-water system, this pressure may be determined on the basis of the mixed-ion, diffuse double-layer theory (see eqn [2]).

Equation [2] can be regarded as a reasonably good first-order approximation of the water-retention forces acting within the textural porosity. It predicts that for a given hydrostatic pressure, *P*, the volume of water retained by the clay mass (which is proportional to the distance between the clay particles) increases as the equilibrium solution becomes more diluted (C decreases) and as the value of SAR increases. In the structural pores, which are sensitive to external forces arising from interactions between clay particles within the textural porosity, the $h(\theta)$ function can be determined from the pressure within the water-filled pores relative to that within the air-filled pores. In the former pores, at equilibrium, the pressure increases as the pore radius decreases or, for a given pressure, higher values of $\theta$ are associated with a larger number of pores having a given corresponding radius. Thus, the functional relationships between $h$ and $\theta$ in the structural porosity depend on the total porosity and the pore-size distribution. Changes in the total structural porosity and its distribution result from changes in the volume occupied by the clay mass within the textural porosity. For a given pressure, the space between the clay particles increases as the value of SAR increases and as C decreases. This results in an increase in the amount of water retained by the clay mass, and, concurrently, in the volume that is occupied by the clay mass. For a constant-volume system, changes in the volume of the clay mass are at the expense of the structural porosity and its distribution. Thus, as can be observed in **Figure 1**, the amount of water retained by the soil for a given capillary pressure head increases with decreasing C and increasing SAR.

## Effects of Salts on Soil Hydraulic Conductivity

The retention curves in **Figure 1** may be viewed as representatives of the soil's pore-size distribution. The data in **Figure 1**, therefore, suggest that swelling of the clay fraction within the soil, induced by relatively low C and high SAR, may alter the geometry of the soil pores and thus affect the intrinsic soil permeability, $\kappa(\theta)$. Investigations have confirmed that soil permeability or hydraulic conductivity, $K(\theta) = \kappa(\theta)\gamma/\eta$ (where $\gamma$ and $\eta$ are the specific gravity and the dynamic viscosity of water, respectively), is highly affected by the soil-solution concentration, C, and composition, expressed in terms of SAR. It has been suggested that the swelling of clay particles in a confined soil system causes the size of large soil pores to decrease, and that breakdown of domains of clay platelets (tactoids) and the subsequent movement and dispersion of the clay platelets may further block soil pores. Low permeability can result from such geometric restrictions. There is a positive correlation between relative soil permeability and the swelling of extracted soil clays, which, in turn, is affected by the soil-solution concentration and composition. Inasmuch as soil-solution composition may affect the breakdown of tactoids (and the subsequent movement and dispersion of clay platelets) on the one hand, and soil-swelling on the other hand, in much the same way, the relative importance of swelling and dispersion in affecting soil permeability remains essentially unresolved. There are situations (relatively coarse-textured soils associated with relatively high influx rates) in which the demixing of adsorbed ions, which concentrates Na on external particle surfaces, may lead to the breakdown of tactoids and the subsequent movement and dispersion of clay platelets.

The effect of concentration and composition of the soil solution on the hydraulic conductivity function $K(\theta)$ in a soil containing 20% clay (dominated by montmorillonite) is depicted in **Figure 2**. The functional relationships are given in terms of water

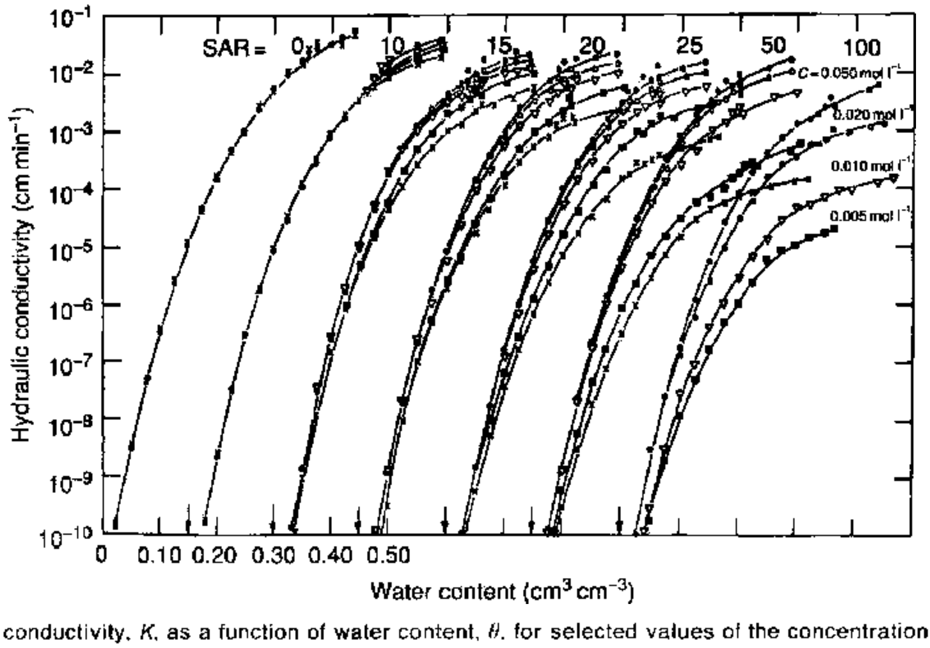


**Figure 2** Hydraulic conductivity, $K$, as a function of water content, $\theta$, for selected values of the concentration, $C$, and the sodium adsorption ratio, SAR, of the equilibrium solution. Arrows indicate that the point $\theta = 0$ is shifted and the data are translated along the $\theta$-axis. (Reproduced with permission from Russo D and Bresler E (1977) Effect of mixed Na-Ca solutions on the hydraulic properties of unsaturated soils. *Soil Science Society of America Journal* 41: 714–717.)

content, $\theta$, and equilibrium solution concentration, $C$, and composition, SAR. The points in **Figure 2** represent actual measurements for various values of SAR and $C$; the solid lines represent best visual fits between measured points. Note that the abscissa scale in **Figure 2** has been shifted to the right and the data are accordingly translated along the $\theta$-axis to distinguish among different sets of concentrations.

In general, each of the $K(\theta)$ functions in **Figure 2** can be represented by a single continuous curve. Furthermore, for the calcium-saturated system (SAR $= 0$ and $2\,\text{mEq}\,l^{-1} \leq C \leq 50\,\text{mEq}\,l^{-1}$), a single-valued $K(\theta)$ is obtained. In the mixed Na-Ca system, however, values of $K(\theta)$ decrease with both $\theta$ and $C$ for a given SAR. The differences among the various functions increase as $\theta$ becomes larger and as $C$ becomes smaller. These differences are pronounced when SAR and $\theta$ are greater than 15 and 0.25, respectively, and when $C$ is equal to or less than $10\,\text{mEq}\,l^{-1}$. In other words, relatively low values of soil-water content compensate for the negative effects of high SAR and low $C$ on the hydraulic conductivity.

The experimental data depicted in **Figure 2** can be explained qualitatively in terms of soil water–clay interactions and the porous nature of the soil. The unsaturated hydraulic conductivity function $K(\theta)$ depends upon the size distribution of water-filled pores and the total water-filled porosity. As mentioned previously, the diffuse double-layer theory for mixed-electrolyte systems predicts that, for a given pressure, the space between clay particles and, concurrently, the amount of water retained by the clay mass increase as SAR increases and $C$ decreases. For a constant-volume system, changes in the volume of the clay mass are at the expense of the quantity and distribution of the soil pores. Thus, for a given pressure head, $h$, the amount of water retained by the soil increases (**Figure 1**) and the hydraulic conductivity decreases (**Figure 2**) as SAR becomes larger and solution concentration, $C$, decreases. Since the swelling of clay decreases as $h$ increases, the increase in the amount of water retained and the decrease in hydraulic conductivity become smaller as $h$ increases. In addition, both water content and hydraulic conductivity, and, concurrently, water velocity, may considerably decrease with increasing $h$. Consequently, increasing $h$ may reduce the possibility that movement and dispersion of clay platelets will lead to additional blocking of the pore space available for water flow.

## Modeling Hydraulic Conductivity and Water Retentivity in Salt-Affected Soils

In the preceding sections it was shown that in salt-affected soils, both the hydraulic conductivity, $K(h)$, and the water retentivity, $\theta(h)$ may depend on soil-solution concentration and composition. In the present analysis, ion precipitation or dissolution is not taken into account; thus, the simple but credible soil-water system containing only Na, Ca, and Cl ions is considered. The equilibrium solution is quantified in terms of the sum of the molar concentrations of the cations, $C_0 = c_{\text{Na}} + c_{\text{Ca}}$, and the ratio between the molar concentration of mono- and divalent cations, $R = c_{\text{Na}}/c_{\text{Ca}}$.

For the modeling of the hydraulic conductivity, $K(h, R, C_0)$, and the water retentivity, $\theta(h, R, C_0)$, functions, it is further assumed: (1) that the soil clay fraction is mostly montmorillonite (having a constant effective surface charge density, $\Gamma_s$), which may exist in packets of platelets or domains called tactoids; (2) that the ions on the external surfaces of the tactoids (the 'exchange phase') are distributed according to the planar mixed-ion diffuse double-layer theory; (3) that the number of clay platelets per tactoid depends on the pressure acting on the system and on the exchangeable sodium percentage (ESP); and (4) that, due to the continuity of the pressure in the soil solution, the pressure difference between two adjacent clay particles and the bulk solution (given by the swelling pressure, $P$ (eqn [2]) is equal to the pressure difference across the air–water interface within the soil pores, i.e., the difference between the pore-water pressure, $P_{\text{w}}$, and the atmospheric pressure, $P_{\text{a}}$, as given by the capillary pressure, $P_{\text{c}} = P_{\text{w}} - P_{\text{a}} = -\gamma h$.

### Calculations of the *b*(*h*, *R*, $C_0$) and *N*(*h*, *R*, $C_0$) Functions

For the Na/Ca-Cl soil system considered here, the general equation of the interacting double layer is obtained from the first integration of the one-dimensional Poisson–Boltzmann equation given by:

$$\begin{aligned}\frac{\mathrm{d}Y}{\mathrm{d}x} = \frac{G}{\sqrt{R+1}}&[R(\exp^{Z_{\text{Na}}Y} - \exp^{Z_{\text{Na}}Y_{\text{d}}}) \\ &+ (\exp^{Z_{\text{Ca}}Y} - \exp^{Z_{\text{Ca}}Y_{\text{d}}}) \\ &+ (R+2)(\exp^{Z_{\text{Cl}}Y} - \exp^{Z_{\text{Cl}}Y_{\text{d}}})]^{1/2} \qquad [1]\end{aligned}$$

where $Y = \epsilon\psi/kT$ is the scaled electrical potential, $Z$ is the valence of the ion shown subscript, $\psi$ is the electrical potential, $\epsilon$ is the charge on the electron, $k$ is the Boltzmann constant, $T$ is the absolute temperature, $x$ is the distance to the charged surface (where $x = s$ and $Y = Y_s$), $Y_{\text{d}}$ is $Y$ at the midplane between two adjacent tactoids (at the plane $x = d$, where $\mathrm{d}Y/\mathrm{d}x = 0$), $G = (8\pi C_0 L\epsilon^2/10^3 DkT)^{1/2}$, $L$ is the Avogadro number, and $D$ is the dielectric constant of the bulk solution.

For dilute solutions, the swelling pressure, $P$, i.e., the pressure difference between the equilibrium solution (where $Y = 0$) and the midplane between two adjacent tactoids (at the plane $x = d$, where $Y = Y_d$), is obtained by substituting the Boltzmann distribution law into the Van't Hoff relation. For the Na/Ca-Cl soil system considered here, with tactoids that are at a distance of $2b$ apart (where $b = d - s$), the swelling pressure, $P$ is given by:

$$P = B\left[R\left(e^{Z_{Na}Y_d} - 1\right) + \left(e^{Z_{Ca}Y_d} - 1\right) + (R+2)\left(e^{Z_{Cl}Y_d} - 1\right)\right] \qquad [2]$$

where $B = kLTC_0/10^3(R+1)$. Substituting $P = \gamma h$ in eqn [2], one can obtain, after rearrangement:

$$a_1 e^{Z_{Na}Y_d} + a_2 e^{Z_{Ca}Y_d} + a_3 e^{Z_{Cl}Y_d} + a_4 = 0 \qquad [3]$$

where $a_1 = R$, $a_2 = 1$, $a_3 = R + 2$, and $a_4 = -[a_1 + a_2 + a_3 + 10^3\gamma h(R+1)kLTC_0]$.

Thus, the solution of the nonlinear eqn [3] yields $Y_d$ as a function of $h$, $R$, and $C_0$. Estimation of $b(h, R, C_0)$ also requires the knowledge of the scaled electrical potential at the charged surface, $Y_s$, which can be estimated from the integration of eqn [1] and from the condition of electroneutrality requiring that the surface charge density, $\Gamma_s$, be equal to the total space charge density, $\delta(x)$, within the range $s \le x \le d$, thus:

$$\left(\frac{dY}{dx}\right)_{x=s} = \frac{4\pi\pi\Gamma_s}{DkT} \qquad [4]$$

Evaluating eqn [1] at the plane $x = s$ (where $Y = Y_s$), using eqn [4], one obtains, after rearrangement:

$$a_1 e^{Z_{Na}Y_s} + a_2 e^{Z_{Ca}Y_s} + a_3 e^{Z_{Cl}Y_s} + a_4 = 0 \qquad [5]$$

where $a_4 = -\{[2\Pi 10^3(R+1)\Gamma_s^2/LDkTC_0] + a_1 e^{Z_{Na}Y_d} + a_2 e^{Z_{Ca}Y_d} + a_3 e^{Z_{Cl}Y_d}\}$, and $a_i$, $i = 1, 2, 3$, as before.

The solution of the nonlinear eqn [5] yields $Y_s$ as a function of $\Gamma_s$, $R$, $C_0$, and $Y_d(h, R, C_0)$. From the values of $Y_d$ and $Y_s$, the distance from the midplane $x = d$ (where $dY/dx = 0$ and $Y = Y_d$) to the charged surface (where $x = s$ and $Y = Y_s$), $b = d - s$, can be obtained by direct integration of eqn [1]. Rearranging eqn [1] yields:

$$\frac{G}{\sqrt{R+1}}\int_d^s dx = \int_{Y_d}^{Y_s}\left(a_1 e^{Z_{Na}Y} + a_2 e^{Z_{Ca}Y} + a_3 e^{Z_{Cl}Y} + a_4\right)^{-1/2} dY \qquad [6]$$

where $a_4 = -[a_1 e^{Z_{Na}Y_d} + a_2 e^{Z_{Ca}Y_d} + a_3 e^{Z_{Cl}Y_d}]$ and $a_i$, $i = 1, 2, 3$, as before. Thus a numerical integration of eqn [6] yields $b = d - s$ as a function of $h$, $R$, and $C_0$.

Once $Y_d(h, R, C_0)$ has been estimated from eqn [3], the fractional excess of Na on the exchange phase, $\Gamma_{Na}/\Gamma_s$ can be estimated from:

$$\frac{\Gamma_{Na}}{\Gamma_s} = \frac{R\sqrt{C_0}}{\Gamma_s\sqrt{\alpha(R+1)}} \sin h^{-1}\left(\frac{\Gamma_s}{\sqrt{C_0}}\frac{\sqrt{\alpha(R+1)}}{(R + 4\cosh Y_d)}\right) \qquad [7]$$

where $\Gamma_{Na}$ is the quantity of sodium ionic charges per unit area of the charged surface, and $\alpha = 8000\Pi/L\epsilon\kappa T$. Note that, inasmuch as $Y_d = Y_d(h, R, C_0)$, the exchangeable sodium percentage, $ESP = 100\Gamma_{Na}/\Gamma_s$ is also a function of $h$, $R$, and $C_0$. For given pressure head, $h$, and $ESP(h, R, C_0)$, the number of clay platelets per tactoid, $N = N[h, ESP(h, R, C_0)]$, is calculated by using experimental data on the dependence of $N$ on $h$ in a calcium system, and by experimental data on the dependence of $N$ on ESP in clay suspensions.

### Calculations of the $\theta(h, R, C_0)$ and $K(h, R, C_0)$ Functions

To estimate the $K(h, R, C_0)$ and the $\theta(h, R, C_0)$ functions, a confined volume element, $V$, of a homogeneous and isotropic porous medium with uniform porosity, $\varphi^0$, is considered. This soil volume, which is large compared with the typical pore size of the porous medium, is subdivided into $n$ equal subvolumes, $v_i$, $i = 1$–$n$, in such a way that only one value of the average pressure head, $h_i$, the equivalent radius of the water-filled pores, $r_i^0 = 2\sigma/\gamma h_i$, (with $\sigma$ being the surface tension at the air–water interface), and the average number of pores having a unit length, $l_i = \varphi^0 v_i/\Pi(r_i^0)^2$, can be defined in each subvolume, $v_i$. Using the functions $b(h_i, R, C_0)$ and $N(h_i, R, C_0)$, the water content in each $v_i$ is calculated from:

$$\theta_i(h_i, R, C_0) = \theta_i^0(h_i) + [\nu_i(h_i, R, C_0) - \nu_i^0(h_i)] \qquad [8]$$

where $\theta_i^0$ is the water content in a reference state (that is, a soil in equilibrium with a concentrated calcium chloride solution), at pressure head $h_i$, and $\nu_i^0$ and $\nu_i$ are the volumes of water retained by the clay particles in a unit volume of the soil, in the reference state, and in equilibrium with a soil solution of a given $R$ and $C_0$, given by:

$$\nu_i^0(h_i) = \frac{A_s}{2\,N_i^0(h_i)}\left\{2b_i^0(h_i) + 2b_0[N_i^0(h_i) - 1]\right\}\rho_b \qquad [9]$$

and

$$\nu_i(h_i, R, C_0) = \frac{A_s}{2\,N_i^0(h_i)}\left\{2b_i(h_i, R, C_0)[1 + N_i^0(h_i) - N_i(h_i, R, C_0)] + 2b_0[N_i(h_i, R, C_0) - 1]\right\}\rho_b \qquad [10]$$

respectively, where $A_s$ and $\rho_b$ are the specific surface area and the bulk density, respectively, of the soil at the reference state, the superscript zero refers to the reference state, and $2h_0$ is the thickness of the water film between two adjacent clay platelets within a tactoid ($2h_0 = 9 \times 10^{-8}$ cm, regardless of $h$, $R$, or $C_0$). Similarly, since in the confined soil element the increment in the water retained by the clay fraction is at the expense of the pore space in which water flow takes place, the effective porosity, $\varphi_i$, in each subvolume, $v_i$, is given by:

$$\varphi_i(h_i, R, C_0) = \theta_i^0(h_i) - [\nu_i(h_i, R, C_0) - \nu_i^0(h_i)] \qquad [11]$$

and the mean equivalent radius of the pores, $r_i$, in each subvolume, $v_i$ is given by:

$$r_i(h_i, R, C_0) = \left[\frac{\varphi_i(h_i, R, C_0)v_i}{\pi l_i}\right]^{1/2} \qquad [12]$$

Given the values of $\varphi_i(h_i, R, C_0)$ and $R_i(h_i, R, C_0)$, the relative hydraulic conductivity can be calculated on the basis of the capillary bundle theory. Using a modified series-parallel model of the porous medium, the hydraulic conductivity relative to the reference state is given by:

$$K_i^r(h_i, R, C_0) = \frac{[\varphi_i(h_i, R, C_0)]^2 \sum_{j=1}^{n}[2j-1][r_j(h_i, R, C_0)]^2}{[\varphi_i^0(h_i)]^2 \sum_{j=i}^{n}[2j-1][r_j^0(h_i)]^2} \qquad [13]$$

where $\varphi_i^0$ is the soil effective porosity in subvolume $v_i$, in the 'stable' reference state, $h_i < h_{i+1}$, and $i = 1$–n. Thus the soil hydraulic conductivity is given by:

$$K_i(h_i, R, C_0) = K_i^0(h_i)K_i^r(h_i, R, C_0) \qquad [14]$$

where $K^0(h)$ is the hydraulic conductivity function of the soil at the 'stable' reference state.

A comparison between the calculated and measured water retention curves in a soil containing 20% clay (dominated by montmorillonite), for different combinations of $C = 10^{-3}C_0[1 + (R+1)^{-1}]$ and $\text{SAR} = R[10^3C_0(R+1)^{-1}]^{1/2}$, is presented in **Figure 3**. The abscissa scale in **Figure 3** is shifted to the right and the data are accordingly translated along the $\theta$-axis to distinguish between the calculated and the measured curves for different sets of concentrations. For calcium-saturated soil (SAR = 0), the experimental retention curves are independent of the solution concentration, whereas the calculated curves are slightly dependent on it. The small discrepancy between the calculated and the measured curves can serve as a measure of the error in the assumptions and of the ability of the calculations to predict soil-water retention. For the mixed systems, there is reasonably good agreement between measured and calculated retention curves. The agreement improves for lower sodic systems (SAR decreases) and as the system becomes more concentrated.

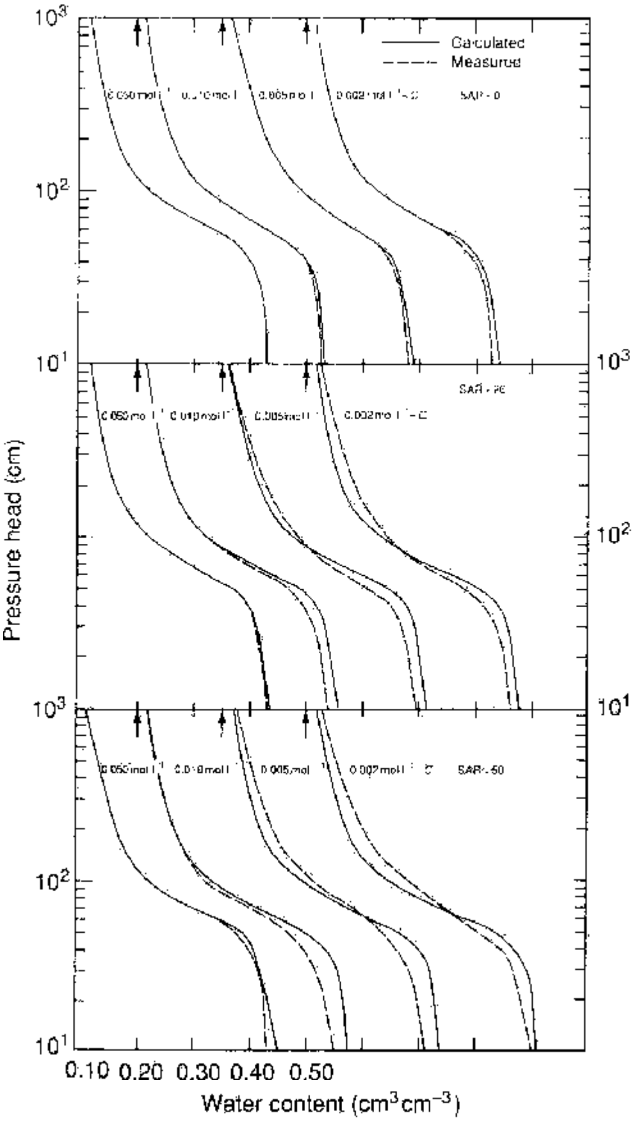


**Figure 3** Capillary pressure head, $h$, as a function of water content, $\theta$, for selected values of the concentration, $C$, and the sodium adsorption ratio, SAR, of the equilibrium solution. Computed results (solid lines) are compared with measured data (dashed lines). Arrows indicate that the point $\theta = 0.10$ is shifted and the data are translated along the $\theta$-axis. (Reproduced from Russo D (1976) MSc Thesis (in Hebrew), *Analysis of the Saturated–Unsaturated Hydraulic Conductivity in a Mixed Sodium-Calcium Soil System*. Faculty of Agriculture in Rehovot. Jerusalem, Israel. The Hebrew University Press.)

A comparison between the calculated and measured relative hydraulic conductivity, $K^r = K(\Theta, \text{SAR}, C)/K^0(\Theta)$, where $\Theta = (\theta - \theta_r)/(\theta_s - \theta_r)$ is the effective water saturation, and $\theta_s$ and $\theta_r$ are the saturated and residual values of $\theta$, respectively, in a soil containing 20% clay (dominated by montmorillonite), for

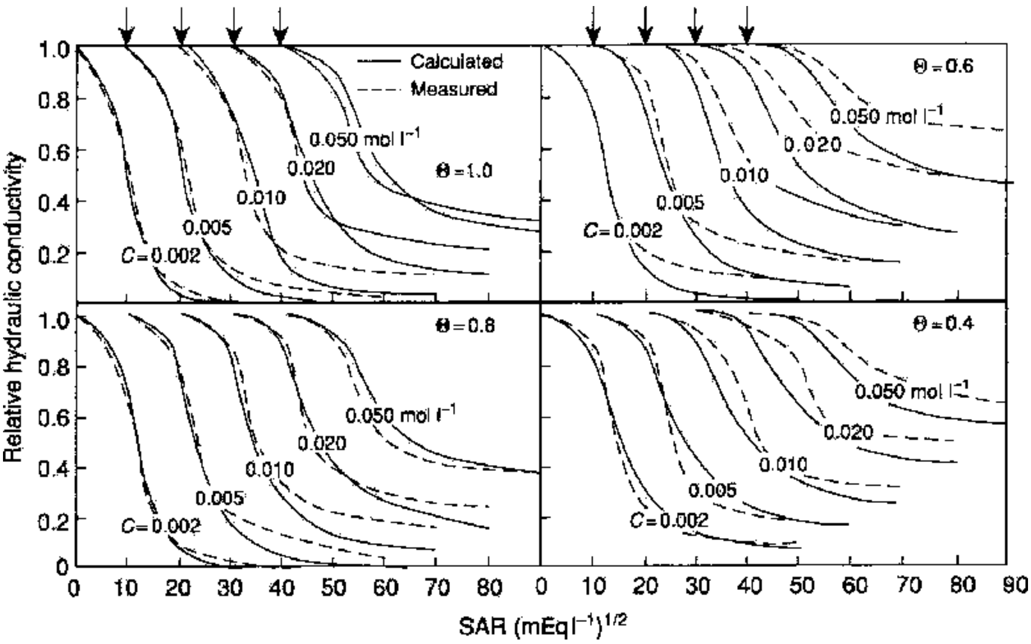


**Figure 4** Relative hydraulic conductivity, $K^r$, as a function of the sodium adsorption ratio, SAR, of the equilibrium solution, for selected values of the solution concentration, $C$, and degree of water saturation, $\Theta$. Computed results (solid lines) are compared with measured data (dashed lines). Arrows indicate that the point SAR = 0 is shifted and the data are translated along the SAR axis. (Reproduced with permission from Russo D and Bresler E (1977) Analysis of the saturated–unsaturated hydraulic conductivity in a mixed Na-Ca soil system. *Soil Science Society of America Journal* 41: 706–710.)

various combinations of $\Theta$, $C$, and SAR, is presented in **Figure 4**. The abscissa scale in **Figure 4** is shifted to the right and the data are accordingly translated along the SAR axis to distinguish between the calculated and the measured curves for different sets of concentrations. There is relatively good agreement between calculated and measured data in most cases. For the combinations of $\Theta$, SAR, and C depicted in **Figure 4**, the calculated $K^r$ deviates by less than twice the value of the measured $K^r$.

The comparison between calculated and experimental results indicates that a reasonable first-order approximation to the soil-water retentivity and conductivity functions can be obtained for relatively wide ranges of soil-solution concentrations and compositions. This can be done by taking into account the porous nature of the soil, the interactions between the charged solid phase and the ions in the liquid phase within the soil, and the effects of the water pressure and the soil solution composition on the organization and structure of the clay platelets.

Soils of differing textures, containing clay fractions that are mostly montmorillonite, have essentially the same surface charge density ($\Gamma_s \approx 4.5 \times 10^4$ esu cm$^{-2}$); consequently, they are characterized by essentially the same $\Gamma_{Na}(h, R, C_0)/\Gamma_s$ (eqn [7]) and $\Gamma_{Cl}(h, R, C_0)$ (see eqn [16]) relationships. On the other hand, the $\theta(h, R, C_0)$ and the $K^r(h, R, C_0)$ functions (**Figure 5**), and, concurrently, the $K(h, R, C_0) = K^0(h)K^r(h, R, C_0)$ function, associated with soils of different texture (see **Table 1**) may be quite different.

The functions depicted in **Figure 5** suggest that the interaction between the soil matrix and the soil solution affects the differences among the various soil types, in their hydraulic properties. In the case of the water retention curves, the differences between the values of $\theta$ associated with the various soils generally increase as $h$ and SAR increase and as C decreases. As for the hydraulic conductivity function, the differences between the values of $K^r$ and, concurrently, between values of $K = K^rK^0$, associated with the various soils increase as SAR increases and as $h$ and C decrease. **Figure 5** suggests that, because of soil-solution–soil-matrix interactions, the capability of the coarse-textured soil to transmit water and solutes faster than the finer-textured soils is extended to a larger pressure head (smaller degree of saturation) as the soil solution SAR increases and as C decreases.

Both the experimental and the theoretical results presented in the previous sections suggest that, for a given soil solution concentration and composition, the effect of the soil matrix–soil solution interactions on the hydraulic conductivity decreases with decreasing water saturation. This is demonstrated in **Figure 6**, in which combinations of soil-solution concentration (C) and composition (SAR) required to maintain $K^r \geq 0.75$ (that is, no more than a 25%

reduction in $K$ relative to the 'stable' reference state) are given for various degrees of water saturation, and for soils of several textures (Table 1). Note that the curves in **Figure 6** generalize the concept of

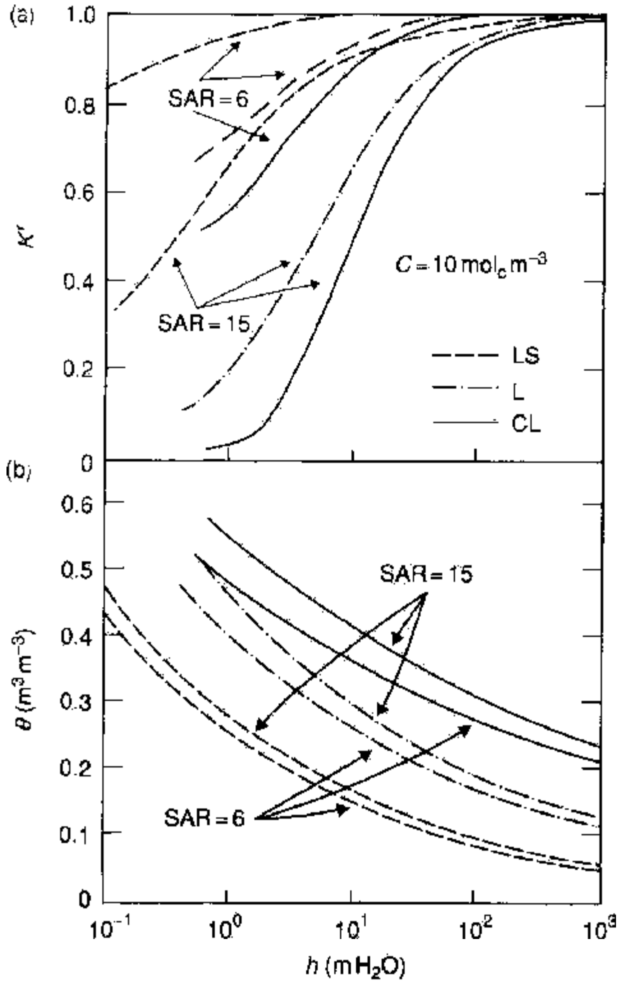


**Figure 5** Relative hydraulic conductivity, $K^r$ (a) and water content, $\theta$ (b) as a function of the capillary pressure head, $h$, for selected values of the concentration, $C$, and the sodium adsorption ratio, SAR, of the equilibrium solution, for three soils of different texture: loamy sand (LS), loam (L), and clay loam (CL). (Reproduced with permission from Russo D (1988) Numerical analysis of the nonsteady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284. American Geophysical Union.)

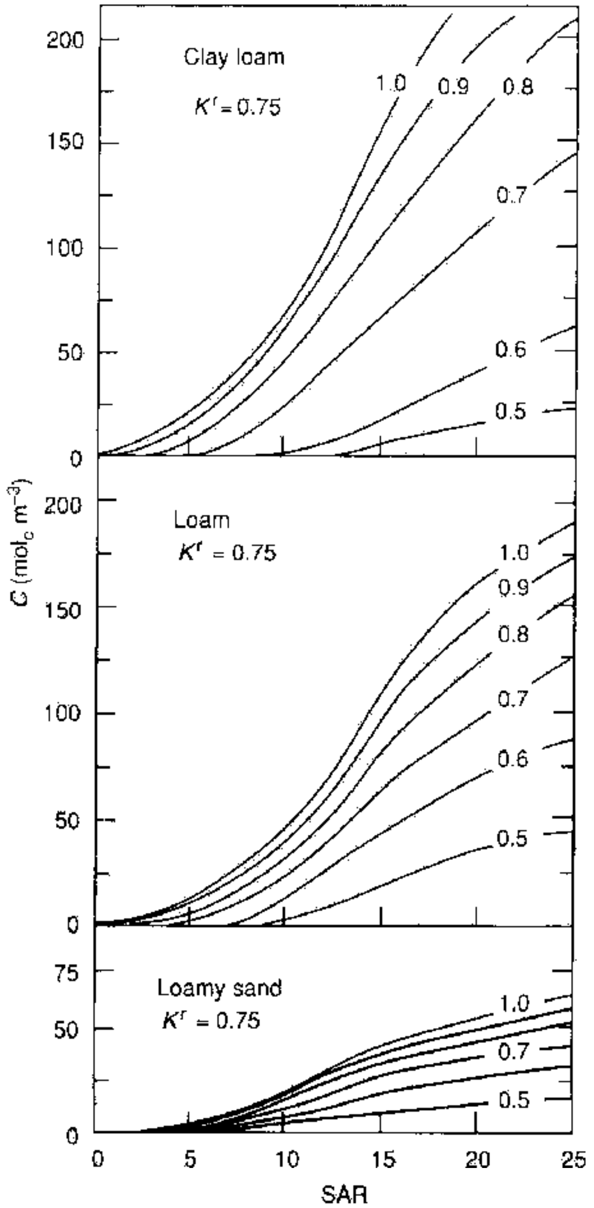


**Figure 6** Combinations of the concentration, $C$, and the sodium adsorption ratio, SAR, of the equilibrium solution, and degree of water saturation, Θ (the numbers labeling the curves) at which 25% reduction in hydraulic conductivity occurs, for three soils of different texture: loamy sand (LS), loam (L), and clay loam (CL). (Reproduced with permission from Russo D (1988) Numerical analysis of the nonsteady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284. American Geophysical Union.)

**Table 1** Representative values for relevant soil physical properties

| *Soil texture* | *Clay fraction* | $A_s$ ($cm^2 g^{-1}$) | *CEC* ($mEq g^{-1}$) | $K_s$ ($cm h^{-1}$) | $h_b$ (cm) | $\beta$ | $\theta_s$ ($cm^3 cm^{-3}$) | $\theta_r$ ($cm^3 cm^{-3}$) | $\rho_b$ ($g cm^{-3}$) |
|---|---|---|---|---|---|---|---|---|---|
| *Loamy sand* | 0.06 | $3.6 \times 10^5$ | 0.056 | 5.56 | 9.1 | 0.30 | 0.410 | 0.015 | 1.53 |
| *Loam* | 0.19 | $9.5 \times 10^5$ | 0.145 | 2.50 | 48.0 | 0.23 | 0.451 | 0.042 | 1.43 |
| *Clay loam* | 0.34 | $1.7 \times 10^6$ | 0.248 | 0.77 | 63.4 | 0.15 | 0.476 | 0.068 | 1.33 |

$A_s$, soil specific surface area; CEC, cation exchange capacity; $\rho_b$, bulk density; $K_s$, soil saturated hydraulic conductivity; $\theta_s$, saturated water content; $\theta_r$, residual water content; $h_b$, soil bubbling pressure head; $\beta$, pore-size distribution index.

Reproduced with permission from Russo D (1988) Numerical analysis of the nonsteady transport of interacting solutes through unsaturated soil. 1 Homogeneous systems. *Water Resources Research* 24. 271–284. American Geophysical Union.

the 'threshold concentration' in terms of soil properties and soil-water status. Consequently, they may be used for water-quality classification in relation to soils of different textures, and for water and soil management. For a given SAR, the solution concentration required to maintain $K^r \geq 0.75$ (that is, the 'threshold concentration') decreases as the soil-water content decreases and as the soil texture becomes coarser. **Figure 6** suggests that a decrease in the water application rate and, concurrently, in water saturation, may improve the efficiency of the reclamation of sodic soils, especially fine-textured ones.

## Implications for Flow and Transport

Soil-solution–soil-matrix physicochemical interactions, enhanced by dilute soil solutions with relatively high sodium-to-calcium ratios, can affect the flow parameters considerably (i.e., hydraulic conductivity and water retention functions). Furthermore, these interactions may affect the transport parameters (i.e., retardation and elution factors). In other words, soil-solution–soil-matrix physicochemical interactions may affect both water flow and solute transport.

Generally, for a given soil, and given boundary and initial conditions, the water flow is governed by the soil hydraulic properties. Thus, the effect of the soil-matrix–soil-solution interactions is to reduce the water flow relative to that of a 'stable' reference soil. For a given soil, the retardation of the water movement by the soil matrix–soil solution interaction will increase as either the initial soil solution SAR or the SAR of the applied water increases, as either the initial soil-solution concentration or the solution concentration of the applied water decreases, and as the surface water application rate increases.

On the other hand, the solute transport is governed by the water flow as well as by cation exchange and anion exclusion processes. Thus, in this case, the soil-matrix–soil-solution interactions may act in two opposing directions: the retardation of the solute movement because of reductions in the soil hydraulic conductivity and because of cation exchange, and the acceleration of the solute movement because of anion exclusion. The retardation effect caused by the reductions in the soil hydraulic conductivity and, concurrently, in water velocity, increases with increasing SAR and decreasing $C$ and $h$, especially in relatively fine-textured soils. The retardation effect because of Na/Ca exchange is given by the retardation factor, $R_f$:

$$R_f(h, R, C_0) = 1 + \frac{[R+1]\Gamma_{Na}(h, R, C_0) F A_{ad} \rho_b}{Z_{Na} R\, C_0 \theta(h)} \quad [15]$$

where $F$ is the Faraday constant, $A_{ad}$ is the specific adsorption surface area of the soil, and $\Gamma_{Na}$ is the quantity of sodium ionic charges per unit area of the charged surface (eqn [7]). The retardation caused by Na/Ca exchange (expressed in terms of $R_f$) increases with increasing SAR and decreasing $C$ and $h$, especially in relatively fine-textured soils.

Anion exclusion can be calculated from the distribution of the scaled electrical potential, $Y(x)$ (eqn [1]) within the domain $s \leq x \leq d$, which, in turn, for given values of $b(h, R, C_0)$ and $Y_s(h, R, C_0)$, is obtained from the solution of eqn [1], by the Runge–Kutta method. Applying the Boltzmann distribution law for the Na/Ca-Cl system considered here, the amount of excluded chloride per unit surface area is calculated from:

$$\Gamma_{Cl}(h, R, C_0) = \frac{|Z_{Cl}| C_0 (R+2)}{(R+1)} \times \int_s^d \{1 - \exp[-Y(x; h, R, C_0)]\} dx \quad [16]$$

which, in turn, is used to calculate the elution factor, $E_f$, given by:

$$E_f(h, R, C_0) = 1 - \frac{[R+1]\Gamma_{Cl}(h, R, C_0) A_{ex} \rho_b}{|Z_{Cl}| C_0 [R+2] \theta(h)} \quad [17]$$

where $A_{ex}$ is the specific exclusion surface area.

The acceleration effect due to the chloride exclusion, which is determined by the second term on the right-hand side of eqn [17], increases with increasing SAR and decreasing $C$ and $h$, especially in relatively fine-textured soils. The net effect of the soil matrix–soil solution interactions on the solute transport is determined, therefore, by the relative magnitudes of all of these processes, which, in turn, depend on the soil type and the boundary and initial conditions imposed on the flow system.

The effect on water flow and solute transport in a salt-affected soil of the conditions at the soil surface pertinent to an irrigation event and a rain event is demonstrated in **Figures 7** and **8**, respectively. The simulated profiles of water content, $\theta$ (top), salinity (in terms of electrical conductivity, EC; middle), and hydraulic conductivity, $K$ (bottom) in **Figures 7** and **8** are presented for three different soils ignoring (**Figures 7a** and **8a**) and considering (**Figures 7b** and **8b**) soil solution–soil matrix interaction. **Figures 7** and **8** clearly demonstrate that in, salt-affected soils, the retardation of water and solute movement by soil matrix–soil solution interaction increases as the soil texture is finer and the clay fraction is larger, particularly during a rain storm (**Figure 8**) associated with diluted solution and relatively high application rate.

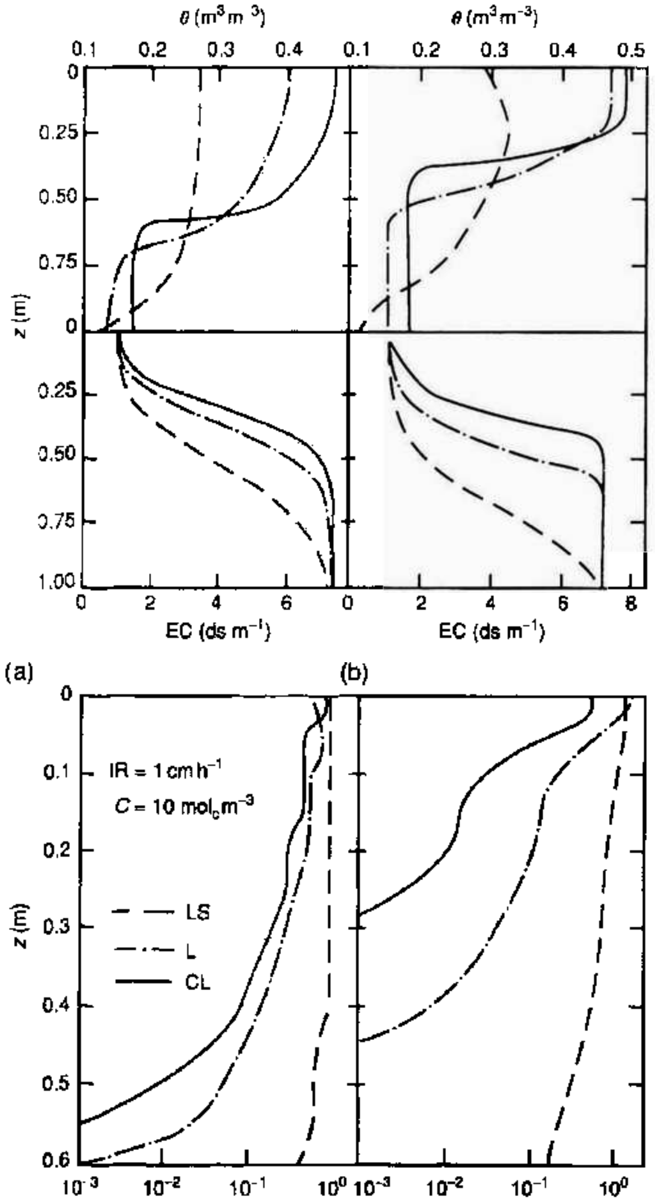


**Figure 7** Profiles of simulated soil-water content, $\theta$, soil-solution electrical conductivity, EC, and soil hydraulic conductivity, $K$, after 12h of infiltration (irrigation), and for the three different soils (**Table 1**). Initial profiles are given by $\theta(z, 0) = \theta_s - 0.3$, $C(z, 0) = 70\,\mathrm{mEq\,l^{-1}}$, and $\mathrm{SAR}(z, 0) = 25\,(\mathrm{mEq\,l^{-1}})^{1/2}$. (Reproduced with permission from Russo D (1988) Numerical analysis of the nonsteady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284.)

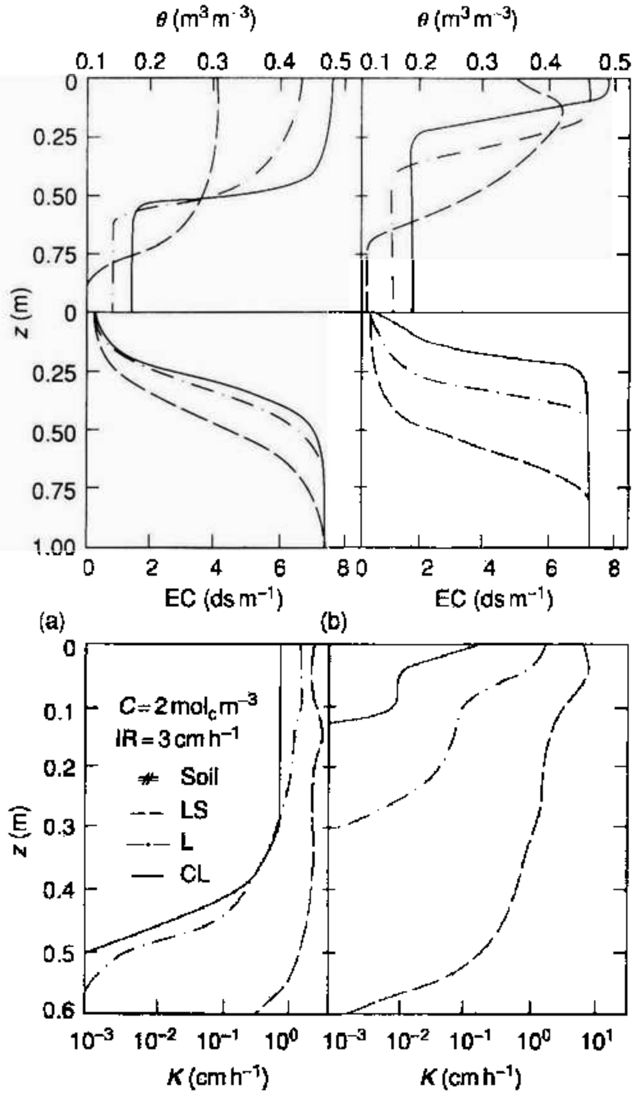


**Figure 8** Profiles of simulated soil-water content, $\theta$, soil-solution electrical conductivity, EC, and soil hydraulic conductivity, $K$, after 4h of infiltration (rain), and for the three different soils (**Table 1**). Initial profiles are given by $\theta(z, 0) = \theta_s - 0.3$, $C(z, 0)$ 70 $\mathrm{mEq\,l^{-1}}$, and $\mathrm{SAR}(z, 0) = 25\,(\mathrm{mEq\,l^{-1}})^{1/2}$. (Reproduced with permission from Russo D (1988) Numerical analysis of the nonsteady transport of interacting solutes through saturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284.)

## Summary

Soil solution–soil matrix interactions can alter the geometry of the soil pores and thus may affect the soil hydraulic conductivity, $K(h)$, and water retentivity, $\theta(h)$. The effect of soil solution–soil matrix interactions on $K(h)$ and $\theta(h)$, the retardation factor, $R_f$, and the elution factor, $E_f$, and, concurrently, on water flow and solute transport may be significant and generally increases as the soil solution sodium adsorption ratio increases; as both the soil solution concentration and the capillary pressure head decrease; as the clay fraction of the soil increases; and as the soil texture becomes finer.

It is worthwhile emphasizing that the experimental and the theoretical findings discussed here are relevant to soils whose clay fraction is dominated by smectite minerals (e.g., montmorillonite), with solutions containing Na, Ca, and Cl ions only. Furthermore, the model of the hydraulic functions in salt-affected soils presented here is based on a simplified description of the porous medium (the series-parallel model), and of the physicochemical interactions between the soil solution and the soil matrix (the mixed-ion diffuse double-layer theory). The model concentrates on the change in the soil pore-size distribution caused by the swelling of clay particles in a confined soil system, which, in turn, may be magnified by the breakdown of the domains of the clay platelets (tactoids) when the soil ESP exceeds a critical value. However, the model disregards the subsequent dispersion and movement of the clay platelets which may occur in coarse-textured soils subjected to a relatively large influx rate and, which, in turn, may further block the soil pores. Nevertheless, the data presented here are sufficiently reliable to indicate appropriate trends.

## List of Technical Nomenclature

| | |
|---|---|
| $\alpha = 8000\Pi/LDkT$ | (eqn [7]) (erg mol$^{-1}$) |
| $\beta$ | pore-size distribution index (dimensionless) |
| $\Gamma_{Cl}$ | quantity of the excluded chloride per unit area of the charged surface (mEq cm$^{-2}$) |
| $\Gamma_{Na}$ | quantity of sodium ionic charges per unit area of the charged surface (esu cm$^{-2}$) |
| $\Gamma_s$ | surface charge density (esu cm$^{-2}$) |
| $\gamma$ | specific gravity of water (g cm$^{-2}$ s$^{-2}$) |
| $\delta$ | space charge density (esu cm$^{-3}$) |
| $\epsilon$ | electronic charge (4.803 × 10$^{-10}$ esu electron$^{-1}$) |
| $\eta$ | dynamic viscosity of water (g cm$^{-1}$ s$^{-1}$) |
| $\Theta$ | effective water saturation (dimensionless) |
| $\theta$ | volumetric water content (cm$^3$ cm$^{-3}$) |
| $\theta_s$, $\theta_r$ | saturated and residual values of $\Theta$ (cm$^3$ cm$^{-3}$) |
| $\ell$ | average number of pores having a unit length (cm) |
| $\kappa$ | permeability (cm$^2$) |
| $\nu$ | volume of water retained by clay particles in a unit volume of soil (cm$^3$ cm$^{-3}$) |
| $\rho_b$ | bulk density (g cm$^{-3}$) |
| $\sigma$ | surface tension at the air–water interface (erg cm$^{-2}$) |
| $\varphi$ | porosity (cm$^3$ cm$^{-3}$) |
| $\psi$ | electrical potential (erg esu$^{-1}$) |
| $A_d$ | specific surface area for adsorption (cm$^2$ g$^{-1}$) |
| $A_e$ | specific surface area for exclusion (cm$^2$ g$^{-1}$) |
| $A_s$ | specific surface area (cm$^2$ g$^{-1}$) |
| $a_i(i = 1–4)$ | coefficients in eqns [3], [5], and [6] (dimensionless) |
| $B = kLTC_0/10^3 (R+1)$ | (eqn [2]) (erg cm$^{-3}$, where 1 erg = 1 g cm$^2$ s$^{-2}$) |
| $b$ | half-distance between two adjacent tactoids (cm) |
| $C_{Na}$ | concentration of the ion identified subscript (mEq l$^{-1}$) |
| $C = C_{Na} + C_{Ca}$ | (mEq l$^{-1}$) |
| $C_0 = c_{Na} + c_{Ca}$ | (mol l$^{-1}$) |
| $c_{Na}$ | molar concentration of the ion identified subscript (mol l$^{-1}$) |
| $D$ | the dielectric constant of water (dimensionless) |
| $d$ | position of the midplane between two adjacent tactoids (cm) |
| $E_f$ | elution factor (dimensionless) |
| $F$ | Faraday constant (2.893 × 10$^{11}$ esu mEq$^{-1}$, where 1 esu = 1 g$^{1/2}$ cm$^{3/2}$ s$^{-1}$) |
| $G = (8\Pi C_0 L\epsilon^2/10^3 DkT)^{1/2}$ | (molecules electron$^{-1}$ cm$^{-1}$) |
| $h$ | capillary pressure head (cm) |
| $h_b$ | air entry value of $h$ (cm) |
| $i, j$ | indices (dimensionless) |
| $K$ | hydraulic conductivity (cm s$^{-1}$) |
| $K^r$ | relative hydraulic conductivity (dimensionless) |
| $k$ | Boltzmann constant (1.38 × 10$^{-16}$ erg deg$^{-1}$ molecule$^{-1}$) |
| $L$ | Avogadro number (6.0225 × 10$^{23}$ molecules mol$^{-1}$) |
| $P$ | swelling pressure (erg cm$^{-3}$) |
| $P_w$, $P_a$, and $P_c = P_w - P_a$, | water, air, and capillary pressures, respectively (erg cm$^3$) |
| $R = c_{Na}/c_{Ca}$ | (dimensionless) |
| $R_f$ | retardation factor (dimensionless) |
| $r$ | equivalent radius of water-filled pore (cm) |
| $s$ | position of the plane of the charged surface (cm) |

| | |
|---|---|
| SAR | sodium adsorption ratio $(C_{Na}/\sqrt{C_{Ca}/2})((\text{mEq l}^{-1})^{1/2})$ |
| $T$ | absolute temperature (degrees Kelvin) |
| $V$ | volume (cm$^{3}$) |
| $v_i$ | subvolume ($\Sigma v_i = V$) (cm$^{3}$) |
| $x$ | distance (cm) |
| $Y$ | scaled electrical potential (molecule electron$^{-1}$) |
| $Y_d$ | value of $Y$ at the midplane between two adjacent tactoids (molecule electron$^{-1}$) |
| $Y_s$ | value of $Y$ at the charged surface (molecule electron$^{-1}$) |
| $Z_{Na}$ | valence of the ion subscript (electron molecule$^{-1}$) |

*See also:* **Nitrogen in Soils:** Cycle; Nitrates; Plant Uptake

## Further Reading

Bear J (1972) *Dynamics of Fluids in Porous Media*. New York: Elsevier.

Blackmore AV and Miller RD (1961) Tactoid size and osmotic swelling in Ca-montmorillonite. *Soil Science Society of America Proceedings* 25: 169–173.

Bresler E (1972) Interacting diffuse layers in mixed mono-divalent ionic systems. *Soil Science Society of America Proceedings* 36: 891–896.

Bresler E (1978) Theoretical modeling of mixed-electrolyte solution flows for unsaturated soils. *Soil Science* 125: 196–203.

Bresler E, McNeal BL, and Carter DL (1982) *Saline and Sodic Soils: Principles-Dynamics Modeling*. Berlin, Germany: Springer-Verlag.

Brutsaert W (1967) Some methods of calculated unsaturated permeability. *Transactions of the American Society of Agricultural Engineers* 10: 400–404.

Childs EC and Collis-George N (1950) The permeability of porous materials. *Proceedings of the Royal Society of London, Series A: Mathematical and Physical Sciences* 201: 392–405.

James RV and Rubin J (1986) Transport of chloride ion in a water-unsaturated soil exhibiting anion exclusion. *Soil Science Society of America Journal* 50: 1142–1149.

Marshall TJ (1958) A relation between permeability and size distribution of pores. *Journal of Soil Science* 9: 1–8.

Nielsen DR, van Genuchten MTh, and Biggar JW (1986) Water flow and solute transport processes in the unsaturated zone. *Water Resources Research* 22: 89S–108S.

Russo D (1988) Numerical analysis of the nonsteady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284.

Russo D and Bresler E (1977) Effect of mixed Na-Ca solutions on the hydraulic properties of unsaturated soils. *Soil Science Society of America Journal* 41: 714–717.

Russo D and Bresler E (1977) Analysis of the saturated–unsaturated hydraulic conductivity in a mixed Na-Ca soil system. *Soil Science Society of America Journal* 41: 706–710.

Shainberg I (1984) The effects of electrolyte concentration on the hydraulic properties of sodic soils. In: Shainberg I and Shalhevet J (eds) *Soil Salinity Under Irrigation: Processes and Management*, pp. 49–64. New York: Springer-Verlag.

Shainberg I and Kaiserman A (1969) Kinetics of the formation and breakdown of Ca-montmorillonite tactoids. *Soil Science Society of America Proceedings* 33: 547–551.

# Salt Balance of Soils *See* Salination Processes

# SALT-AFFECTED SOILS, RECLAMATION

**R Keren**, The Volcani Center, Bet Dagan, Israel

## Introduction

Saline and sodic soils exist in arid and semiarid regions and in regions of poor natural drainage. These soils have poor physical properties owing to high sodicity and high pH. Under these conditions, slow water infiltration into the soil, soil erosion, slow internal drainage, poor aeration, compaction, and waterlogging on lower lands commonly occur due to soil swelling and dispersion.

Considerable progress has been made in managing and controlling salinity, sodicity, and alkalinity in irrigated lands. However, because of natural hydrological and geochemical factors, as well as irrigation-induced activities, soil salinity, sodicity, alkalinity, and associated drainage and soil erosion continue to plague agriculture. Control of salinity and sodicity is essential to establish a sustainable, successful agriculture.

## Effect of Salts on Soil Properties

### Hydraulic Conductivity

Many soils in arid to humid regions have an unstable structure, which makes them difficult to manage because of their tendency to swell and disperse. Since soil permeability decreases with the square of the pore radius, a small reduction in size of the large pores due to swelling and clay migration has a large effect on soil permeability for water and gases. The extent of swelling and dispersion of clays depends on the clay mineralogy, the composition of the adsorbed ions, and the electrolyte concentration (EC) of the soil solution. Whether clay will migrate or seal the soil depends on the extent of swelling. Clay swelling is a continuous process that increases gradually with decreases in EC, whereas clay dispersion is possible only at EC below the flocculation value (FV). The degree of clay swelling before replacing the solution with one that has an EC below the FV of the clay determines whether clay will migrate in the conducting pores, or whether the clay particles will be trapped in the narrow pores (sieving effect) and the hydraulic conductivity (HC) of the soil will further decrease.

Even in arid and semiarid regions where irrigation is practiced, rainstorms occur occasionally. Under such conditions, sodicity damage to the soil will not take place during the irrigation season when the EC in the applied water may be high enough to counter the dispersive effect of the exchangeable Na. The sodicity damage will occur during the rainy season when the EC decreases to values below the threshold concentration determined by the properties of the soil and the exchangeable sodium percentage (ESP). Thus, when soils are exposed to rainwater, susceptibility to low levels of ESP is enhanced. It is obvious, therefore, that any attempt to assess the sodicity hazard in soils, the clay mineralogy, and the total EC of percolating solution should be considered. At any particular ESP, the influence of exchangeable Na is greater with exchangeable Mg as the complementary ion than it is with Ca.

### Infiltration Rate

In soils having stable surface structures, decreases in infiltration rate (IR) result from the inevitable decrease in the matric suction gradient which occurs as infiltration proceeds. Decreases in soil IR from an initially high rate can also result from gradual deterioration of soil structure and the formation of a surface seal. When a seal of very low HC is formed, its reduced permeability determines the IR of the soil, and the wetting-front depth has only a slight effect on it. The presence of a seal increases detachment forces owing to runoff, but at the same time increases the shear strength of the soil surface. Thus, the net effect depends on the forces acting between soil particles at soil surface and the energy involved in water-drop impact and shearing forces of the overland flow.

The IR is much more sensitive to ESP than is HC. The higher sensitivity of the soil surface to ESP when the soil surface is exposed to rainwater is explained by: (1) the mechanical impact of the raindrops, which enhances clay dispersion; (2) the absence of the soil matrix, which slows clay transport; and (3) the low EC in the soil surface. Low EC and high ESP cause clay swelling and dispersion, which enhance seal formation and therefore runoff. Adsorbed Mg has a specific effect on soil erosion and IR for montmorillonitic, noncalcareous and calcareous soils.

## Reclamation of Salt- and Sodium-Affected Soils

### Reclamation of Salt-Affected Soils

Salinity control in the root zone is one of the main problems of irrigation in arid and semiarid regions. In these regions a salinity problem caused or complicated by poor drainage cannot be adequately controlled unless the water table is stabilized and maintained at a safe depth. This requires open or tile drains, or drainage wells to remove a part of the subsurface saline water and transport it to an acceptable sink for safe disposal.

Effective salinity control depends largely on the initial soil salinity, the content of the soil water during leaching, hydrodynamic dispersion, spatial variation, the salt tolerance of the crops, the depth of the water table and, when applicable, layout of the drainage system. The prevailing concept in earlier reclamation trials was to leach all excess salts from the entire root zone. Under well-drained conditions, however, the aim should be to reduce salinity in the upper soil layer (depending on crop type) to below the threshold value of the growing crop.

### Leaching Methods

**Continuous and intermittent ponding** The efficiency of salts leaching can be defined as the quantity of soluble salts leached per unit volume of water applied. The time factor should also be considered when leaching takes place under conditions of high evaporation. The range of soil pore-size distribution also affects the leaching efficiency. Pore sizes may vary enough to cause substantial changes in hydrodynamic dispersion, especially during ponded leaching. The hydrodynamic dispersion decreases under unsaturated flow, resulting in the more efficient transport of salt. Thus, the efficiency increases as the soil water content decreases during leaching.

On the basis of leaching theory, a water volume of one replacement of pore volume (pv) will reduce soil salinity by one-half, and a pv of 1.5–2.0 of water should remove approximately 80% of the soil salt content. This volume corresponds approximately to an equivalent depth of water per unit depth of soil (assuming a pv of 50%).

Based on field data from various parts of the world, the following empirical formula for salt transport efficiency under one-dimensional leaching is:

$$\frac{C}{C_0} = \frac{k}{(D/D_S)} \qquad [1]$$

where C and $C_0$ are the salt concentration in the soil at any given time during leaching and the initial concentration, respectively, $D$ and $D_S$ are the depth of leaching water applied and the soil depth to be leached, respectively, and $k$ is an empirical coefficient that ranges from 0.1 to 0.3 for sandy loam and clay soil, respectively, under ponding leaching as shown in Figure 1. Under intermittent ponding with 50–150 mm of water per application or under sprinkler irrigation, the empirical coefficient is approximately 0.1, irrespective of the soil type. This equation applies in the range of $D/D_S > k$. The salt transport efficiency decreases sharply when $D/D_S$ exceed 0.5 or 0.75 for sandy loam and for clay loam to clay soil, respectively.

When natural drainage conditions are limited and an artificial drainage system is required to remove leaching water, leaching efficiency may decrease because the hydraulic gradient is steeper near the drains than away from the drains. Therefore, a larger fraction of the leaching water will flow through the soil next to the drains. This amount depends on the drain spacing ($S$) and the depth ($Y$). For example, when $S/Y = 2.5$, only 30% of the area between drains will be leached with maximum efficiency. To increase leaching efficiency, drain spacing and depth should be determined by the use of a drainage equation.

The intermittent ponding method is especially suitable in fields with tile drains, as it allows the water table to draw down, which greatly increases leaching efficiency. When soils develop a surface seal, intermittent ponding may also help water to infiltrate by forming cracks. However, the longer period of wet soil exposure increases water evaporation. Combining intermittent ponding with mulching greatly improves the leaching efficiency.

When fields consist of several types of soils, leaching takes place in sandy soils at rates different to those of more clayey soils. If leaching is continued until the clay soil is sufficiently leached, an excessive amount of water percolates into the sandy soil. This problem is pronounced under ponded leaching, during which ponded water percolates predominantly through the sandy soil. To leach much more efficiently, it is possible to use the stochastic analysis of solute transport in spatially variable fields to divide a field into small sections that need different amounts of water.

**Sprinkling** As mentioned above, the leaching efficiency increases as the soil water content decreases. Thus, leaching by sprinkling at rates lower than the infiltration rate of the soil will result in a more efficient removal of accumulated salts than will be leached by ponding. Ponded leaching requires three times as much water as sprinkling to reduce soil salinity by the same increment. Sprinklers can be used advantageously when the field is unprepared for ponded leaching or for sandy soils. This method has a high leaching efficiency, except under high evaporation, because the leaching takes place under unsaturated conditions. This method can also be used advantageously when water for leaching is scarce and costly.

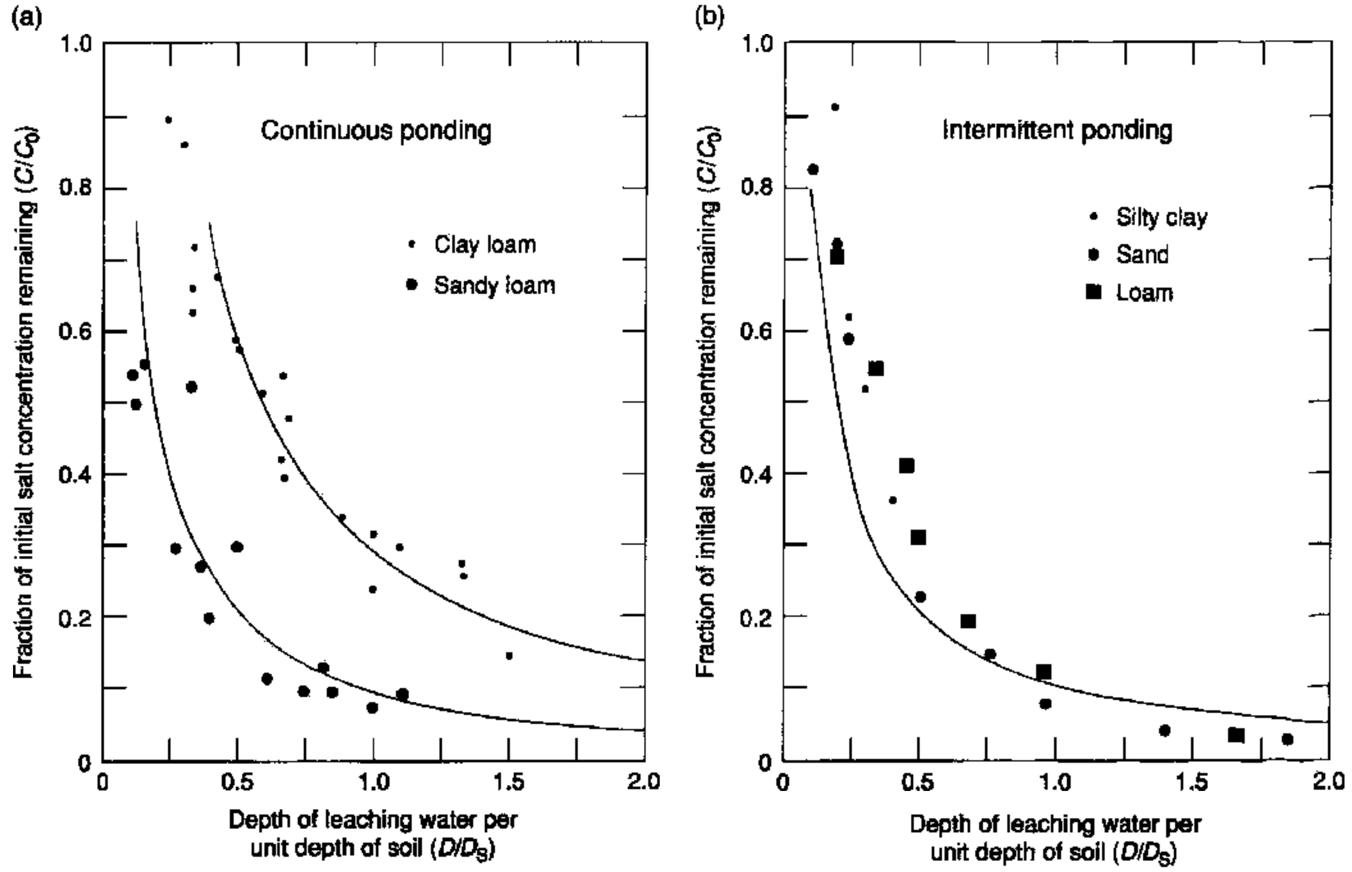


**Figure 1** Relationship between fraction of initial salt concentration remaining in soil, $C/C_0$, and depth of leaching water applied per unit depth of soil, $D/D_S$: (a) for continuous ponding; (b) for intermittent ponding. (Adapted from Hoffman GJ (1980) Guidelines for the reclamation of salt-affected soils. In: O'Connor GA (ed.) *Second Inter-American Conference on Salinity and Water-Management Technology*, pp. 49–64. Juarez, Mexico.)

In summary, the intermittent-ponding method requires less water than the continuous-ponding method to achieve the same degree of leaching, and the sprinkling method is more efficient than the other methods at removing salt from small pores in the soil profile.

**Postleaching operation** Prolonged ponding destroys the structure of the soil. Therefore, the last phase of reclamation is drying the soil and, when necessary, taking steps to enhance soil structure, such as applying gypsum or organic matter. Sprinklers or intermittent ponding generally cause less soil-structure destruction and may not require elaborate postleaching measures.

### Reclamation of Sodium-Affected Soils

In order to reduce runoff and erosion in sodic soils, reclamation is a common practice. Reclamation of sodic soils involves replacement of exchangeable Na by Ca. The source of Ca for replacing adsorbed Na can be external sources such as gypsum, calcium chloride, and irrigation water containing Ca ions, or the soil itself, involving the dissolution of Ca-containing minerals.

Although chemical amendments may be effective in restoring water penetration after reduction caused by either excessive adsorbed sodium or a low salt concentration in the applied water, they will not be effective if low permeability is due to soil texture, compaction, or water-restricting layers. Thus, reclamation requires a certain flow of water through the soil profile, and an appropriate IR and HC must be achieved.

### Reclamation by Adding Soil Amendments

**Gypsum** Gypsum is the most commonly used amendment for sodic soil reclamation and for reducing the harmful effects of high-sodium irrigation waters because of its solubility, low cost, and availability. Gypsum added to a sodic soil can cause permeability changes by increasing EC and by cation exchange effects. The relative significance of the two effects is of interest for several reasons. If the electrolyte effect is sufficiently high to prevent dispersion and swelling of soil clays, the surface application of gypsum may be worthwhile. In this case the amount of gypsum required depends on the amount of high-quality water applied and the rate of gypsum dissolution. It is somewhat independent of the amount of exchangeable Na in the soil profile. Conversely, in soils where the EC effect is not enough due to the high ESP level, or in order to obtain permanent improvement, then the amount of gypsum required depends on the amount of exchangeable Na in a selected depth of soil.

The benefit of gypsum for reclaiming sodic soils depends not only on the infiltration characteristics of the soil, but also on the gypsum-dissolution properties. The main mechanism by which gypsum affects the IR of soils exposed to rainwater is by its dissolution and release of electrolytes into the soil solution. The dissolution rate of gypsum is an important factor in IR because of the short contact time between rainwater and gypsum particles at the soil surface. Some of the factors that influence the dissolution rate of gypsum are the surface area of the gypsum particles, the soil water velocity during leaching, and the electrolyte composition of the soil solution.

Gypsum source, application rate, and particle size all have their effect on crust formation and IR (**Figure 2**). Although the IR for soil without gypsum decreases sharply as the cumulative rainfall increases, the IR remains relatively high in the presence of phosphogypsum (PG), whereas the effectiveness of mined gypsum (MG) on IR is significantly less. The higher efficiency of PG in maintaining high IR is attributed to its high rate of dissolution (**Table 1**). Application of gypsum on sodic calcareous soil is very effective in lowering surface runoff under field conditions (**Table 2**). Spreading the gypsum over the soil is more effective than mixing it in the upper 10 cm of soil. Although gypsum affects the ESP of the soil in the upper 0- to 15-cm layer very little (**Table 3**), it has a large effect on IR (**Figure 2**). The gypsum has a significant effect on the amount of soil erosion. The effectiveness is greater when the gypsum particles are spread over the soil rather than mixed in the top 10 cm of soil. The presence of gypsum on the soil surface increases soil stability by increasing EC in the percolated solution and by decreasing ESP. This effect of gypsum on soil erosion is due to two reasons: (1) reducing runoff by increasing IR, and (2) changing the erodibility factor. The erodibility of the soil is essentially related to the stability of soil aggregates and to the cohesive forces holding the aggregates together.

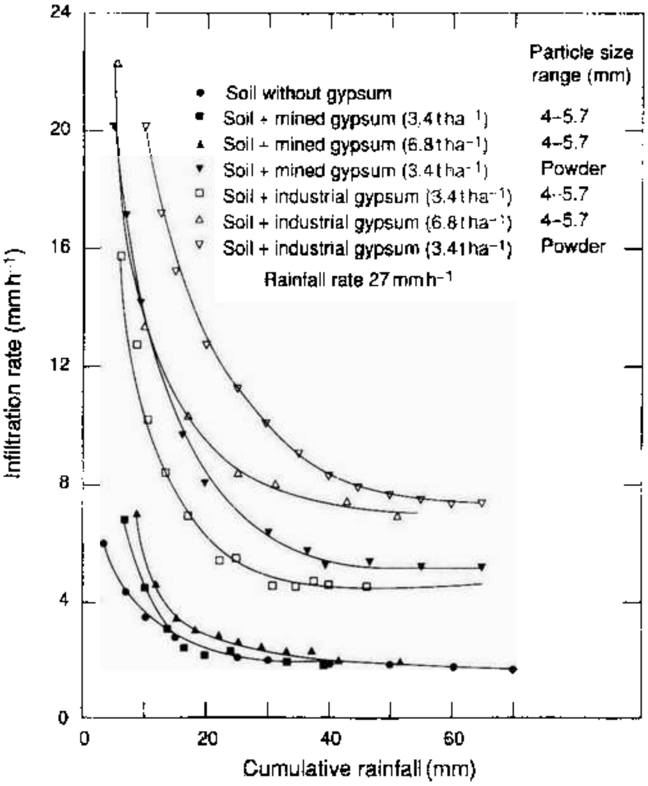


**Figure 2** Effect of industrial and mined gypsum (at amounts and fragment sizes as indicated) on infiltration rate of loess soil as a function of cumulative rainfall. (Adapted from Keren R and Shainberg I (1981) The efficiency of industrial and mined gypsum in reclamation of a sodic soil – rate of dissolution. *Soil Science Society of America Journal* 45: 103–107, with permission.)

**Table 1** The dissolution coefficients of industrial gypsum, $K_{(PG)}$ and mined gypsum, $K_{(MG)}$, with different fragment sizes

| | *Particle size (mm)* | |
|---|---|---|
| *Dissolution coefficients ($s^{-1}$)* | *1.0–2.0* | *4.0–5.7* |
| $K_{(PG)} \times 10^4$ | 198 | 58 |
| $K_{(MG)} \times 10^4$ | 19 | 6 |
| $K_{(PG)}/K_{(MG)}$ | 10.4 | 9.7 |

Adapted from Keren R and Shainberg I (1981) The efficiency of industrial and mined gypsum in reclamation of a sodic soil – rate of dissolution. *Soil Science Society of America Journal* 45: 103–107, with permission.

In order to obtain permanent improvement, however, the cation exchange effect is most important. The amount of exchangeable sodium to be replaced during reclamation depends on the initial ESP ($ESP_i$), the cation exchange capacity (CEC; in moles of cation charge per megagram), soil bulk density ($\rho_b$; in megagrams per cubic meter), the desired final exchangeable sodium fraction ($ESP_f$) and the depth of soil to be reclaimed ($L$; in meters). Once the above parameters have been determined, the amount of exchangeable Na to be replaced per unit area of land ($Q_{Na}$; in moles of cation charge per hectare) can be calculated from:

$$Q_{Na} = 10^4 L\rho_b(\mathrm{CEC})(\mathrm{ESP_i} - \mathrm{ESP_f}) \qquad [2]$$

The $ESP_f$ value depends on the response of the soil in terms of its physical conditions.

**Table 2** Surface runoff from loess soils at two levels of exchangeable sodium percentage (ESP) during rainstorms as affected by phosphogypsum (PG)

| | | | *Surface runoff (% of rainfall)*[a] | | | | |
|---|---|---|---|---|---|---|---|
| | | | *ESP 4.6* | | *ESP 19.3* | | |
| | | | | | | *PG*[b] | |
| *Storm no.* | *Rainfall (mm)* | *Time between rainstorms (days)* | *Control* | *PG*[b] *spread over* | *Control* | *Mixed in* | *Spread over* |
| 1 | 16 | - | 0 | 0 | 23.1 | 2.6 | 0 |
| 2 | 20 | 40 | 6.5 | 0 | 13.3 | 3.0 | 0.7 |
| 3 | 19 | 20 | 1.3 | 0.5 | 27.3 | 10.5 | 3.8 |
| 4 | 59 | 6 | 21.5 | 3.8 | 41.0 | 21.2 | 17.2 |
| 5 | 44 | 9 | 13.4 | 2.1 | 45.0 | 22.5 | 12.4 |
| 6 | 12 | 1 | 12.5 | 2.5 | 40.0 | 25 | 12.5 |
| 7 | 12 | 15 | 10.1 | 2.5 | 31.7 | 13.8 | 6.7 |

[a] Percentage of annual rainfall (182 mm): ESP 4.6: Control, 12.5%; PG spread over, 2.1%; ESP 19.3: Control, 35.3%; PG spread over, 10.3%; PG mixed in, 16.5%.

[b] The PG rates were 5 Mg ha$^{-1}$ and 10 Mg ha$^{-1}$ for the soils having ESP values of 4.6 and 19.3, respectively.

Adapted from Keren R, Shainberg I, Frenkel H, and Kalo Y (1983) The effect of exchangeable sodium and gypsum on surface runoff from loess soil. *Soil Science Society of America Journal* 47: 1001–1004, with permission.

**Table 3** Exchangeable sodium percentage (ESP) in the profile of the sodic loess soil as affected by gypsum treatments

| | *ESP*[a] | | | | |
|---|---|---|---|---|---|
| | *Soil layer depth (cm)* | | | | |
| *Treatment* | *0–15* | *15–30* | *30–45* | *45–60* | *60–90* |
| *Control*[b] | | | | | |
| Autumn | 19.3±4.8 | 19.0±4.9 | 16.4±5.1 | 16.4±4.8 | 14.2±3.9 |
| Spring | 16.9±2.9 | 19.4±4.0 | 19.0±2.8 | 19.3±3.4 | 19.6±4.0 |
| *Gypsum mixed in* | | | | | |
| Autumn | 17.4±5.0 | 20.1±4.6 | 20.3±6.0 | 20.8±6.0 | 21.3±6.7 |
| Spring | 12.8±5.0 | 17.7±3.6 | 19.6±3.8 | 17.4±2.9 | 18.0±2.3 |
| *Gypsum spread over* | | | | | |
| Autumn | 16.9±5.8 | 17.0±5.8 | 19.3±6.0 | 20.9±4.5 | 21.1±4.5 |
| Spring | 11.5±4.1 | 16.3±3.6 | 19.4±3.5 | 19.3±7.0 | 20.4±4.3 |
| CEC (cmol$_c$ kg$^{-1}$ soil) | 18.2±1.5 | 17.8±1.3 | 18.5±2.0 | 18.9±2.2 | 18.1±1.1 |

[a] Mean values with standard deviation (12 and 43 repetitions for ESP and CEC, respectively).

[b] Autumn 1981 (after irrigation with 400 mm of saline water, before gypsum application) and spring 1982 (after the rainy season).

CEC, cation exchange capacity.

Adapted from Keren R, Shainberg I, Frenkel H, and Kalo Y (1983) The effect of exchangeable sodium and gypsum on surface runoff from loess soil. *Soil Science Society of America Journal* 47: 1001–1004.

The amount of gypsum needed to reclaim a sodic soil, in metric tonnes per hectare, can be calculated using eqn [3]:

$$GR = 8.61 \times 10^{-5} Q_{Na} \qquad [3]$$

The efficiency and rate of exchange, namely, the percentage of applied Ca that exchanges for adsorbed Na, varies with ESP, being greater at high ESP values. Removal of Na at ESP levels below 10 is slow, and some of the applied Ca displaces exchangeable Mg so that the efficiency declines to about 30%. Efficiency may also be low (20–40%) in fine-textured soils, because of the slowness of exchange of Na inside the structural units.

Generally, soil water penetration is too low to allow reclamation of sodic soils to be attained in a single leaching. For example, a 50-cm depth of water applied for leaching can only dissolve about 12 Mg ha$^{-1}$ of gypsum (reaching a saturation solution in respect to gypsum). Larger gypsum applications, therefore, will not be effective unless soil water penetration is sufficient for larger water applications. Thus, sodic soil can normally only be reclaimed to a shallow depth the first year (depending on the depth of water applied for leaching), but this will often permit a shallow-rooted crop to be grown after leaching (**Figure 3**). Subsequently annual applications of amendments and leaching volumes can be applied to reclaim the entire profile over a period of years.

In estimating the amount of gypsum to be applied, little attention has been directed to the possibility that

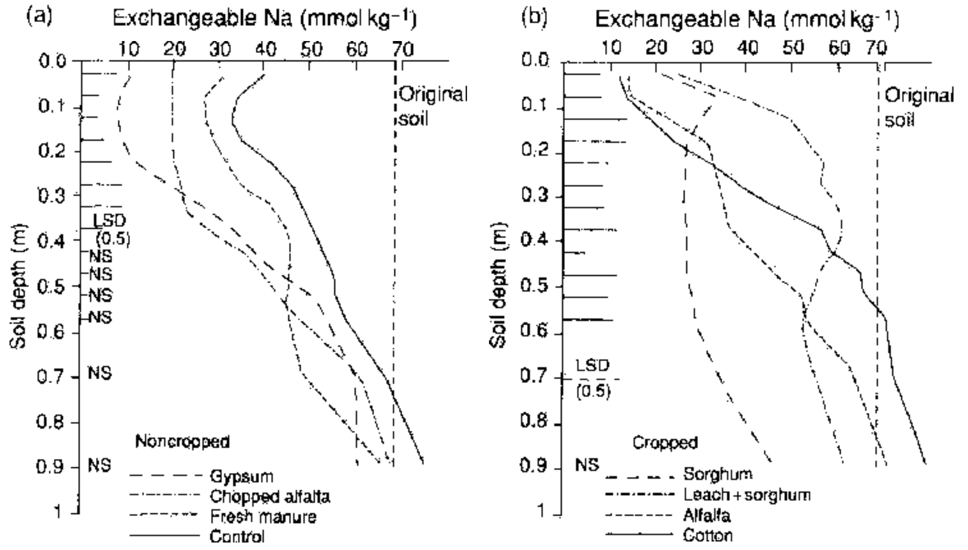


**Figure 3** Final distribution of exchangeable Na levels with soil depth resulting from reclamation treatments for (a) noncropped and (b) cropped treatments. LSD, least significant difference; NS, no significance. (Adapted from Robbins CW (1986) Sodic calcareous soil reclamation as affected by different amendments and crops. *Agronomy Journal* 78: 916–920, with permission.)

incorporating large amounts of gypsum powder into the soil may cause temporary reduction in HC because excess small particles of gypsum may block conducting pores in the early stage of the dissolution process.

The soil solution in the gypsum layer reaches equilibrium (in the presence of excess gypsum) only when the contact time between an elemental volume of solution and gypsum particle surface is sufficiently long, or when the surface area of the gypsum fragments per volume of water is sufficiently large. Some of the factors that influence the rate of gypsum dissolution are the surface area of the gypsum particles, the soil water velocity during leaching, and the electrolyte composition of the soil solution.

The gypsum dissolution rate is written as:

$$(\mathrm{d}C/\mathrm{d}t) = k(C_s - C) \qquad [4]$$

where $\mathrm{d}C/\mathrm{d}t$ is the net rate of dissolution, $k$ is the dissolution coefficient, and $C_s$ and $C$ are the solution concentration at saturation and at time $t$, respectively. Increasing solution flow velocity increases the dissolution rate coefficient, but decreases the contact time between gypsum and the flowing solution; the net effect is a decreasing dissolution rate with increasing soil solution flow velocity.

Integrating eqn [4] yields:

$$-\ln\left(1 - \frac{C}{C_s}\right) = kt \qquad [5]$$

Since the thickness of the stationary film of solution around the gypsum particles changes with the soil water velocity, the dissolution rate coefficient for a given surface of gypsum particle also changes. Thus, the lines obtained from eqn [5] are not linear. The left-hand side of eqn [5] is empirically related linearly to the square root of time:

$$-\ln\left(1 - \frac{C}{C_s}\right) = \alpha t^{1/2} + \beta \qquad [6]$$

where $\alpha$ and $\beta$ are the slope and intercept, respectively, and $t$ is the time for an increment of solution to leave a given soil depth:

$$t = \frac{L}{V} \qquad [7]$$

where $L$ is the length of the layer of soil-gypsum mixture and $V$ is the soil water velocity. Combining eqns [6] and [5] and introducing eqn [7] yields:

$$k = (V/L)^{1/2}\left[\alpha + (V/L)^{1/2}\beta\right] \qquad [8]$$

The dissolution rate coefficient as a function of water flow velocity can be evaluated for a given surface area of gypsum. Since the permeability of a sodic soil is low, the dissolution rate is relatively high during the initial stage of reclamation. During the reclamation process, however, the permeability increases and the dissolution rate decreases. Thus, the soil regulates gypsum-dissolution rate during reclamation. When it is possible to control water flow velocity in soils (e.g., sprinkler irrigation), reduced velocities are to be preferred for increased gypsum dissolution rate and greater efficiency of cation exchange reaction due to increase in contact time between water and gypsum particles (Figure 4).

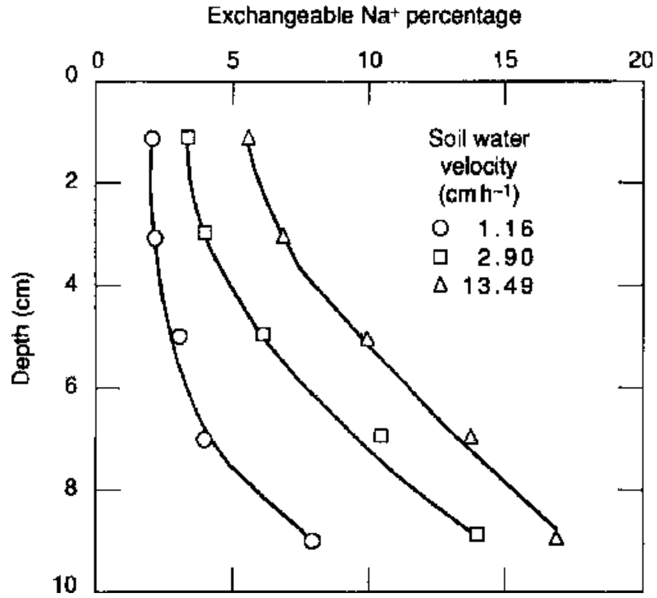


**Figure 4** Distribution of exchangeable sodium percentage with soil depth as a function of soil water velocity for a given amount of dissolved gypsum. (Adapted from Keren R and O'Connor GA (1982) Gypsum dissolution and sodic soil reclamation as affected by water flow velocity. *Soil Science Society of America Journal* 46: 726–732, with permission.)

Several chromatographic models have been used for simulating reclamation of sodic soils with gypsum. The models include many of the known chemical reactions occurring in natural soils as a result of $CaCO_3$ and $CaSO_4$ dissolution, salt leaching, and ion exchange reactions. These models have been field-tested with data from sodic-soil reclamation leaching plots. The difference between measured and computed results is no more than the horizontal variations typically found in salt-affected lands. These models provide powerful tools for quantitative predictions of water and gypsum required to reclaim soil profiles to prescribed levels of salinity and ESP.

In another approach, a source term for gypsum dissolution has been introduced into a steady-state solute-transport model. However, considerable difficulties have been encountered in matching calculated concentration–time curves to column experimental data. Miscible displacement models coupled with chemical reaction models provide insight into the processes of reclamation. However, they have not been used extensively in the field, because solving them requires extensive data specific to the site.

It is clear from the literature that the assumption of chemical equilibrium for dissolution of gypsum may not be applicable to soil-reclamation modeling. Another approach is to couple dissolution kinetics of gypsum and exchange-cation reactions to a salt-transport model.

**Acids and sulfur** These amendments are of value for ameliorating soils containing $CaCO_3$, with which the acid interacts to form gypsum (using $H_2SO_4$) or calcium chloride (using HCl). In some areas, the availability of acid waste products arising from mining and industrial activities is increasing markedly. Use of these waste products as soil amendments may provide a safe means of disposal and could greatly influence the economics of any amendment proposal. Being highly corrosive, neither $H_2SO_4$ nor $SO_2$ should be added to water that flows through metal or concrete irrigation systems. Sulfur (S) requires an initial phase of microbiological oxidation to produce $H_2SO_4$. Other amendments such as pyrites and polysulfides, which must first be oxidized by soil microorganisms, will be slower-acting than $H_2SO_4$. Their effectiveness in field experiments has been variable, perhaps being related to the presence or absence of appropriate microbial populations.

### Reclamation without Adding Soil Amendments

**Calcium carbonate** Reclamation of a sodic soil without amendments is possible when a soil source of Ca exists (e.g., calcium carbonate), the drainage through the soil profile is good, and adequate leaching water is present. The rate at which $CaCO_3$ dissolution in water approaches equilibrium is dependent upon a number of factors, which include the surface area-to-solution volume ratio, the ionic composition of the solution, the ion composition of the adsorbed phase, affinity of the clay minerals to cations, the temperature, and the partial pressure of $CO_2$ in the soil atmosphere. In general, a supersaturation with respect to calcite is observed in solution extracts from calcareous soils.

In soils, the kinetics of $CaCO_3$ dissolution is not a simple diffusion-controlled or first-order reaction. In these calcareous systems, the transfer of atmospheric $CO_2$ to solution is a limiting step in the kinetics of dissolution. There are several uncertainties in modeling the kinetics of carbonate chemistry: at low pH the rate depends significantly on the thermodynamic transport constant for $H^+$, which is not well defined, and reaction-site density and controls on pH at the surface interface between the crystal and the bulk solution are also not well understood. Moreover, kinetics of precipitation and dissolution of $CaCO_3$ have not yet been modeled for soil water systems.

Soil $CaCO_3$ may be dissolved slowly to contribute Ca but its solubility is enhanced in saline sodic soils. Moreover, incorporation of fresh organic matter in calcareous soil increases native calcium carbonate mineral dissolution, owing to elevation of $CO_2$ partial pressure in the soil atmosphere, and therefore significantly enhances sodic soil reclamation rate (**Figure 3**).

**Successive dilutions of highly saline water** When soil physical conditions have deteriorated and HC of the soil is so low that the time required for reclamation or the amount of amendment required is excessive or if a sodic soil is to be leached with a water so low in salinity that water infiltration decreases adversely, the use of the high-salt water-dilution method may be applied. The method involves application of successive dilutions of highly saline water containing divalent cations. This principle has been tested successfully in experiments, and reclamation at various stages of leaching can now be predicted. In the early phase, the high salinity of the water prevents clay dispersion and induces flocculation of the soil colloids, the Ca content provides a source of Ca for exchange with Na, and the sodicity level decreases. On dilution with high-quality water, the sodium adsorption ratio of the irrigation water is reduced by the square root of the dilution factor. When sufficient water is applied at each step of dilution, then, to ensure exchange equilibrium to the desired depth, the total equivalent surface depth of water for reclamation should be about nine times the depth of soil to be reclaimed. The method is particularly effective for soils of the expanding-clay type, which have extremely low soil HC.

## Further Reading

Ayers RS and Westcot DW (1985) *Water Quality for Agriculture*. FAO Irrigation and Drainage paper No. 29. Rome, Italy: UN FAO.

Bresler E, McNeal BL, and Carter DL (1982) *Saline and Sodic Soils: Principles, Dynamics Modeling*. Berlin, Germany: Springer-Verlag.

Hillel D (2000) *Salinity Management for Sustainable Irrigation*. Washington, DC: The World Bank.

Hoffman GJ (1980) Guidelines for the reclamation of salt-affected soils. In: O'Connor GA (ed.) *Second Inter-American Conference on Salinity and Water Management Technology*, pp. 49–64. New Mexico, NM, USA: New Mexico State University.

Jury WA, Jarrell WM, and Devitt D (1979) Reclamation of saline-sodic soils by leaching. *Soil Science Society of America Journal* 43: 1100–1106.

Keren R and Miyamoto S (1990) Reclamation of saline, sodic and boron-affected soils. In: Tanji KK (ed.) *Agricultural Salinity Assessment and Management*, pp. 410–431. New York: American Society of Civil Engineering.

Keren R and O'Connor GA (1982) Gypsum dissolution and sodic soil reclamation as affected by water flow velocity. *Soil Science Society of America Journal* 46: 726–732.

Keren R and Shainberg I (1981) The efficiency of industrial and mined gypsum in reclamation of a sodic soil – rate of dissolution. *Soil Science Society of America Journal* 45: 103–107.

Keren R, Shainberg I, Frenkel H, and Kalo Y (1983) The effect of exchangeable sodium and gypsum on surface runoff from loess soil. *Soil Science Society of America Journal* 47: 1001–1004.

Misopolinos ND (1985) A new concept for reclaiming sodic soils with high-salt water. *Soil Science* 140: 69–74.

Nielsen DR, Bigger JW, and Luthin JN (1966) Desalinization of soils under controlled unsaturated flow conditions. In: *Sixth International Congress Communication on Irrigation and Drainage*, pp. 19.15–19.24. New Delhi, India.

Oster JD (1982) Gypsum usage in irrigated agriculture: a review. *Fertilizer Research* 3: 73–89.

Quirk JP (1971) Chemistry of saline soil sand their physical properties. In: Talsma T and Philip JR (eds) *Salinity and Water Use*. New York: Macmillan.

Reeve RC and Doering EJ (1966) The high-salt-water dilution method for reclaiming sodic soils. *Soil Science Society of America Journal* 30: 498–504.

Robbins CW (1986) Sodic calcareous soil reclamation as affected by different amendments and crops. *Agronomy Journal* 78: 916–920.

Shainberg I and Shalhevet J (1984) *Soil Salinity Under Irrigation – Processes and Management*. Berlin, Germany: Springer-Verlag.

Sumner M (1999) *Handbook of Soil Science*. Boca Raton, FL: CRC Press.

Tanji KK and Deverel SJ (1984) Simulation modeling for reclamation of sodic soils. In: Shainberg I and Shalhevet J (eds) *Soil Salinity Under Irrigation – Processes and Management*, pp. 238–251. Berlin, Germany: Springer-Verlag.

# SAND DUNES

**H Tsoar**, Ben Gurion University of the Negev, Beer Sheva, Israel

## Introduction

Sand dunes are found in all continents and in all world climates. They cover approximately 20% of the world's arid and extremely arid zones but are also found in humid regions of North America and Europe and along most of the coasts of the world. In total they cover an area of more than 5 000 000 km$^2$, of which 99% is desert. Sand dunes appear in many different shapes as a result of the factors that form them, such as variability in wind direction, grain size, and vegetation cover.

As a soil, dune sand is usually considered to be devoid of any beneficial characteristics for flora and fauna because sand is known to have low fertility, low water-retaining capacity, and sand-sized particles are cohesionless. Wind magnitudes of 4 Beaufort force (5–7 m s$^{-1}$) and above can easily erode sand, thus causing another limitation for flora and fauna. As a result, aeolian sand encroachment has become synonymous with desiccation and desertification in arid and semiarid lands.

## Dune Sand Texture

Soil particles are grouped according to their sizes into four classes: gravel, sand, silt, and clay. Sedimentologists and geomorphologists use a grade scale consisting of a series of classes in which the boundaries between adjacent size classes differ by a factor of 2. This scale is modified by transforming the grade boundaries into a phi ($\varphi$) scale (Table 1), which is defined as:

$$\varphi = -\log_2 d \qquad [1]$$

where $d$ is the grain diameter in millimeters.

Wind should be above the threshold velocity to initiate sand transport. The threshold velocity depends directly on grain size and the cohesiveness of the grains (Figure 1). Grain particles less than 0.10 mm tend to cohere, as the rate of cohesiveness increases with decreasing size. Accordingly, particles of fine-sand size have threshold velocities of erosion that are lower than those of silt and clay by factors of 2–3. The mode size of most dune sand all over the world is in the class of fine sand (0.125–0.250 mm), which has the lowest wind-threshold velocities. Fine sand can be transported most easily on the windward slope of sand dunes under storm winds of average intensity. Coarser grains of medium, coarse, and very coarse sand are unable to climb slopes. This is the main reason for the formation in deserts of sand sheets composed of bimodal sand in which the coarser mode is always greater in grain size than fine sand.

The cohesionless nature of sand causes it to be swept away easily by the wind in a leaping movement, known as 'saltation,' which describes a ballistic path very close to the bed, descending at high speeds and low angles. Less-exposed grains or coarser grains are either pushed in a short-distance movement, ensuing from the scatter of the impact of the saltating grains, known as 'reptation,' or driven forward along the surface in a creep mode through the impact of the saltating grains.

The low cohesiveness and the ensuing form of transportation by saltation have a major implication for plants. The easy erosion of sand by the wind exposes plant roots, causing damage to the plants through grain bombardment, thus making the plants vulnerable to disease. Some plants overcome this by growing minute roots that produce a sticky liquid. Sand particles are glued to these roots, forming a shield which protects them from damage.

The most common adaptational response of animals to low cohesiveness of sand is to increase the surface area of the soles and fingers, either by enlarging or adding scales to the fingers (as do many lizards and skinks), or by growing stiff hairs on the sole (as do many mammals, particularly rodents and carnivores). The other technique used to overcome low cohesiveness of soil is sand-swimming, engaged in by animals such as skinks and some insects, which have relatively long and narrow bodies and short legs that help them to move while winding their bodies between the sand particles.

**Table 1** Size classes of grain particles

| *Grain size (mm)* | *Grain size (μm)* | *Grain size (φ)* | *Classification* |
|---|---|---|---|
| >2 | >2000 | <−1 | Gravels |
| 1–2 | 1000–2000 | 0 to −1 | Very coarse sand |
| 1/2–1 | 500–1000 | 1 to 0 | Coarse sand |
| 1/4–1/2 | 250–500 | 2 to 1 | Medium sand |
| 1/8–1/4 | 125–250 | 3 to 2 | Fine sand |
| 1/16–1/8 | 62–125 | 4 to 3 | Very fine sand |
| 1/256–1/16 | 4–62 | 8 to 4 | Silt |
| 1/4096–1/256 | 0.243–4 | 12 to 8 | Clay |

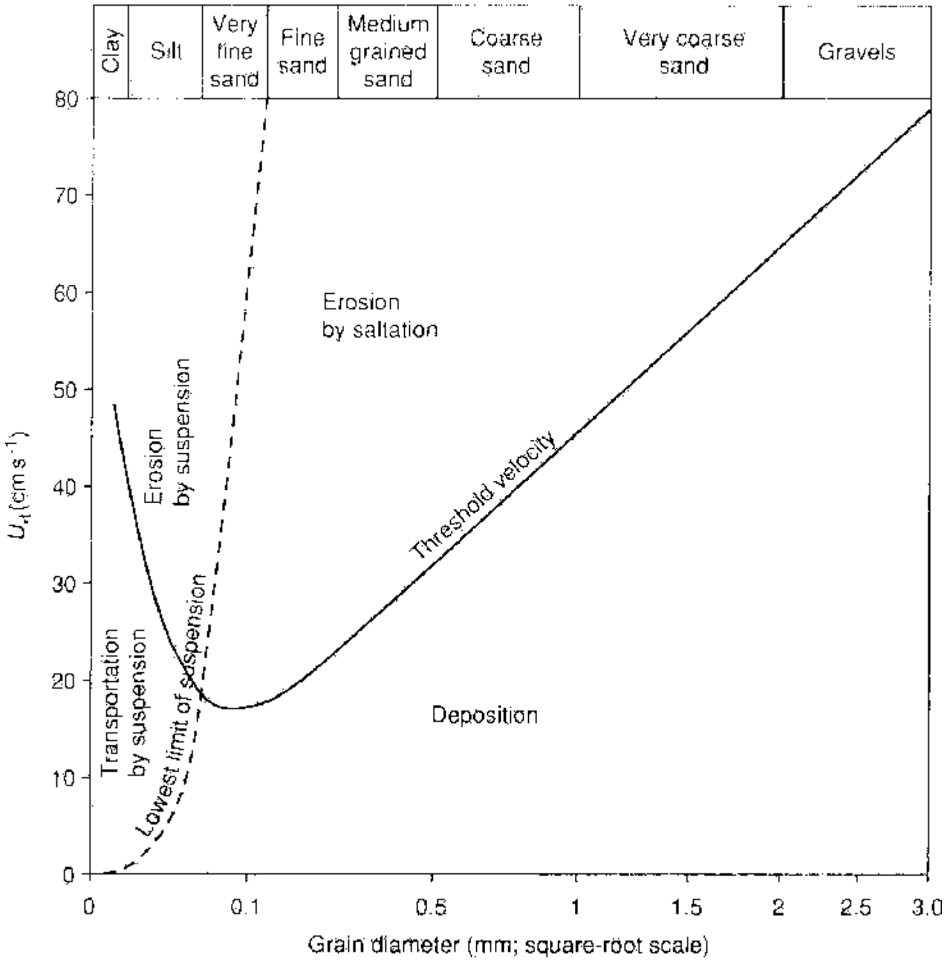


**Figure 1** Threshold friction velocity ($U_{*t}$) curve for quartz grains of different diameters (continuous line). The dashed line separates saltation from suspension. Adapted from Pye K and Tsoar H (1990) Aeolian Sand and Sand Dunes, London: Unwin Hyman.

Particle sizes of clay, silt, and very fine sand are so small that they have low settling velocities relative to the velocities of the vertical component of the turbulent wind flow. As a result, these particles remain in suspension or quasisuspension in their transportation mode and are considered as dust (**Figure 1**). Dust does not easily accumulate together with sand unless the latter is covered by vegetation, thus serving as a trap for the suspended particle load in the lower atmosphere. Biogenic crust is formed on sand dunes in some desert areas when vegetation spreads and reduces the sand movement. Filamentous cyanobacteria are known to form crusts in the northwestern Negev sand dunes (**Figure 2**). This crust serves as trap for atmospheric dust, resulting in a feedback mechanism between the amount of dust that is deposited on the surface and the mass of the cyanobacteria. There is a significant correlation in the northwestern Negev between the percentage of fines (silt + clay) and the percentage of organic matter in the crust. For approximately 1.5% of organic matter, there are more than 30% of fines. Rainwash leaches part of the fines to a depth of 30–50 cm, which is the average penetration depth of rain for small, frequent storms in the Negev desert.

## The Effect of Dune Sand Texture on Water Infiltration

One of the most important attributes of sand dunes, resulting from the sand texture and the relatively big pore spaces, is low field capacity and a low value for soil moisture availability for plants (**Figure 3**), as well as a high rate of infiltration and leaching, resulting in the washing away of nutrient elements necessary for plant growth.

The hydraulic conductivity at saturation for sand dunes is very high (between $10^{-3}$ and $10^{-4}\,\mathrm{m\,s^{-1}}$). Unlike most soils composed of silt and clay, the hydraulic conductivity of sand responds very little to changes in moisture content. The high rate of hydraulic conductivity determines the depth of water infiltration. The maximal depth of infiltration in linear sand dunes in the northwestern Negev, during a year with rainfall that is 10.5% lower than the climatological mean (95 mm), reached 60–90 cm,

**Figure 2** Crust on sand dunes in the northwestern Negev desert, where the annual rainfall is 85 mm. The crust is formed by filamentous cyanobacteria. Note an area with exposed sand in the middle lower part of the picture.

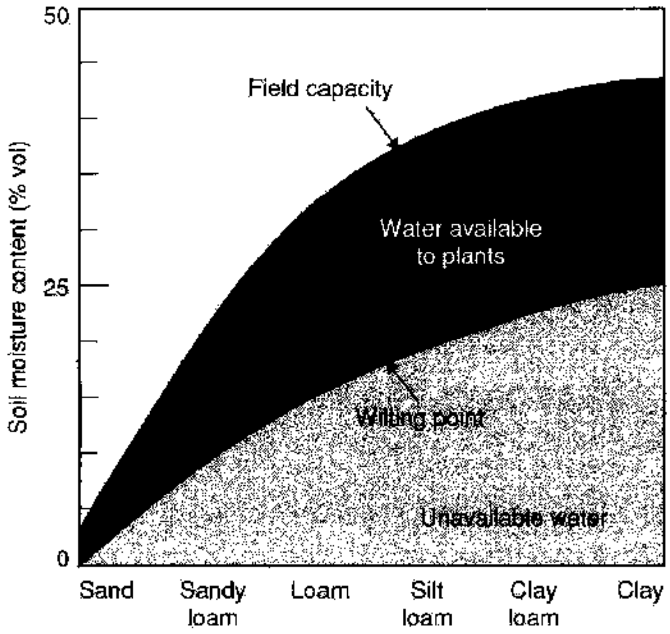


**Figure 3** The effect of soil texture on the wilting point, field capacity, and the emanative moisture that is available to plants. Note that sand has the lowest wilting point but also a very low field capacity, which leaves very little amount of moisture available to plants as most of the water infiltrates easily towards the groundwater.

while, in a nearby interdune area covered mostly by silt and clay, the infiltration was limited to a depth of 30–40 cm. During a wetter year, with rainfall 38% higher than the average, the range of infiltration in the same linear dune was in the range of 90–120 cm and in some cases down to 180–210 cm. In the fine-grained interdune layer, it was limited to a depth of 30–60 cm. Deep infiltration to a depth of 180–400 cm occurs in sand only in response to rain that accumulates to 100 mm in less than 2 months.

Sand dunes are known to have an internal laminar structure of varying geometry. The laminae result from two main modes of deposition: grainflow and grainfall. These two types of deposition produce different arrangements of grain packing and distinct concentrations of fine and coarse sand. The internal structure of aeolian sand dunes affects the direction of infiltration into the dune. If the aeolian sand is composed of grainflow deposits that dip at angles of 32–34°, infiltration will deviate by 50–60° from vertical downward movement ('piston infiltration'). The capillary tension caused by closely packed or fine-grained laminae that are near-horizontal or parallel to the surface inclination brings about subsurface lateral water flow. Such a flow can be found in sand dune areas where the mean annual rainfall is more than 200 mm. Areas with approximately 100 mm annual rainfall experience lateral movement only during periods with rainfall much above the average.

The biogenic crust formed on top of ungrazed sand dunes in the western Negev (**Figure 2**) has a lower hydraulic conductivity than dune sand because of the high proportion of fines. However, the biogenic crust on the Negev sand dunes is not hydrophobic and therefore absorbs rainwater until saturation is reached and the infiltration rate drops to 9–12 $\mathrm{mm\,h^{-1}}$. Since the prevailing sealing process of bare crust by the impact of intensive raindrops is not effective in the Negev, there is no correlation of runoff yield with rain amount or intensity. Runoff is generated in the Negev on saturated crust by rain intensity of 9–12 mm $h^{-1}$ lasting for 1–2 min only, and its yields are low to very low.

## The Effect of Sand Texture on the Thermal Properties of Sand Dunes

When sand is devoid of or has only meager vegetative cover, its texture and mineralogical composition play an important role in determining the thermal properties of its environment. Heat is absorbed from solar shortwave radiation and is concentrated in the sand's surface to a thickness of approximately 1 mm.

The rate at which modification of temperature takes place within sand ($k$) is directly proportionate to its thermal conductivity ($K$) or to its ability to transmit heat, but is inversely proportionate to its specific heat ($c$) or the amount of heat required to achieve the modification of temperature. This relationship is described in the following equation:

$$k = K/c\gamma \qquad [2]$$

where $k$ is known as the thermal diffusivity, and $\gamma$ is the bulk density of the sand (approx. 1600 kg $m^{-3}$). The units of $k$ are meters squared per second. The smaller the value of $k$, the more slowly heat will diffuse through the substance. The thermal diffusivity of sand is approx. $3.75 \times 10^{-7}$ $m^2 s^{-1}$, which is relatively low and induces a very slow penetration of surface temperature changes. This results in a very acute and thin near-surface accumulation of heat. As a consequence, the maximal sand surface temperatures are higher by 10–20°C than the air temperature, at 2 m high, during the afternoon (**Figure 4**). This steep lapse rate gives rise to local up-rushes of hot air and descent of cooler air, which increases the air turbulence, and the erosivity of the wind. Under certain conditions the ascending hot air develops into whirlwinds known as 'dust devils.' The minimal sand surface temperature is only lower by 1–3°C than the minimal air temperature in the early morning (**Figure 4**).

These thermal qualities of sand result in three significant, diurnal changes in temperature at deeper levels of dune sand: (1) the maximal temperature is lessened with depth; (2) the diurnal amplitude is degraded with depth; (3) there is a lag between the daily maximum and minimum temperature from the surface to various depths (**Figure 4**).

The unique thermal qualities of sand dunes have important implications for animals living in this habitat. The most common animal adaptation to high surface temperatures is avoidance: most animals living in sand dune habitats in deserts are nocturnal or are active during the early morning or late evening, when soil temperature decreases to tolerable temperatures. During the rest of the day, most of these animals take shelter below the sand surface, where the temperatures are moderate, either in their own burrows or in holes dug by other animals. Many of the diurnal animals have relatively long legs (lizards, beetles) and can elevate themselves high above the hot surface, thus keeping their bodies at relatively low temperatures while only some parts of their legs are in contact with the hot sand.

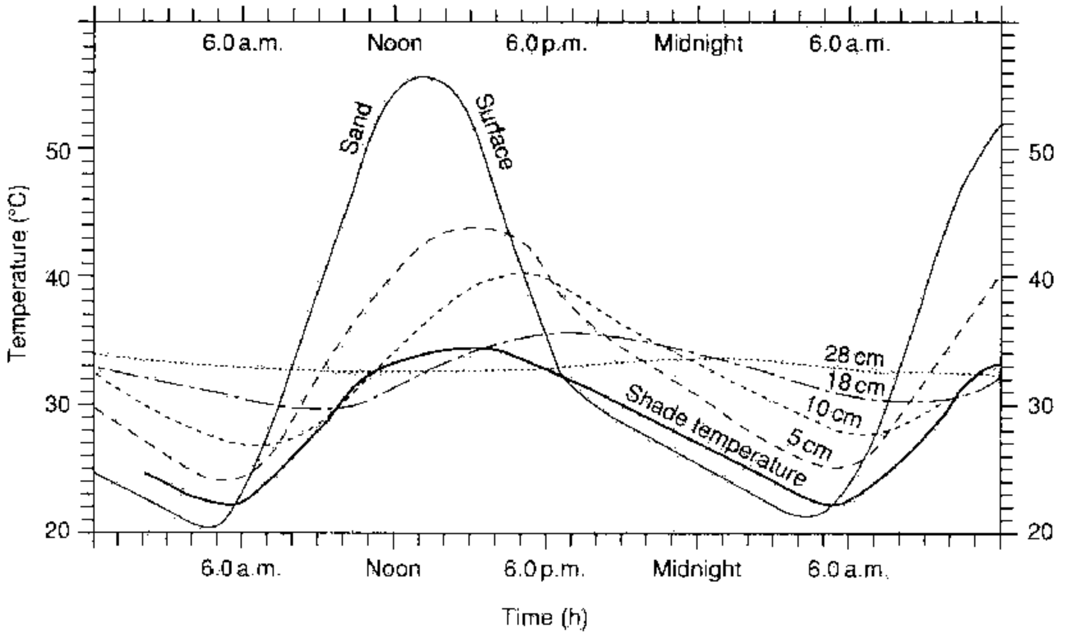


**Figure 4** Variation in sand temperature with depth as measured in August. Shade temperature was measured at an elevation of 2 m above the surface. Adapted from Pye K and Tsoar H (1990) Aeolian Sand and Sand Dunes. London: Unwin Hyman.

## The Advantage and Disadvantage of Sand Texture on Vegetation Growth in Deserts

Vegetation can grow on sand dunes in arid areas with less than 100 mm of annual mean rainfall (Figure 2). The primary limitation of vegetation growth on dune sand is human impact; the most dominant natural limitation is wind power. Rainfall is a limiting factor only where the annual mean is very low (<50 mm).

Two unique and distinct attributes of dune sand affect vegetation growth on dunes: (1) dune sand has low fertility and low water-retaining capacity (**Figure** 3); (2) sand grains are not cohesive and therefore are easily carried by the wind (**Figure** 1). For this reason, plants on sand are sparsely distributed and they are absent where wind erosion is severe, such as the windward slopes of dunes. Accordingly, the amount of vegetation on dunes depends on the wind power. There are arid and humid areas where the wind power is so high that it does not enable vegetation to thrive on these dynamic dunes, such as the Oregon coastal dunes.

The attribute of high hydraulic conductivity of sand relative to silt and clay is advantageous in deserts. In humid areas with abundant precipitation, the main cause of soil-moisture loss in sand is the high rate of infiltration beyond the root layer. In arid and semiarid lands, the situation is quite different. These areas have little precipitation, resulting in infiltration to depths of only 60–120 cm. Perennials can use the moisture at these depths, which are deep enough to protect the sand moisture from the high rate of evaporation typical of deserts during the long dry and hot rainless periods. As a result, moisture retention is assured in desert sand below the upper 30–60 cm. In other desert soils such as loess, which are mostly composed of silt and clay, the infiltration reaches a much shallower depth of 30–40 cm, where evaporation is effective. The delineation between the advantage of coarse-textured sand (low evaporation) over fine-textured soil (silt and clay) versus its inherent disadvantage (high hydraulic conductivity) lies somewhere between a mean of 300 and 500 mm of precipitation.

The moisture content of sand at the wilting point is very low. Therefore, a small amount of water input brings sand moisture to values much above the wilting point, whereas the same amount of water added to loamy soils does not, making the water unavailable for plants (**Figure** 3).

The effect of soil texture on biomass has been tested on the three soil types (loam, sandy loam, and sand) located in the same area in the Negev desert where the annual mean rainfall is 100 mm. Biomass measurements (annuals only) were taken during two winters with approximately average rainfall (104.6 and 111 mm), but with dissimilar distribution. The highest biomass for the 2 years was formed on the sand soil (647.18 g m$^{-2}$), with the second highest on the sandy loam (569.95 g m$^{-2}$), while the loam had the least amount of biomass (115.53 g m$^{-2}$).

## Mobility and Stability of Sand Dunes

There is no direct relationship between the amount of rainfall and the vegetation cover on the dunes. Active dunes with no vegetation cover are found in humid areas (e.g., the Oregon coastal dunes, or the Alexandria coastal dunes in South Africa) while dunes fully stabilized by microphytes and macrophytes are found in the arid part of the Negev desert. Several wind erosion or mobility indices ($M$) have been developed for various parts of the world for the purpose of climatic changes and their implications on sand dune mobility. All of these indices are based on two factors that increase or decrease dune mobility. The first one deals with the degree of windiness. Most dunes will be mobilized if windiness were to increase. The second factor is the ratio between the mean precipitation ($P$) and the evaporation ($E$), known as precipitation-effectiveness index ($P-E$). Based on the above factors, a simple and widely accepted $M$-index is:

$$M = \frac{W}{(P/\mathrm{PET})} \qquad [3]$$

where $W$ is the percentage of days during the year with sand-moving winds and $P$/PET is the ratio of mean annual precipitation to mean annual potential evapotranpsiration. The $M$-index of eqn [3] is calibrated in South Africa. Dunes are fully active when values for $M$ exceed 200 and most dunes are completely stabilized by vegetation when $M$ is less than 50. Because of its simplicity, the $M$-index of eqn [3] is commonly used by scientists studying the effect of climate change on sand mobility and stability in North America. The sand dunes along the coastal plain of Israel and in the northwestern Negev desert should be active with no vegetation cover according to eqn [3] ($M > 200$). In fact, the northwestern Negev desert dunes in areas of annual mean rain of 100 mm are fully stabilized by vegetation and biogenic crust (**Figure** 2), and the more humid coastal dunes are partially vegetated.

Dune sand has the lowest moisture content (**Figure** 3). Thus, chiefly in humid climates, most of the rain is lost to the groundwater and is not available to the plants. Hence the amount of rainfall is not a decisive factor in sand dune stabilization except by increasing the cohesiveness of the sand. The deep infiltration of rainwater in dune sand reduces the effect of evaporation (PET in eqn [3]) mostly in arid areas relative to other soils composed of fine particles of silt and clay.

Wind power is the most important factor in sand dune mobility because of the noncohesiveness of the sand. Wind above a certain wind power can erode sand to such an extent that it prevents seeds from germinating in the sand and stabilizing it. According to the sand transport equation, the sand flux ($q$) is directly proportional to the cube of the wind:

$$q = 5.2 \times 10^{-4}(U - U_t)^3 \qquad [4]$$

where $q$ is in metric tons per meter width per hour, $U$ is the wind velocity, measured at a height of 1 m, and $U$ is the threshold wind velocity, which is equal to $4\,\mathrm{m\,s^{-1}}$. According to eqn [4], 1 h of wind velocity at $16\,\mathrm{m\,s^{-1}}$ can carry the same amount of sand as 72 days with wind velocity of $5\,\mathrm{m\,s^{-1}}$, or 27 h of wind velocity at $8\,\mathrm{m\,s^{-1}}$. Hence, the factor $W$ of eqn [3] that refers to the wind by the percentage of days during the year with sand-moving winds is a poor representation of the sand-moving power of the wind.

A much better index of wind power is the drift potential (DP) of the wind, which is based on eqn [4]:

$$\mathrm{DP} = \Sigma q = \frac{U^2(U - U_t)}{100}t \qquad [5]$$

where $\sum q$ is related to the total sand flux from all wind directions, and $t$ is the amount of time the wind blows above the threshold (in percent). The drift potential (DP) is given in vector units and it is customary to base the calculations on wind velocities measured in knots. An index of wind direction variability is illustrated by the ratio between the resultant drift potential and the drift potential (RDP/DP), where values close to one indicate narrow unidirectional DP, and values close to zero indicate multidirectional DP.

The total yearly average DP and the ratio RDP/DP can explain the mobility and stability of sand dunes mostly because the limiting factor for vegetation on dune sand is wind erosion. When RDP/DP is low, wind energy is distributed on more than one slope of the dune and the energy exerted on each slope is lower. Sand dunes in areas where the annual mean rainfall is ≥50 mm are unvegetated and mobile under wind conditions in which $M$ (according to eqn [6]) exceeds 1:

$$M = \frac{\mathrm{DP}}{1000 - (750(\mathrm{RDP/DP}))} \qquad [6]$$

Table 2 displays the DP, RDP, and $M$ (according to eqn [6]) in various dune sites of the world. It shows that rainfall is not as decisive a factor for dune mobilization or stabilization as is the DP. In the Negev desert, dunes are fully stabilized where the mean yearly rainfall is 90 mm. The dunes in Sinai (east of Al-Ismailiya) and the coastal dunes in Gaza are fully active due to human impact. Dunes that were artificially stabilized, such as along the Netherlands coast, will become active once vegetation is artificially destroyed.

## Dune Shape and Dynamics

The tremendous variety of sand dunes in the world deserts makes their classification a difficult task. Sand accumulates into piles of various shapes because of three main causes:

1. Obstruction to the wind flow such as a cliff, boulder, etc.
2. Vegetation cover;
3. Self accumulation of sand into piles.

Classification according to the above three factors is depicted in **Figure 5**. The first two causes result from the immobilization of sand by vegetation or topographic obstructions. The last one is related to aerodynamics. In addition, other factors, such as

**Table 2** Drift potential (DP), directional variability of the wind (resultant drift potential (RDP/DP), $M$-index (eqn [6]), and rainfall of several dune field sites

| *Location* | *Country* | *DP* | *RDP/DP* | *Average yearly rainfall (mm)* | M | *Dunes status* |
|---|---|---|---|---|---|---|
| Al Ismailiya | Egypt (western Sinai) | 62 | 0.47 | 50 | 0.10 | Fully active |
| Nizzana | Israel (Negev) | 108 | 0.70 | 90 | 0.23 | Fixed (fully vegetated) |
| Gaza | Palestine (coast) | 158 | 0.73 | 400 | 0.35 | Active (partly vegetated) |
| Ashdod | Israel (coast) | 147 | 0.71 | 500 | 0.31 | Semiactive (partly vegetated) |
| Upington | South Africa (Kalahari) | 520 | 0.66 | 183 | 1.03 | Stabilized linear dunes |
| Port Elizabeth | South Africa (coast) | 951 | 0.49 | 660 | 1.5 | Fully active with no vegetation |
| Lüderitz | Namibia (coast) | 2300 | 0.85 | <100 | 6.34 | Fully active with no vegetation |
| Ijmuiden | Netherlands (beach) | 3999 | 0.51 | 768 | 6.48 | Fully stabilized by vegetation; active when vegetation is destroyed |

Reproduced with permission from Tsoar H (2001) Types of aeolian sand dunes and their formation. In: Balmforth NJ and Provenzale A (eds.) Geomorphological Fluid Mechanics. Lecture Notes in Physics. Berlin: Springer.

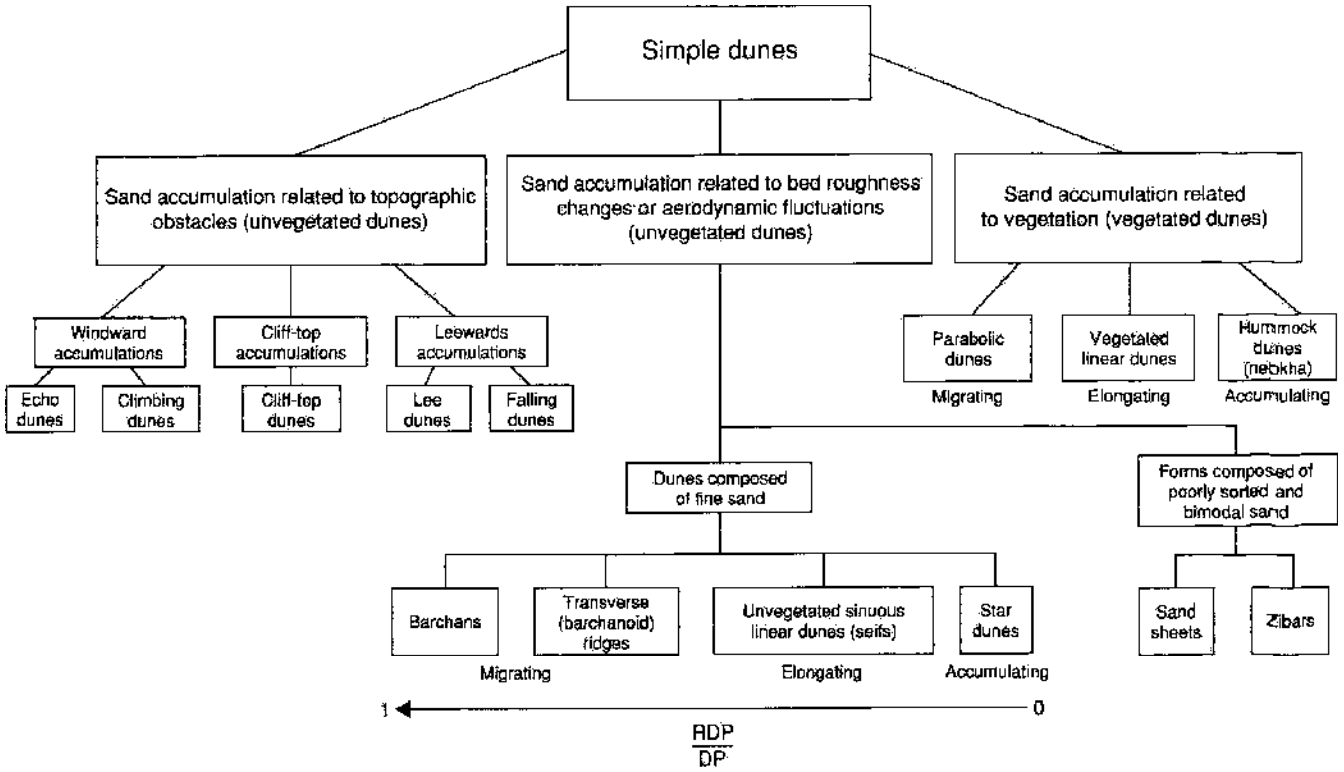


**Figure 5** Classification of major dune types based on the way sand is accumulated into dunes and the effect of wind direction variability. RDP, resultant drift potential; DP, drift potential. Adapted from Pye K and Tsoar H (1990) Aeolian Sand and Sand Dunes. London: Unwin Hyman.

variation in wind direction, climatic changes manifested by dramatic shift in wind direction and magnitude, velocity, and frequency of storms, sand availability and thickness of sand cover, and sudden removal of vegetation cover, can have an effect on dune morphology.

From a dynamic point of view, sand dunes can be classified into three distinct groups:

1. Migrating dunes, where the whole dune body advances with little or no change in shape and dimension; best represented by transverse and barchan dunes.
2. Elongating dunes, where the dunes elongate and become extended in length with time; best represented by linear dunes.
3. Accumulating dunes, where the dunes have little or no net advance or elongation; best represented by star dunes.

These three groups are distinguished according to the wind directional variability for dunes without vegetation (**Figure 5**). The migrating dunes are formed by a wind regime that is unidirectional or close to unidirectional. The wind directions of the elongating dunes are bidirectional with a 70–90° separation between the two directions. The accumulating dunes are formed under multidirectional winds, where the two main directions form an obtuse angle that is approximately 180°.

The migrating and elongating dunes are under the influence of two different types of sand movement and deposition. Migrating dunes such as barchan and transverse dunes have two distinct slopes, a moderate windward and a steeper leeward. The windward slope undergoes erosion and the sand is carried to the lee slope where it is deposited on the upper part of the slope. When the upper lee slope reaches the critical angle of internal friction, a failure begins. The slide of grains from the upper slope downward reduces the slope inclination and the avalanche ceases at the angle of repose (33°), forming a slip-face (**Figure 6**).

Elongating dunes have a more complicated mechanism of sand transport and deposition because, when winds encounter the dune crest line at an acute angle, the flow at the lee side is deflected to parallel the crest line. Accordingly, the leeward slope is not merely a zone of deposition, but also a zone of erosion by the diverted lee flow. The occurrence of erosion and transport or deposition of sand depends upon the angle of incidence between the wind and the crest line. When this angle is <40°, the deflected wind has the power to erode sand and carry it along the lee slope, with little deposition on the upper lee slope. When the angle of incidence is close to 90°, the velocity of the wind in the lee slope drops and deposition largely occurs as it does with migrating dunes. For angles between 40° and 90°, there is a gradual transition in the lee slope, from dominant erosion

**Figure 6** A field of barchans in the Namibian desert looking upwind. Note the avalanche tongues on the slip-face. Reproduced with permission from Tsoar H (2001) Types of aeolian sand dunes and their formation. In: Balmforth NJ and Provenzale A (eds.) Geomorphological Fluid Mechanics. Lecture Notes in Physics. Berlin: Springer.

**Figure 7** Oblique aerial picture of seif dunes. Note the dunes' tortuosity and the sharp crest-line. Reproduced with permission from Tsoar H (2001) Types of aeolian sand dunes and their formation. In: Balmforth NJ and Provenzale A (eds.) Geomorphological Fluid Mechanics. Lecture Notes in Physics. Berlin: Springer.

**Figure 8** Aerial photograph of star dunes in Erg Oriental in Algeria. Note the radiating arms from the dune central peak. Reproduced with permission from Tsoar H (2001) Types of aeolian sand dunes and their formation. In: Balmforth NJ and Provenzale A (eds.) Geomorphological Fluid Mechanics. Lecture Notes in Physics. Berlin: Springer.

and transport to deposition. For that reason the linear seif dunes tend to meander (**Figure 7**) and consequently change the angle of incidence and the function of the lee side from an erosional to a depositional zone.

Star dunes as representative types of the accumulating dunes are recognized by their distinctive shape of sinuous arms radiating from a central pyramidal peak (**Figure 8**). Both mechanisms of migration and elongation act upon the star dunes from several directions. Star dunes are formed in a wind regime with high directional variability (RDP/DP $< 0.4$), and for that reason they are found mainly in the high latitudes of deserts with marked seasonal changes in wind direction.

The transverse, barchan, seif, and star dunes are formed without the presence of vegetation. There are also several typical vegetated dune types, which are located in arid and humid areas (mostly along the coasts).

Isolated clumps of vegetation act as sand traps and thus lead to the formation of nebkhas (coppice dunes) – hummocks that can reach up to 30 m high and 100 m across (**Figure 9**). The most effective pioneer plant forming nebkhas is *Ammophila arenaria*, which grows in clumps and can grow up through an accumulation of sand.

The most commonly found vegetated sand ridges on sandy backshores are foredunes that are formed where pioneer vegetation can grow and trap aeolian sand. Foredunes develop into continuous vegetated ridges, which lie parallel to coastlines exposed to onshore, sand-transporting winds (**Figure 10**). The foredune is a static dune that can only grow upward, and the only dune type that involves the exchange of sand with the beach. Other coastal dunes are mostly transgressive types (barchan, transverse, and parabolic dunes), formed when sand is transferred inland where foredunes are absent, or through blowouts (wind-excavated gaps through which sand is transported landward) in the foredune ridges.

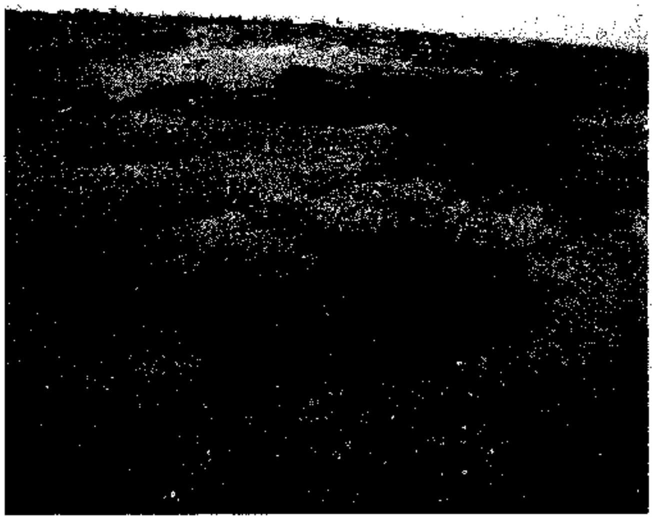


**Figure 9** Nebkha (coppice dune) formed by vegetation that traps sand.

Parabolic dunes, as their name implies, are U-shaped, with vegetated arms pointing upwind. Parabolic dunes can be active and transgressive or fully stabilized and inactive (**Figure 11**). The mechanism of parabolic dune formation in coastal humid areas is due to the fact that vegetation is more easily established at the base of the dune near the water table. Vegetation or dampness along the lower sides of the dune retards sand motion and both are considered to be anchors. Vegetation in parabolic dune formation is said to protect the less-mobile arms against the force of the wind action, thereby allowing the central part to advance downwind. In this way the advancing apex leaves behind trailing ridges that elongate and turn the dune into a hairpin form.

It has recently been found that barchan and transverse dunes turn into parabolic dunes in areas of low wind power where human impact is reduced. According to the profile of barchan or transverse dunes, erosion on the windward slope of the dune diminishes gradually toward the crest, which is an area

**Figure 10** Foredune of approximately 5 m height formed on the backshore, where vegetation can thrive.

**Figure 11** Parabolic (U-dune) formed by the effect of vegetation on sand transport.

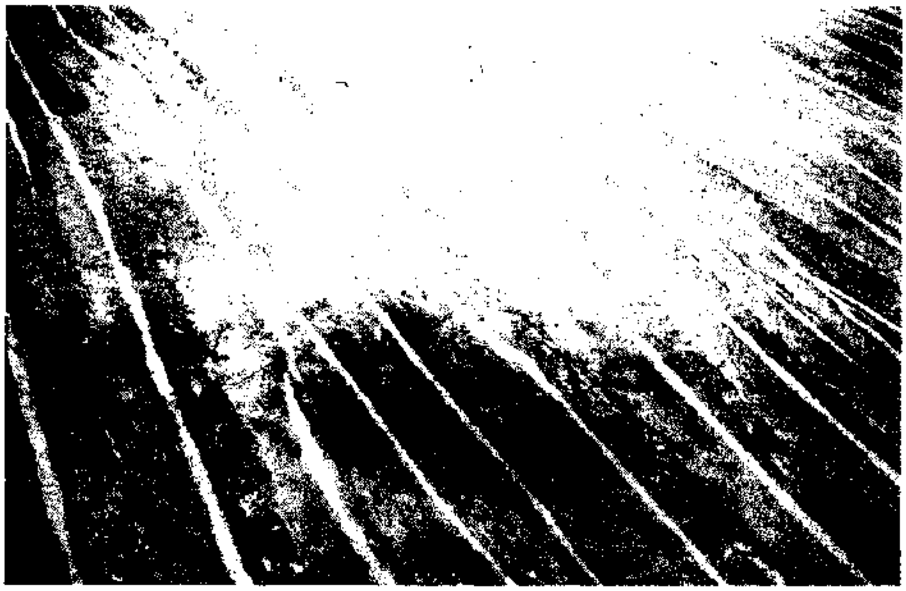


**Figure 12** Aerial photograph of vegetated linear dunes. This linear dune elongates in the direction of the dominant, strong wind. Note the straight alignment of the dune (different from the meandering seif of **Figure 7**) and the Y-junctions, which indicate wind direction from lower left to upper right.

of neither erosion nor deposition. Hence, once human impact stops, vegetation will recover on the barchan crest, thereby starting the process of transformation of these dunes into parabolic ones.

Vegetated-linear dunes belong to the group of elongating sand dunes found in many deserts of the world (Australian deserts, the Kalahari, Indian deserts, and the Negev). Vegetated-linear dunes are low with rounded profiles (**Figure 12**). They range in height from a few meters up to dozens of meters. Vegetation covers them, sometimes entirely, and sometimes abundantly on the plinth and lower slopes, but is very sparse or absent on the crest. Those that are fully covered by vegetation have become partly or wholly stabilized. Vegetated-linear dunes may run in parallel for scores of kilometers.

An exclusive attribute of vegetated-linear dunes is the tendency for two adjacent dunes to converge and continue as a single ridge. Convergence is in the form of a Y-junction (a tuning-fork shape) commonly open to formative winds. Vegetated-linear dunes are distinguished from seif dunes by: (1) coextension along the strongest dominant-wind direction; (2) the cover of vegetation; (3) straight alignment with no tortuosity; and (4) Y-junctions.

## Further Reading

Bagnold RA (1941) *The Physics of Blown Sand and Desert Dunes*. London, UK: Methuen.

Danin A (1996) *Plants of Desert Dunes*. Berlin, Germany: Springer-Verlag.

McKee ED (1979) *A Study of Global Sand Seas*. Washington, DC: US Government Printing Office.

Pye K and Tsoar H (1990) *Aeolian Sand and Sand Dunes*. London, UK: Unwin Hyman.

**Saturated and Unsaturated Flow** *See* **Hydrodynamics in Soils**; **Vadose Zone**: Hydrologic Processes

# SCALING

Contents

## Physical Properties and Processes

**G Sposito**, University of California–Berkeley, Berkeley, CA, USA

### Introduction

The physical properties of field soils exhibit a high degree of spatial variability that complicates any attempt to predict their temporal evolution using conventional models based on nonlinear partial-differential equations. This complexity can be reduced by assuming that the varying local values of soil physical properties in a field are connected to one another by arbitrary scale factors, while the form of the governing partial-differential equation remains the same everywhere in the field ('scale invariance'). This assumption turns out to require soil hydraulic properties to be power-law functions of volumetric wetness (water content) and, in particular, it requires a power-law relation between hydraulic conductivity and matric potential. These power-law relationships have often been observed in controlled field experiments, indicating the applicability of scale invariance and opening the way for fractal models of soil physical properties. However, types of relationship other than power-law have also frequently been observed, pointing to the limitation of scaling as a means of reducing the complexity of spatial variability.

### Scale Invariance

The physical properties of natural soils are inexorably variable when observed from place to place. Volumetric wetness – to choose perhaps the simplest example – can vary substantially over the span of an agricultural field, even if only the surface horizon of the field soil is examined and care has been taken to provide a uniform input of water prior to examination. The complexity that this kind of variability would bring to any attempt to understand and predict the temporal evolution of volumetric wetness across an entire field is evident, and the same conclusion follows for other important soil physical properties. (*See* **Spatial Variation, Soil Properties.**)

Now suppose that a simplifying hypothesis is made: that the underlying cause of spatial variability in soil physical properties has to do with the vicissitudes of soil texture or structure, but not with the fundamental mechanisms by which the physical properties may change in response to inputs of matter and energy. If this supposition is true, then the underlying physical laws that govern the behavior of these properties must be uniform across a field, even if the properties themselves are not. Moreover, functional relationships between the properties (e.g., the relationship of volumetric wetness to matric potential) should also be uniform, in the sense that each such relationship is quantified by the same parametric equation everywhere in a field, albeit with different (and highly varying) values of the parameters at different places.

That spatial uniformity in the physical laws describing the behavior of soil properties implies a similar uniformity in the functional relationships between these properties may not seem obvious, but this conclusion follows once the physical laws are expressed mathematically in the form of nonlinear partial-differential equations. Fundamentally, imposing the condition of spatial uniformity on a governing partial-differential equation places a strong constraint on the relationships between its dependent variable and the nonconstant coefficients that appear in the equation, with the result that these relationships necessarily must also be uniform.

To make these ideas concrete, let us consider the well-known, nonlinear partial-differential equation that governs isothermal, isohaline, isochoric infiltration processes in a uniform pedon (i.e., the Richards equation in the 'mass picture'):

$$\partial\theta/\partial t = \partial[D(\theta)\partial\theta/\partial z]/\partial z - (\partial K/\partial\theta)(\partial\theta/\partial z) \qquad [1]$$

where $\theta$ ($L^3 L^{-3}$) is volumetric wetness, $t(T)$ is time, $z(L)$ is vertical position, $D(\theta)$ ($L^2T^{-1}$) is water diffusivity under a gradient in $\theta$, and $K(\theta)$ ($LT^{-1}$) is hydraulic conductivity. Eqn [1] describes the temporal evolution of volumetric wetness in terms of two empirical transport coefficients, $D(\theta)$ and $K(\theta)$, whose relationship to $\theta$ must be specified in order to apply the equation in a predictive manner. Eqn [1] could be used to model infiltration in an instrumented field plot, with the expectation that $D(\theta)$ and $K(\theta)$ would have widely varying values across an array of such plots, even at the same value of $\theta$. The hypothesis of uniformity requires that eqn [1] apply to observations of $\theta(t, z)$ anywhere in the array, irrespective of the high spatial variability. This latter supposition, in turn, requires that $D(\theta)$ and $K(\theta)$ each preserve the same functional relationship to $\theta$ anywhere in the field.

The uniformity hypothesis being imposed on eqn [1] is expressed formally by defining three arbitrary parameters that connect two sets of its dependent and independent variables evaluated at any two places where the equation applies:

$$\theta' = \mu\theta, \quad t' = \delta t, \quad z' = \alpha z \qquad [2]$$

where $\mu$, $\delta$, and $\alpha$ are (positive-valued) scale factors. These scale factors relate $\theta'(t', z')$ in one place to $\theta(t, z)$ in another, the premise being that, since the governing physical law is the same at both places, all of the spatial variability in the volumetric wetness across a field can be captured quantitatively solely by changes in the scales ('magnification' or 'reduction') over which the dependent and independent variables are observed. To be consistent with this premise, eqn [1] must still hold at the 'prime' location, i.e., after it is transformed by multiplying $\theta$ on both sides of the equation by $\mu$, then changing $t$ to $t'$ on the left side and $z$ to $z'$ on the right side, using eqn [2]:

$$\begin{aligned}\partial\theta'/\partial t' =& (\alpha^2/\delta)\partial[D(\theta)\partial\theta'/\partial z']/\partial z' \\ &- (\mu\alpha/\delta)(\partial K/\partial\theta')(\partial\theta'/\partial z')\end{aligned} \qquad [3]$$

Eqn [3] will have the same mathematical form as eqn [2] only if the scale transformations:

$$\begin{aligned}D'(\theta') &\equiv D(\mu\theta) = (\alpha^2/\delta)D(\theta), \\ K'(\theta') &\equiv K(\mu\theta) = (\mu\alpha/\delta)K(\theta)\end{aligned} \qquad [4]$$

are also imposed on the two transport coefficients. Eqn [4] connects values of the water diffusivity and hydraulic conductivity at any two places. When they hold, the governing partial differential equation for $\theta(t, z)$ will have the same mathematical form under arbitrary scaling anywhere in a field. In this sense, the equation (and the physical behavior which underlies it) is said to be 'scale-invariant.' It is thus immune to the vagaries of spatial variability, these being epitomized instead in the three scale factors (eqn [2]) and the required scale transformations of the transport coefficients (eqn [4]).

## Scaling Relationships

It is apparent that eqn [4] cannot be true for just any functional dependence of the two transport coefficients on $\theta$, unless the scale factor $\mu$ always happens to have the value 1.0. In other words, if the transport coefficients do not change compatibly with the scaling of eqn [1], then scale invariance of the equation is vacuous. This critical issue can be investigated rigorously with methods in the mathematical theory of Lie groups. This investigation shows that $D(\theta)$ and $K(\theta)$ must have the following functional forms in order that eqn [4] be satisfied:

$$D(\theta) = a(\theta + b)^{m} \qquad [5]$$

$$K(\theta) = [ca^{\beta}/(1 + m\beta)](\theta + b)^{1+m\beta} \qquad [6]$$

where $a$, $b$, $c$ are constants to be determined by the units of $D$ and $K$ and the datum of $\theta$, while $m$ and $\beta$ are empirical parameters related to the scale factors in eqn [2]. Their relation to the scale factors emerges clearly after imposing eqn [4] on eqns [5] and [6]:

$$D'(\theta') \equiv a(\mu\theta + \mu b)^{m} = \mu^{m}D(\theta) = (\alpha^2/\delta)D(\theta) \qquad [7]$$

$$\begin{aligned}K'(\theta') &\equiv [ca^{\beta}/(1 + m\beta)](\mu\theta + \mu b)^{1+m\beta} \\ &= \mu^{1+m\beta}K(\theta) = (\mu\alpha/\delta)K(\theta)\end{aligned} \qquad [8]$$

(In eqns [7] and [8], the arbitrary constant $b$ is assumed to be determined by the datum for $\theta$ and, therefore, to scale as $\theta$ does.)

It follows that:

$$(\alpha^2/\delta)^{1/m} = (\alpha/\delta)^{1/m\beta} \qquad [9]$$

and, therefore:

$$\delta = \alpha^{(2\beta-1)/(\beta-1)} \qquad [10]$$

Either eqn [9] or [10] then yields the relation:

$$\mu = \alpha^{1/m(1-\beta)} \qquad [11]$$

for the scale factor of $\theta$. Thus, only one of the three scale factors can be chosen independently once the values of $m$ and $\beta$ have been measured. (Note also that $m \neq 0$, $m\beta \neq -1$, must be finite and $\beta \neq 1$ or 1/2. These conditions are required in order that eqns [6], [9], and [10] be determinate.)

The physical significance of the scale factor $\alpha$ can be appreciated more incisively after using the definition of the water diffusivity:

$$D(\theta) \equiv K(\theta)\frac{\partial\psi}{\partial\theta} \qquad [12]$$

along with eqns [5] and [6] to derive an equation for the matric potential $\psi(\theta)(L)$ that is consistent with scale invariance:

$$\psi(\theta) = [(1 + m\beta)/cm(1 - \beta)](\theta + b)^{m(1-\beta)} \qquad [13]$$

The corresponding equation for $\psi'(\theta') \equiv \psi(\mu\theta)$ then follows from eqn [11]:

$$\begin{aligned}\psi'(\theta') &= [(1 + m\beta)/cm(1-\beta)](\mu\theta + \mu b)^{m(1-\beta)} \\ &= \mu^{m(1-\beta)}\psi(\theta) = \alpha\psi(\theta)\end{aligned} \qquad [14]$$

The scale factor $\alpha$ is found to transform the matric potential from one place to another. Similarly, the transformation of the hydraulic conductivity follows from eqns [4], [10], and [11] as:

$$K'(\theta') = \alpha^{-\eta}K(\theta) \qquad [15]$$

where

$$\eta \equiv (1 + m\beta)/m(\beta - 1) \qquad [16]$$

The right side of eqn [15] is more conventionally expressed in terms of the square of a scale factor $\omega$, defined by:

$$\omega \equiv \alpha^{2/\eta} \qquad [17]$$

such that the scale transformation of $K(\theta)$ takes the form:

$$K'(\theta') = \frac{K(\theta)}{\omega^2} \qquad [18]$$

Eqns [14] and [18] are known as 'Warrick–Nielsen scaling relationships,' and the exponent $\eta$ that relates them in eqn [17] is the 'Russo–Jury exponent.'

A further connection between eqns [6] and [13] and conventional practice in soil physics can be made after defining the datum constant $b$ to equal $-\theta_r$, where $\theta_r$ is the 'residual volumetric wetness,' a measure of the volume of 'immobile water' that does not participate in flow on the time scale over which eqn [1] is applied. Adding to this the conventional definitions:

$$\psi_s \equiv \psi(\theta_s), \qquad K_s \equiv K(\theta_s) \qquad [19]$$

where $\theta_s$ is porosity, one can rewrite eqns [6] and [13] in a more convenient power-law form:

$$\psi(S) = \psi_s S^{m(1-\beta)} \qquad [20]$$

$$K(S) = K_s S^{1+m\beta} \qquad [21]$$

where

$$S \equiv (\theta - \theta_r)/(\theta_s - \theta_r) \qquad [22]$$

is the degree of saturation. Eqns [20] and [21] indicate that this scale-invariant variable is a natural one with which to express the relationship between either matric potential or hydraulic conductivity and volumetric wetness. If $S$ is eliminated between eqns [20] and [21], hydraulic conductivity can be connected directly to matric potential:

$$K(\psi) = K_s(\psi_s/\psi)^{\eta} \qquad [23]$$

upon noting eqn [16]. The Russo–Jury exponent is now seen as the parameter that determines the power-law dependence of hydraulic conductivity on matric potential, as implied by eqn [17].

### Example: Hamra Soil Field Experiment

Eqns [20–23] can be illustrated by their application to a 0.8-ha field comprising an Alfisol (Hamra series) on which 30 instrumented plots were established. **Table 1** lists field-wide mean values and coefficients of variation (CVs) for $\psi_s$, $K_s$, $m$, and $\beta$ as deduced from measurements of $\psi(\theta)$ and $K(\theta)$ made at four depths in the soil profile. Notable in the table are the typical decline in $K_s$ with depth and the relatively large CVs (mean divided by standard deviation) for

**Table 1** Mean values and coefficients of variation (CV) of the parameters in eqns [20–23] as measured in the Hamra field soil near Bet Dagan, Israel[a]

| | $\psi_s$ (m) | | $K_s$ ($10^{-5}$ ms$^{-1}$) | | m | | β | |
|---|---|---|---|---|---|---|---|---|
| *Depth (m)* | *Mean* | *CV* | *Mean* | *CV* | *Mean* | *CV* | *Mean* | *CV* |
| 0.0 | −0.072 | 0.22 | 6.12 | 0.41 | 1.86 | 0.68 | 1.46 | 0.68 |
| 0.3 | −0.072 | 0.23 | 4.47 | 0.36 | 2.50 | 0.65 | 1.60 | 0.65 |
| 0.6 | −0.074 | 0.21 | 2.17 | 0.65 | 3.19 | 0.82 | 1.69 | 0.82 |
| 0.9 | −0.082 | 0.27 | 1.98 | 0.75 | 3.35 | 0.81 | 1.70 | 0.81 |
| 0.0–0.9 | −0.074 | 0.24 | 3.66 | 0.69 | 2.48 | 0.88 | 1.60 | 0.88 |

$m$, $\beta$, empirical parameters related to scale factors in eqn [2].

$\psi_s$ and $K_s$, 'air-entry' matric potential and saturated hydraulic conductivity, respectively.

[a] Adapted from Russo D and Bresler E (1981) Soil hydraulic properties as stochastic processes. *Soil Science Society of America Journal* 45: 682–687

this parameter and the two exponents $m$ and $\beta$. The lesser spatial variability of $\psi_s$ parallels that observed for $\theta_s$ (all-depth field-wide mean of 0.37, with CV = 0.11). The field-wide mean values of the Russo–Jury exponent also tended to decline with depth: 4.3 ± 1.6, 3.3 ± 0.9, 2.9 ± 0.8, 2.9 ± 0.7; the all-depth mean being 3.3, with CV = 0.36. Other field studies show that $\eta$ usually lies in the interval (2.0, 4.5) with a typical CV in the range 0.2–0.5.

The complete data set for the Hamra soil has been used to compute values of $\psi(S)$ and $K(S)$ at fixed values of $S$ in the interval (0.750, 0.975) based on eqns [20] and [21]. The resulting all-depth mean values and CVs of the dimensionless ratios, $\psi(S)/\psi_s$ and $K(S)/K_s$, are listed in **Table 2** for 10 values of the degree of saturation $S$. The very large CVs at smaller values of $S$ are immediately apparent. It is this extreme spatial variability that the scale factors $\alpha$ (eqn [14]) and $\omega$ (eqn [18]) are designed to mediate. In particular, it should be possible to substitute the all-depth mean values of $\psi(S)/\psi_s$ and $K(S)/K_s$ into the left side of eqns [14] and [18] in order to calculate $\alpha$ and $\omega$ anywhere in the field (subject to imposing an all-depth mean value of 1.0 for each scale factor, since only the ratios $\psi(S)/\psi_s$ and $K(S)/K_s$ are used). The result of such a calculation is summarized in **Table 3**, which indicates that $\alpha$ and $\omega$ have virtually identical CVs over the field and that the field-wide mean value of $\eta$ linking them (eqn [17]) is the same as found from examining values of the matric potential and hydraulic conductivity directly.

**Table 2** All-depth, field-wide mean values and coefficients of variation (CV) for the relative matric potential ($\psi(S)/\psi_s$) and hydraulic conductivity ($K(S)/K_s$) in the Hamra soil[a]

| S | $\psi(S)/\psi_s$ | *CV* | $K(S)/K_s$ | *CV* |
|---|---|---|---|---|
| 0.750 | 25 | 3.9 | 0.14 | 0.92 |
| 0.775 | 13 | 3.3 | 0.17 | 0.86 |
| 0.800 | 7.7 | 2.8 | 0.20 | 0.80 |
| 0.825 | 4.8 | 2.2 | 0.24 | 0.74 |
| 0.850 | 3.3 | 1.6 | 0.28 | 0.68 |
| 0.875 | 2.4 | 1.2 | 0.34 | 0.61 |
| 0.900 | 1.9 | 0.8 | 0.41 | 0.53 |
| 0.925 | 1.5 | 0.5 | 0.50 | 0.45 |
| 0.950 | 1.3 | 0.3 | 0.61 | 0.34 |
| 0.975 | 1.1 | 0.13 | 0.77 | 0.20 |

[a] Adapted from Jury WA, Russo D, and Sposito G (1987) The spatial variability of water and solute transport properties in unsaturated soil. II. Scaling models of water transport. *Hilgardia* 55: 33–36.

**Table 3** All-depth, field-wide mean values and coefficients of variation (CV) for the scale factors $\alpha$ and $\omega$ along with the Russo–Jury exponent (eqn [17])[a]

| *Parameter* | *Mean* | *CV* |
|---|---|---|
| $\alpha$ | 1.00[b] | 0.435 |
| $\omega$ | 1.00[b] | 0.446 |
| $\eta$ | 3.36 | 0.354 |

[a] Adapted from Jury WA, Russo D, and Sposito G (1987) The spatial variability of water and solute transport properties in unsaturated soil. II. Scaling models of water transport. *Hilgardia* 55: 33–56.

[b] Unit normalization imposed.

## Physical Significance of Scaling

Now let us step back to take stock of what the scaling of soil physical properties and processes actually means. The basic premise is that fundamental physical laws governing the behavior of the properties are scale-invariant. The example taken was the Richards equation for infiltration (eqn [1]), which describes volumetric wetness changes when water moves downward in a soil profile whose water diffusivity and hydraulic conductivity are known. Scale invariance of this equation implies that it is equally valid from place to place in a field, with uniform functional relationships between its two transport coefficients and the volumetric wetness. These latter three variables may fluctuate wildly across a field, but they are none the less connected by scale factors that relate their values between any two places. Their spatial variability is thus epitomized solely by differences in the local scales of volumetric wetness, position, and time (eqn [2]) that prevail at different locations, or if the degree of saturation (eqn [22]) is chosen as the dependent variable in the Richards equation, solely by differences in the scales of position and time that attend water movement at different locations in a field. Evidently these differences of scale arise from localized soil textural or structural characteristics, although the precise cause of spatial variability is left unspecified.

Implicit in the scale invariance of the Richards equation is the fact that this equation must also apply on average to an entire field of soil:

$$\partial S/\partial t = \partial[D_m(S)\partial S/\partial z]/\partial z - (\partial K_m/\partial S)(\partial S/\partial z) \qquad [24]$$

where $D_m(S)$ and $K_m(S)$ are field-wide mean values of the transport coefficients. Eqn [24] is ultimate testimony to the hypothesis that the mechanism of infiltration is the same irrespective of the scale of observation. Indeed, the dependent variable in eqn [24] is scale-invariant, and the mean transport coefficients are related to local values at any place 'P' through the Warrick–Nielsen scaling relationships:

$$\alpha_{\mathrm{P}}^{1} D_{\mathrm{P}}(S)/\omega_{\mathrm{P}}^{2} = D_m(S) \qquad [25a]$$

$$K_{\mathrm{P}}(S)/\omega_{\mathrm{P}}^{2} = K_m(S) \qquad [25b]$$

with a corresponding equation for matric potential:

$$\alpha_P \psi_P(S) = \psi_m(S) \qquad [25c]$$

(eqn [25a] follows from eqns [12], [14], and [18].)

In field experiments like the one performed on the Hamra soil, $\psi_m$ and $K_m$ are first calculated, then they are used along with local values of the matric potential and hydraulic conductivity to estimate a set of $\alpha_P$ and $\omega_P$, thus giving pragmatic substance to the reality of eqn [24]. Indeed, using this approach, one need only solve the mean Richards equation for a chosen boundary-value problem and have in hand a set of scale factors $\{\alpha_P, \omega_P\}$ in order to apply the solution at every point in a field – clearly an advantage in predictive applications of the flow equation. However, this computational efficiency can only be realized if there is truly no scale dependence of the mechanism of water flow, a matter for continuing field experimentation to evaluate.

The compatibility conditions in eqns [5], [6], and [13] ensure that the scale invariance of the Richards equation is not a vacuous hypothesis. The matric potential and the two transport coefficients are required to be power-law functions of the volumetric wetness or, equivalently, the degree of saturation. This power-law form has been widely used to describe the functional relationship between matric potential and volumetric wetness in field soils, under which application it is termed the 'Brooks–Corey model':

$$S = (\psi_s/\psi)^\lambda \qquad [26]$$

where the Brooks–Corey exponent $\lambda = 1/m(\beta - 1)$ according to eqn [20]. Representative values of the parameters $\psi_s$ and $\lambda$ have been calculated using hydraulic data on hundreds of soils according to texture. These data show that the 'air-entry value' $\psi_s$ tends to increase, while the Brooks–Corey exponent $\lambda$ tends to decrease, as the clay content of a soil becomes larger. Eqn [26] can also be derived from a fractal model of soil pore space in which the pore-size distribution is scale-invariant and $\lambda \equiv 3 - D$, where $D < 3$ is the fractal dimension that characterizes the degree of 'lacunarity' of a soil, i.e., the extent to which its pores are space-filling at the expense of the solid matrix. Given the typical range of values observed for the Brooks–Corey exponent ($0.1 < \lambda < 0.7$), it follows that $2.3 < D < 2.9$, with larger fractal dimensions corresponding to clayey soils. Thus, in clayey soils, 'lacunarity,' or the degree to which pores are space-filling, is expected to be larger than in sandy soils, in agreement with the respective porosities of these soils.

A power-law relation is also expected between hydraulic conductivity and volumetric wetness or, more generally, between hydraulic conductivity and matric potential (eqn [23]), if the Richards equation is scale-invariant. This relationship is often observed, but other models of $K(\psi)$ are abundant in the literature of soil physics. For example, the Gardner model postulates an exponential relationship instead of a power-law relationship for $K(\psi)$:

$$K(\psi) = K_s \exp\left[-\frac{\psi_s - \psi}{D_s/K_s}\right] \qquad [27]$$

If this equation accurately describes the behavior of hydraulic conductivity in a soil, the Richards equation cannot be scale-invariant over a field containing that soil. Indeed, application of the scaling relations in eqns [14] and [15] to both sides of eqn [27] is impossible. The reason for the failure is signaled by the presence of the ratio, $D_s/K_s$, in eqn [27]. This ratio has the dimensions of length and serves as a measure of how quickly $K(\psi)$ declines as $\psi$ decreases below the air-entry value in a 'Gardner soil.' It is, therefore, an intrinsic length scale in the Gardner model, i.e., a length scale that is determined uniquely by the values of $D_s$ and $K_s$ at each separate location in a field soil. In terms of conventional water-desorption methods of measuring the cumulative pore-size distribution in a soil (by combining data on $\psi(S)$ with the Young–Laplace equation to calculate the effective pore radius as a function of the degree of saturation), the length scale $D_s/K_s$ is interpreted as that of the very first pores to empty when $\psi$ drops below the air-entry value, $\psi_s$. (*See* **Porosity and Pore-Size Distribution**.) The existence of such an intrinsic length scale stands in direct contradiction to the use of an arbitrary scale factor for relating matric potential at two places in a field (eqn [14]). Therefore, in a Gardner soil, the arbitrary scaling of matric potential (and, correspondingly, of all the variables in eqn [1]) is not possible because the intrinsic length scale $D_s/K_s$ preempts that possibility. This example of broken scaling symmetry is valuable, in that it illuminates by contrast the underlying significance of eqns [14–18]: that the flow equation operates everywhere in a field soil without prejudice as to the local scales of length, time, or water content.

*See also:* **Infiltration**; **Porosity and Pore-Size Distribution**; **Spatial Variation, Soil Properties**

## Further Reading

Anderson AN, McBratney AB, and Crawford JW (1998) Applications of fractals to soil studies. *Advances in Agronomy* 63: 1–76.

Baveye P, Parlange J-Y, and Stewart B (1998) *Fractals in Soil Science*. Boca Raton: CRC Press.

Crawford JW and Matsui N (1996) Heterogeneity of the pore and solid volume of soil. *Geoderma* 73: 183–195.

Crawford JW and Sleeman BD (1998) From particles to architecture: fractals, aggregation and scaling in soil science. In: Huang PM, Senesi N, and Buffle J (eds) *Structure and Surface Reactions of Soil Particles*, pp. 41–75. Chichester, UK: John Wiley.

Edwards MP (1994) Classical symmetry reductions of nonlinear diffusion-convection equations. *Physics Letters A* 190: 149–154.

Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Hillel D and Elrick DE (eds) (1990) *Scaling in Soil Physics*. Madison, WI: Soil Science Society of America.

Jury WA, Russo D, and Sposito G (1987) The spatial variability of water and solute transport properties in unsaturated soil. II. Scaling models of water transport. *Hilgardia* 55: 33–56.

Millán H and Orellana R (2001) Mass fractal dimensions of soil aggregates from different depths of a compacted vertisol. *Geoderma* 101: 65–76.

Pachepsky YaA, Crawford JW, and Rawls WJ (eds) (2000) *Fractals in Soil Science*. Amsterdam, the Netherlands: Elsevier.

Perrier E, Rieu M, Sposito G, and de Marsily G (1996) Models of the water retention curve for soils with a fractal pore size distribution. *Water Resources Research* 32: 3025–3031.

Rawls WJ and Brakensiek DL (1995) Utilizing fractal principles for predicting soil hydraulic properties. *Journal of Soil and Water Conservation* 50: 463–465.

Rieu M and Sposito G (1991) Fractal fragmentation, soil porosity, and soil water properties. *Soil Science Society of America Journal* 55: 1231–1244.

Russo D and Bresler E (1981) Soil hydraulic properties as stochastic processes. *Soil Science Society of America Journal* 45: 682–687.

Sposito G (ed.) (1998) *Scale Dependence and Scale Invariance in Hydrology*. Cambridge, UK: Cambridge University Press.

Wiltshire R and El-Kafri M (2004) Non-classical and potential symmetry analysis of Richards equation for moisture flow in soil. *Journal of Physics A: Mathematical and General* 37: 823–839.

# Transport Processes

**R P Ewing**, Iowa State University, Ames, IA, USA

## Introduction

'Scaling of transport properties' refers to the common and often troublesome observation that the model parameters describing dispersion at one scale are frequently incorrect at another scale. Worse, sometimes one model is applicable at one scale, and a different model at another scale. The word 'scale' denotes extent, most often of space or time, so to say that dispersion scales is to say that different results are obtained from measuring it at different spatial or temporal scales. The most common situation is that one measures dispersion at a small scale and predicts it at a larger scale, then finds that the actual dispersion as measured later at the larger scale is greater than predicted. Further evidence of scaling is the observation that dispersion coefficients compiled from many separate experiments show a strong trend of dispersion increasing with distance. Is this scaling inherent to the dispersion process?

Because prediction of solute transport is generally based upon the advection–dispersion equation (ADE), which uses a diffusion-type description of transport, to say that transport scales often implies that it does not conform to a simple diffusion-like description. Scaling of dispersion is often thought to be an exceptional case, sometimes called 'anomalous dispersion,' implying that dispersion which obeys the ADE is the norm. But as more so-called anomalous transport has been found, it is increasingly accepted as both more general and widespread: 'normal' transport actually represents a special and unusual case. In this article, dispersion that conforms to the ADE will be called 'Gaussian.'

This article discusses the simple case of one-dimensional dispersion of a nonreactive, neutral, conservative solute during steady-state isothermal saturated flow, ignoring any density and viscosity effects of solute concentration. This simple case is already sufficiently difficult, but the reader must remember that real-world situations also include reactive, degrading, sorbing, and/or charged solutes moving nonisothermally and transiently under multiple driving and dispersing forces.

## Diffusive–Dispersive Transport

Hydrodynamic dispersion (called 'dispersion' here for convenience) is the result of two mechanisms: mechanical dispersion and molecular diffusion. In simplistic terms, mechanical dispersion is caused by the medium, while diffusion is inherent to the fluid. Mechanical dispersion results from differences in travel times along different streamlines. At the scale of a single pore, these differences result from flow being slower close to a pore wall and faster at pore centers. At a larger scale, streamlines differ in velocity due to heterogeneity of the medium, such as variations in pore size and connectivity, and the presence of structure across multiple scales from pore-scale correlations to large-scale layering and fractures. When flow is rapid, mechanical dispersion contributes most of the total dispersion, but at slow flow rates dispersion is predominantly diffusive.

Molecular diffusion is the random motion of fluid molecules, so diffusion of solutes is a function of the fluid (solvent and solute). The diffusion equation describes net solute flux as being proportional to, and opposite in direction from, the concentration gradient. In one dimension and in bulk fluid (no solid phase):

$$\frac{\partial C}{\partial t} = D_{\mathrm{aq}} \frac{\partial^2 C}{\partial x^2} \qquad [1]$$

where $C$ is solute concentration ($\mathrm{M\,L^{-3}}$), $t$ is time (T), $D_{\mathrm{aq}}$ is the diffusion coefficient ($\mathrm{L^2\,T^{-1}}$) of the solute in water, and $x$ is distance (L). In the presence of a solid phase, the molecular diffusion coefficient can be adjusted by the porosity $\phi$ of the medium and the tortuosity $\tau$ of the diffusive path length, giving an effective diffusion coefficient: $D_{\mathrm{eff}} = D_{\mathrm{aq}}\,\phi/\tau$. If solute is introduced at time $t_0$ in a plane at concentration $M$ per unit area, within an infinite medium having initial concentration $C_0 = 0$, then the resulting concentration at a distance $x$ from that plane at time $t$ is:

$$C(x,t) = \frac{M}{2\sqrt{\pi D t}} \cdot e^{-x^2/4Dt} \qquad [2]$$

which is a Gaussian (normally distributed) concentration profile whose width increases in proportion to the square root of time.

Dispersion is often estimated using data from an experiment in which water in the sample is replaced by water with a tracer. If the output is tested for the presence of the tracer, a steady increase in flux concentration is seen as the original water is replaced by water with the tracer. If instead the tracer's resident concentration is measured inside the sample at a single instant, a decrease in tracer concentration is seen from the inlet to the outlet end of the sample. The output of a flux concentration dispersion experiment can be summarized graphically using a breakthrough curve (BTC) (Figure 1). In these curves, the $x$-axis is presented in units of pore volumes, the amount of water needed to replace the original water in the soil sample precisely once. The $y$-axis is expressed as the ratio of the measured to the final concentration, $C/C_0$. This normalization of the two axes allows curves from different experiments (flow rates, column lengths, etc.) to be compared.

### Sir Geoffrey Taylor and the Spreading of Solute in a Tube

While diffusion is only one component of dispersion, the physicist Sir Geoffrey Taylor showed that

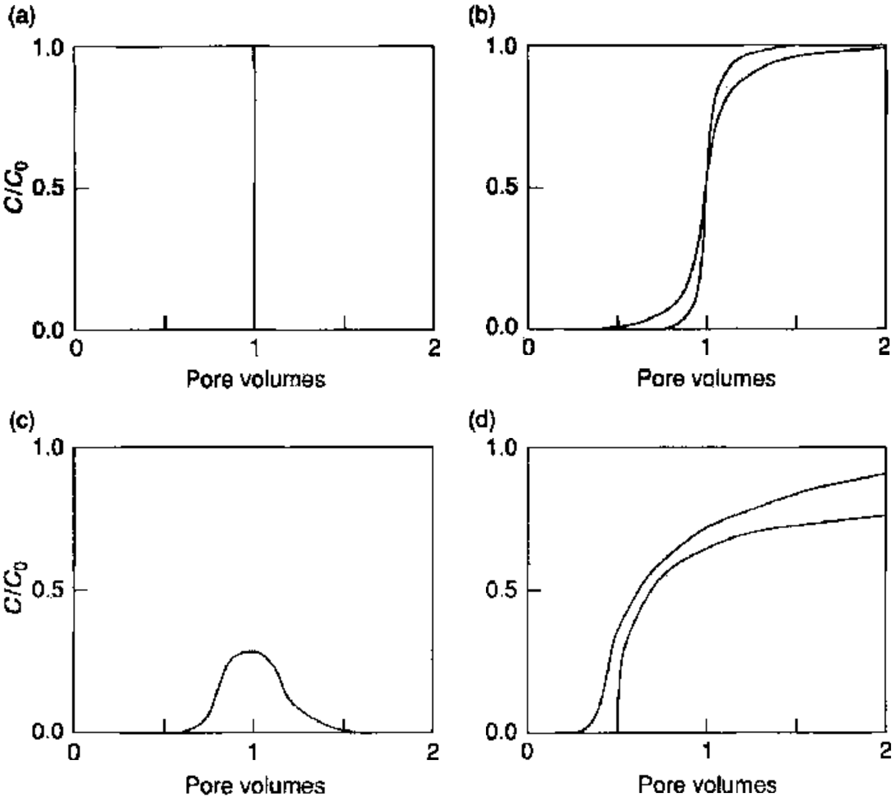


**Figure 1** Sample flux concentration breakthrough curves for: (a) a step input with piston flow; (b) a step input with Gaussian dispersion; (c) a Dirac (pulse) input with Gaussian dispersion; and (d) non-Gaussian dispersion. See text for explanation.

dispersion of a solute in fluid flowing steadily through a tube can – under certain conditions – be precisely described by superimposing the mean fluid velocity on to a diffusion-like dispersion term. The main condition is simply that flow be sufficiently slow that solute molecules have time to diffuse freely between fast (tube center) and slow (tube wall) regions, a condition that is often well approximated, at least at the pore scale, under natural conditions in soils and aquifer materials. Taylor's experimental setup – a uniform tube with steady flow – clearly does not match any naturally occurring porous media. Nonetheless, the analogy between dispersion and diffusion has historically been applied to solute movement in porous media generally. The analogy is particularly attractive because the diffusion equation has been widely studied, and operating within this analogy allows use of a wealth of mathematical solutions.

Sir Geoffrey Taylor's three seminal papers on dispersion of solutes in a tube, combined with Rutherford Aris' subsequent analysis, established much of how both engineers and soil scientists now think about dispersion. An earlier paper by Griffith had stated that a drop of dye, placed into water flowing through a glass tube, moves at the mean water velocity, with its dispersion along the tube being symmetrical. Reading Griffith's paper, Taylor was struck by its two 'remarkable' implications: first, movement with the mean water velocity implied that water in the center of the tube, flowing at twice the mean velocity, must pick up dye as it traveled toward the center of the dye slug, then release it after passing the center of the slug; and second, the longitudinally symmetrical dispersion seemed at odds with the radially asymmetrical (parabolic) velocity distribution of water in the tube. Taylor set to work to examine these phenomena both experimentally and mathematically.

Starting by assuming that there was no diffusion at all, Taylor realized that the parabolic velocity profile:

$$v(r) = v_0(1 - r^2/a^2) \qquad [3]$$

(where $v(r)$ is velocity (L T$^{-1}$) at a distance $r$ from the tube center, $v_0$ is the maximum velocity, $a$ is the tube radius (L), and $r$ is distance from the tube center (L) such that $0 \leq r \leq a$), coupled with a step input of dye, would yield the linear breakthrough curve:

$$C/C_0 = \left(1 - \frac{x}{v_0 t}\right) \qquad [4]$$

where $t$ is time (T), for $0 < x < v_0 t$. He also noticed that the diffusion coefficient $D_{\mathrm{aq}}$ (L$^2$ T$^{-1}$) was small in comparison with the observed dispersion coefficient $D$(L$^2$ T$^{-1}$): apparently longitudinal diffusion (diffusion in the direction of flow) was an insignificant contributor to dispersion. Radial diffusion (normal to the direction of flow), however, could explain the movement of solute molecules between the fast-flowing tube center and the negligible velocities near the wall.

Taylor's analysis of the interaction between radial diffusion and the parabolic velocity profile shows that the dispersion behaves exactly like diffusion, that is, it yields a Gaussian concentration profile the width of which grows in proportion to the square root of time (**Figure 2**). The theoretical dispersion coefficient $D$, which agrees well with his observations, is given by:

$$D = \frac{a^2 v_0^2}{192\, D_{\mathrm{aq}}} \qquad [5]$$

that is, dispersion is proportional to velocity squared. But this relationship only holds if two conditions are met. The first condition states that $av/D_{\mathrm{aq}} \gg \sqrt{48}$, where $v$ is the mean velocity. The quantity $av/D_{\mathrm{aq}}$ actually defines the Peclet number (the ratio of advective to diffusive transport), so this condition can be simply restated as $\mathrm{Pe} \gg 7$. Aris, noting that Taylor's $D$ explicitly excluded longitudinal diffusion, added it back into Eqn [5] to obtain:

$$D = D_{\mathrm{aq}} + \frac{a^2 v_0^2}{192\, D_{\mathrm{aq}}} \qquad [6]$$

thereby eliminating the need for this first condition. Notice that the resulting dispersion coefficient is explicitly the sum of diffusive and mechanical dispersive forces.

The second condition states that $a^2/4D_{\mathrm{aq}}t$ must be small; that is, there must be sufficient time for radial

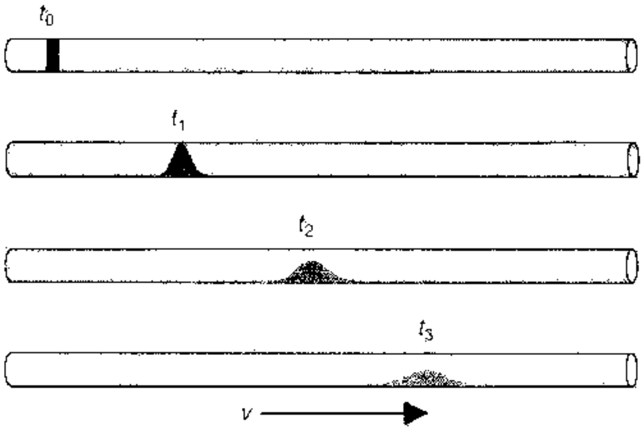


**Figure 2** How the solute flows and disperses with time in a tube into which solute is introduced at a single instant in time.

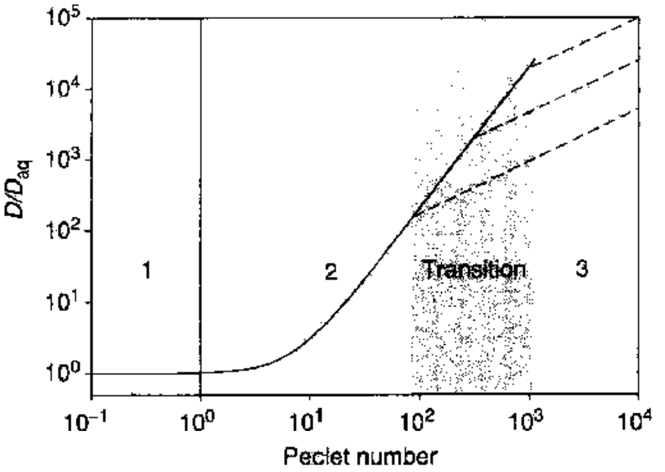


**Figure 3** Mechanical dispersion over total dispersion as a function of the Peclet number. Above some Peclet number whose exact value varies with length, dispersion transitions to being linear with time; below that, it scales with the square of time. See text for explanation.

diffusion to move molecules a distance equal to the radius of the tube. When the elapsed time is too short, radial diffusion is insignificant and the dispersion can be approximated by Eqn [4], Taylor's first analysis, based upon the assumption of no diffusion. In terms of the Peclet number, this second condition may be expressed as $\mathrm{Pe} \ll 4L/a$, where $L$ is the distance from the dye injection point at which the measurement is made.

The conditions for Taylor's analysis can be shown by plotting the ratio of total dispersion to diffusion as a function of the Peclet number (**Figure** 3). The solid line represents the Aris equation (Eqn [6]), which shows different behavior in each of three regions; the dashed lines show the behavior that would be seen if the second condition were violated. In the first region ($\mathrm{Pe} \ll 7$), diffusion is an important contributor to total dispersion and so Taylor's original analysis (Eqn [5]) does not hold, although Aris' (Eqn [6]) does. In the second region, both conditions are met and the slope of the line is approximately 2. The Peclet number above which the second condition fails, $\mathrm{Pe} \ll 4L/a$, is shown as a gray region rather than a single line because it changes for different $L$. In this third, preasymptotic region, dispersion is linearly related to velocity. Because for any given velocity and tube radius the second condition does not hold for some short time or distance, we can refer to short times and/or distances as being 'preasymptotic.'

## Solute Spreading in Soils

Superimposing a mean velocity on to the diffusion equation yields the ADE, also called the convection–dispersion equation (CDE). A one-dimensional form of the ADE for nonreactive solutes describes the change in concentration as resulting from the sum of dispersive and advective terms:

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} - v\frac{\partial C}{\partial x} \qquad [7]$$

where $D$ is the dispersion coefficient ($\mathrm{L}^2\,\mathrm{T}^{-1}$) and $v$ is the mean water velocity ($\mathrm{LT}^{-1}$). Analytical solutions to this equation have been derived for a wide variety of initial and boundary conditions, and numerous computer programs are also available. In a simple case, a point injection of a solute will move at the mean water velocity, while spreading as if it were diffusing with a diffusion coefficient equal in magnitude to the dispersion coefficient $D$. Non-Gaussian behavior is often seen at early times and/or short distances, but such situations will yield increasingly Gaussian behavior with increasing distance and/or time. Accordingly, non-Gaussian behavior is sometimes called 'preasymptotic,' implying that all dispersion eventually becomes Gaussian.

In the field, dispersion is often observed to increase linearly with fluid velocity. The CDE can in this case be generalized with respect to velocity by defining the dispersivity $\alpha(\mathrm{L})$ as $\alpha = D/v$, and replacing the dispersion coefficient $D$ in the ADE with $\alpha v$. This dependence on velocity may seem counterintuitive: if dispersion is the sum of molecular diffusion and mechanical dispersion, then mechanical dispersion (which results from the medium) should be independent of velocity, whereas slower flow (having more time) should result in greater diffusion-based dispersion. But as Taylor's studies of dispersion in a tube have shown, $D$ increases with $v$ within a specific range of Peclet numbers, so dispersion increasing with velocity is not entirely surprising. On the other hand, dispersion in soils is observed to be proportional to $v$ rather than $v^2$, so Taylor's analysis provides only a partial explanation of the whole process.

Unfortunately, the assumption that dispersion can be treated mathematically as if it were diffusion often leads to a poor match between model and data. This is frequently because one or more of the equation's assumptions have been violated: probably either the medium is not uniform (let alone tubular!) or there is insufficient exchange between solute molecules traveling on different streamlines. Because diffusion and flow are not strictly additive with respect to dispersion except under specific conditions, the basic concept of the ADE is not sufficiently general for many situations found in the real world. Much of the recent research on solute dispersion concerns struggles with the poor predictions produced by the ADE, frequently focusing on developing alternative conceptual models.

## Scale

Soil science is concerned with processes taking place from the nanometer to the kilometer scale and beyond, and from the near-instantaneous to geologic timescales. It is common to speak of scales in a colloquial manner – 'pore-scale,' 'lab-scale,' 'overnight' – and the mental images that these terms invoke can be useful, although they are not quantitative. Less useful are terms such as 'microscale,' which naturally carry widely different meanings between, and even within, different disciplines. Because soil science is inherently interdisciplinary, such words will have different meanings even to different soil scientists and should be avoided. If quantitative terms are needed, then using the size itself is best: 'meter-scale,' for example.

Some properties are specific to a defined range of scales. To illustrate, suppose one wants to measure the porosity of a sand using computed tomography (CT). If the CT unit measures with a resolution of 0.01 mm, then most voxels (volume elements from the CT scan) will represent either the inside of a solid particle or the center of a pore. In neither case will it be useful to speak of the porosity of an individual voxel: porosity is a property of a volume and becomes meaningless as that volume shrinks to a point. If we coarsen the scale slightly, each individual voxel will sample just a few grains and pores, and so individual measurements will be highly variable. But when porosity is measured over a sufficiently large number of grains and pores (at least 30 is a good guideline), the measured value is stable. This illustrates the concept of the representative elementary volume (REV), the smallest volume over which a given property can be measured with reasonable confidence. Notice that the REV must be larger than the characteristic scale of variability of the medium (if one exists).

At this point we must also make a distinction between scale and level. Scale is strictly about physical extent, while level involves a conceptual difference or a change in properties. The two are easily confused, because changes in physical size often entail changes in level also. For example, a soil core 10 cm deep may come from a single horizon, but 'scaling up' to a 20-cm core may yield a sample with two distinct horizons in it – a difference in level as well as scale. Likewise, scaling up from a 10-cm core to a pedon may introduce fractures or wormholes into the system, and so, rather than studying a soil matrix, one now faces a macropore-matrix or dual-permeability system. The distinction between scale and level may be quite subtle. For example, buoyancy effects may not be evident during fast flow of a weak salt solution through a 1-cm-diameter soil column, but they may be evident during slower flow through a larger-diameter soil column. In the second case, the increase in size (scale) allows the emergence of a previously negligible mechanism (level), so the second system under study is not equivalent to the first. It can be useful to compare relevant dimensionless groups to help discern whether changing the scale also allows new processes to come into play. Relevant dimensionless groups may include the Peclet number (ratio of advective to diffusive forces), the Bond number (ratio of buoyancy to capillary forces), the Reynolds number (ratio of inertial to viscous forces), and the Schmidt number (ratio of kinematic viscosity to molecular diffusivity).

A further difficulty is that many parameters of interest can each be best measured within a fairly narrow range of scales. For example, dispersion is generally estimated from solute concentration, based upon samples at the scale of a liter or less. Measuring solute concentration over a large area is generally accomplished by collecting many small samples, because collecting the entire outflow of a large area is generally impractical, if not impossible. So studies of dispersion generally operate at many scales: the scale of the individual observations (where different observations may themselves have different scales), the scale of the current composite (for example, the inferred pollutant plume based upon some 100 individual core samples), the scale of the desired prediction, and the characteristic scale, if any, of the soil's heterogeneity. While we are mainly concerned here with the disparity between the middle two, any difference in scale may contribute to scaling.

## Conceptual Models

Several conceptual models are in common use for predicting dispersion. The ADE, described above, is widely used in both subsurface hydrology and soil science. Conceptually equivalent to flow through a tube (**Figure 4a**), it does not adequately describe the complexities encountered in flow through soils and so frequently fails to predict dispersion correctly. One common failure is finding that the value of the dispersion coefficient increases with the distance over which the solute has moved. A second failure is a non-Gaussian output which persists across many length scales, despite the prediction of the ADE that the dispersion will asymptotically become Gaussian. A common instance of this is 'slow tailing,' in which the breakthrough curve resulting from a step input shows early arrival and a slower-than-Gaussian increase to the final concentration. This has been addressed in some models by the use of a fractional ADE (fADE), which uses a fractional form of the error

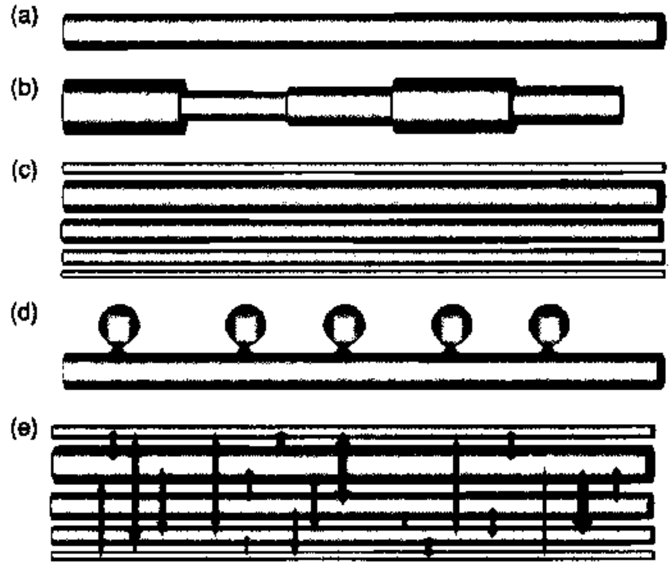


**Figure 4** (a) A single tube with constant cross section; (b) a tube with varying cross section; (c) a bundle of tubes of different but constant cross sections; (d) a single tube with dead-end 'pores'; and (e) a bundle of tubes of different cross sections with solute exchange between tubes.

**Table 1** Scaling with time ($T$) of solute mass mean and standard deviation according to four transport models

| *Model* | *Mean* | *Standard deviation* | *Standard deviation/mean* |
|---|---|---|---|
| ADE | $T$ | $T^{1/2}$ | $T^{-1/2}$ |
| MIM | $T$ | $T^{1/2}$ | $T^{-1/2}$ |
| CLT | $T$ | $T$ | Constant |
| CTRW ($0 < \beta < 1$) | $T^{\beta}$ | $T^{\beta}$ | Constant |
| CTRW ($1 < \beta < 2$) | $T$ | $T^{(3-\beta)/2}$ | $T^{(1-\beta)/2}$ |

ADE, advection–dispersion equation; MIM, mobile–immobile model; CLT, convective lognormal transfer function model; CTRW, continuous-time random walk model.

Adapted from: Jury WA, Gardner WR, and Gardner WH (1991) *Soil Physics*, 5th edn. New York: John Wiley; Skaggs TH and Leij FJ (2002) Solute transport: theoretical background. In: Dane JH and Topp GC (eds) *Methods of Soil Analysis*, part 4, chap. 6.3, *Physical Methods*. Madison, WI: Soil Science Society of America; Berkowitz B, Kosakowski G, Margolin G, and Scher H (2001) Application of continuous time random walk theory to tracer test measurements in fractured and heterogeneous porous media. *Ground Water* 39(4): 593–604.

function to generate a near-Gaussian distribution with 'fat tails.' This approach, being more general, usually results in a better fit than the original ADE. However, it is somewhat more complex mathematically and only accounts for a subset of cases that are not well fit by the ADE. Interestingly, the ADE is unable to account correctly for dispersion resulting from flow through layered soils, conceptually equivalent to different tubes in series, as shown in **Figure 4b**: it predicts identical results regardless of which soil comes first, while experimentally the order of the layers affects the final value of the dispersion coefficient.

A second common conceptual model is the so-called streamtube model or transfer function model, in which solute transport is considered to take place through multiple parallel but noninteracting streamlines (**Figure 4c**). Because mean solute velocity within a tube is proportional to tube radius squared, dispersion results from the differences in velocity among the various streamtubes. Because in principle it is possible to reproduce any possible breakthrough curve by generating the correct streamtube distribution, this method is very flexible. This model is essentially a multiple-tube equivalent of the dispersion that would result from flow in a tube in the absence of diffusion, and so, because their dispersion results from insufficient mixing, streamtube models may be considered to be preasymptotic. Streamtube models are most applicable to media in which flow really does take place through parallel but noninteracting pathways, such as horizontal flow through an aquifer with homogeneous layers of contrasting properties. Some variants of this model allow dispersion within each streamtube, while others assume no local dispersion (i.e., piston flow). One common version, Jury's convective lognormal transfer function model (CLT), specifies the distribution of the solute travel times (or, by analogy, the tube radii) and thus the form of the final BTC.

Knowing the dispersion at one scale only allows correct prediction at a different scale if the chosen model scales the same way as the observations. For example, while the ADE predicts that solute-spreading scales with the square root of time, the CLT (a specific streamtube model) predicts solute-spreading that scales linearly with time (**Table 1**). Solute movement that is consistent with a streamtube-like mechanism will therefore appear to scale if analyzed using the ADE; likewise, solute transport that is well described by the ADE will scale incorrectly if analyzed using the CLT.

The streamtube models assume zero interaction between streamlines, in contrast to the ADE, which assumes that streamlines do interact. These two models therefore represent extremes along a continuum of streamline interaction. In principle, the streamtube model should better represent early or preasymptotic times, before streamlines have had time to intermix adequately, and the ADE later times. Because in the streamtube model dispersion scales linearly with time (Table 1), invoking a streamtube model at short (preasymptotic) times and the ADE at later times provides a plausible explanation for some scaling behavior: one can argue that the observations took place within the transition between the two models. By this reasoning, the time or distance needed to mix – to make the transition from streamtube-like dispersion to ADE-like dispersion – should provide an estimate of the soil's REV with respect to dispersion, and so some indication of the characteristic length of the soil's structure.

A third model in common use is the mobile–immobile (MIM) model, also known as the two-region model. Originally developed to explain dispersion in terms of dead-end pores, this model has been widely adopted to describe dispersion in structured soils. The idea is that the soil is composed of mobile regions (such as interaggregate pores) through which most or all of the flow actually moves, closely interconnected with immobile regions (e.g., the interior of aggregates) that exchange mass with mobile regions by diffusion only (**Figure 4d**). The parameters of the model are therefore the dispersion in the mobile region (identical to the dispersion term in the ADE), the mass exchange coefficient, describing diffusive exchange between the mobile and immobile regions, and the mobile pore fraction. The model can equivalently be thought of as describing chemical sorption to soil particles that comes in two varieties, instantaneous and slow, but, in keeping with the focus here on nonreactive solutes, this view will be ignored. Because diffusive exchange between regions is generally slow relative to advection, a given layer of soil is generally out of equilibrium with the solute flowing through it for some time, so this model is often said to describe nonequilibrium flow. When the immobile region has zero volume, or when the diffusive exchange is instantaneous, the MIM reduces to the ADE; this extreme can be though of as describing a soil of low heterogeneity. As might be expected from its close ties with the ADE, the MIM predicts identical scaling behaviors (**Table 1**). On a practical level, the MIM often suffers from nonunique solutions; that is, a given BTC may be equally well fit by more than one single set of parameters. This nonuniqueness implies that the equation has too many fitting parameters (three for the MIM versus one for the basic ADE), and casts some doubt on the physical interpretation of those parameters.

## Observations and Mechanisms

There is a strong tendency for dispersivity to be greater at larger scales, as documented (for example) in several metastudies, each examining many different field studies of solute transport. Unfortunately, the scatter in the data is great enough that no universal scaling behavior can unambiguously be inferred. While the cause of this trend is not clear, the most common explanation is that, as a solute travels farther, it encounters increasing scales of heterogeneity. This means that at larger scales solute molecules will on average sample a larger range of velocities than they did at a smaller scale. The travel time for each molecule is the sum of many different times at many different velocities; increasing the range of velocities will therefore (assuming random and independent sampling by each molecule) result in a wider distribution of travel times and thereby a larger dispersivity. A fractal medium, with its ever-increasing scale of heterogeneity, results mathematically in dispersivity increasing with distance, i.e., scaling.

But increasing scales of heterogeneity are apparently not necessary for scaling to occur. For example, one laboratory experiment used a two-dimensional flow cell packed with two sands: 16% of the volume consisted of blocks of fine sand randomly placed within a coarse sand matrix, with the hydraulic conductivities differing by a factor of 20. Within approximately 1 m, the fit dispersivity increased from 0.073 to 0.127 m. The skew of the data was fairly constant, suggesting that a change from preasymptotic to asymptotic behavior was not occurring. Similarly, the Borden aquifer tests saw dispersivity increase by a factor of 8 over a distance of some 10 m in what was considered a homogeneous sand. And of course dispersion that occurs primarily through a streamtube-like mechanism, wherein regions that differ in velocity are isolated from each other, will also show scaling of ADE-calculated dispersivity.

Discussion of the 'scale of heterogeneity' also raises the question of the importance of heterogeneity at smaller scales. If much of dispersion is driven by heterogeneity, and that heterogeneity exists at all scales, then presumably the nature of the heterogeneity below the scale of heterogeneity is also important. But because both models and measurements are blind to processes below their respective nonzero-size scales, there are always smaller-scale processes that they cannot take into account, but which still influence the large-scale outcome. An exclusive focus on the largest-scale heterogeneities implicitly assumes that the smaller-scale heterogeneities are either negligible or diffusive in nature, but neither is necessarily the case. The smaller-scale influences can be significant and nondiffusive; for example, the presence of undetected cracks in a soil profile may lead to arrival of solute at the water table well in advance of prediction.

Soils and other geologic porous media are neither purely random nor purely structured; instead they vary in their degree of randomness across most or all scales. But while understanding of soil properties has historically concentrated on the measurements themselves, and only relatively recently on the spatial structure of the properties, generally the topological structure is still ignored. As a consequence, researchers in the theory of contaminant transport have long considered (for example) what proportion of an aquifer is sand or clay, and even how thick the average sand layer might be, but not the question of whether there is a continuous sand pathway through the system. Continuity and interconnectedness are critical to diffusive and advective exchange across streamlines, and to the range of velocities encountered within a given streamline. They therefore make the difference between the ADE and streamtube models, and so are critical determinants of the scaling behavior of a given soil or aquifer.

In most situations, solute transport in soils or other geologic porous media is subject to all of the mechanisms mentioned so far: diffusive exchange between neighboring streamlines, parallel pathways with little exchange between them, and diffusive exchange between more and less mobile regions of the medium. There are also other important dispersive processes, such as advective (and even turbulent) exchange between streamlines, streamline-specific sorption based upon mineralogy, biodegradation based upon differences in microhabitat, and streamlines with nonconstant velocity, that do not show up in these models. Certainly the extent to which each mechanism contributes to dispersion is difficult to determine in advance, nor can one know in advance whether the length scale at which dispersion is to be predicted is greater than or less than the scale of heterogeneity (if any) of the medium.

## Model Selection

An investigator attempting to predict future dispersion from current conditions is faced with the choice of three conceptual models, each based upon quite different mechanisms and assumptions. Prediction therefore involves two tasks: selecting the appropriate model, and fitting it to the data in order to make a prediction. The wrong model choice may result in scaling, such that the predicted behavior may be severely rather than merely slightly wrong. But in fact the choice of models is still quite narrow, in the sense that only two scaling behaviors are available (Table 1), even though intermediate behaviors are commonly observed.

An alternative is to use a new model, originally developed to describe electrical conductivity of amorphous silicon. The continuous-time random walk model (CTRW) has been used to describe solute transport only since the mid-1990s, but it has been shown to be capable of capturing the entire range of behavior seen in the three models discussed above. Each of the three models is a special case of the CTRW, indicating that the CTRW is a generalized transport model that can handle both the extremes and the intermediate cases.

The CTRW conceives of the soil as having a variety of pathways of higher and lower conductivity. What determines the dispersion is not only the range of velocities represented by these pathways, but also the probability with which a solute molecule will transition from one pathway to another (**Figure 4e**); such transitions can take place both through diffusion and through advective pathways merging and separating. Transition probabilities are given by a joint probability distribution function $\psi(s,t)$, which describes the probability of a particle moving a distance and direction $s$ within time $t$. The long-time behavior is dominated by transitions at long (asymptotic) times, and asymptotic forms of probability functions tend to decrease monotonically. (In the context of the CTRW, asymptotic behavior is always reached quickly.) The distribution can then be characterized by exponential decay, which leads to ADE-like behavior, and power-law decay in which $\psi(s,t) \sim t^{-1-\beta}$ for $0 < \beta < 2$. Power-law decay yields anomalous transport except for $\beta > 2$. The value of the exponent $\beta$ controls the rate of decay, and thereby the scaling of the solute mean and standard deviation (**Table 1**).

A wider range of scaling behaviors is attained with the CTRW than with the original three models combined. Significantly, because the choice of $\beta$ changes both the shape of the BTC and its temporal scaling, the CTRW implicitly considers temporal scaling to be an inherent aspect of dispersion, rather than a separate phenomenon. This raises the possibility that in the future the concept of scaling as a distinct issue will cease to exist, being replaced by an understanding that dispersion has inherent scale-related behaviors.

*See also:* **Fractal Analysis; Porosity and Pore-Size Distribution; Scaling:** Physical Properties and Processes; **Spatial Variation, Soil Properties**

## Further Reading

Baveye P and Boast CW (1999) Physical scales and spatial predictability of transport processes in the environment. In: Corwin DL, Loague K, and Ellsworth TR (eds) *Assessment of Non-Point Source Pollution in the Vadose Zone*, pp. 261–280. Geophysical monograph 108. Washington, DC: American Geophysical Union.

Berkowitz B, Kosakowski G, Margolin G, and Scher H (2001) Application of continuous time random walk theory to tracer test measurements in fractured and heterogeneous porous media. *Ground Water* 39(4): 593–604.

Crank J (1975) *The Mathematics of Diffusion*, 2nd edn. Oxford, UK: Clarendon Press.

Gelhar LW, Welty C, and Rehfeldt KR (1992) A critical review of data on field-scale dispersion in aquifers. *Water Resources Research* 28(7): 1955–1974.

Jury WA, Gardner WR, and Gardner WH (1991) *Soil Physics*, 5th edn. New York: John Wiley.

Matheron G and de Marsily G (1980) Is transport in porous media always diffusive? A counter example. *Water Resources Research* 16(5): 901–917.

Molz FJ, Güven O, and Melville JG (1983) An examination of scale-dependent dispersion coefficients. *Ground Water* 21(6): 715–725.

Skaggs TH and Leij FJ (2002) Solute transport: theoretical background. In: Dane JH and Topp GC (eds) *Methods of Soil Analysis*, part 4, *Physical Methods*, chap. 6.3. Madison, WI: Soil Science Society of America.

Sposito G (ed.) (1998) *Scale Dependence and Scale Invariance in Hydrology*. Cambridge, UK: Cambridge University Press.

Taylor GI (1953) The dispersion of matter in a solvent flowing slowly through a tube. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences* 219: 189–203.

# SEPTIC SYSTEMS

**R L Lavigne**, University of Massachusetts, Leeds, MA, USA

## Introduction

Since the advent of societal living, humankind has relied on the soil environment to treat a wide variety of its wastewater residues. The use of soil-based treatment systems continues to be a common practice in essentially all countries of the modern world. For centuries people have relied on the capacity of soil to purify waters contaminated from solid-waste disposal facilities, plague-infested communities, municipal sewage collection systems, industry, agriculture, cemeteries, and more commonly small residential, institutional, and commercial wastewater-producing facilities. Given adequate retention or treatment time, and sufficient spacing from surface and groundwater resources, many soils can effectively degrade organics through microbial processes; adsorb, precipitate, or complex most metals; and subsequently release an effluent that is near potable quality. In many parts of the world, soil-based wastewater-treatment systems function in close proximity to public or private drinking-water supplies, with little or no adverse effect on potability. In densely populated areas, however, where appropriate separation and treatment distance are not provided for soil-based treatment systems, potable water contamination from bacteria, viruses, nutrients, metals, and other contaminants has occurred.

Unfortunately, the generally small size of on-site wastewater-treatment systems ('septic systems') has created a global policy where most of them are built without regulations, technical guidelines, or protocols for maintaining and operating them. In highly regulated areas, more stringent controls may exist, but even where septic-system regulations do occur, their requirements are seldom founded on the principles of sound soil science or soil physics. In the USA, municipalities often use septic-system regulations to control development, and all too often the regulations become counterproductive in protecting the environment and public health. A classic example of this paradox is the encouragement of septic systems in highly permeable soils where groundwater and surface-water resources can quickly become polluted.

## Technical and Historical Overview

The most common septic systems in use worldwide include outhouses and cesspools. In spite of their simple nature, they are often effective in treating human wastewater and protecting groundwater and surface-water resources. During the 1960s, the Commonwealth of Massachusetts embarked on a region-wide program to install cesspools in populated areas surrounding its Boston water supply known as Quabbin Reservoir. The program was largely successful, and the reservoir continues to provide surface water supplies in the Northeast part of the USA.

### Outhouse – Structure and Function

In structure, outhouses are little more than a hole in the ground, usually 1–2 m deep. A wooden structure or 'house' is erected above the hole, with one or more seats. Liquid wastes percolate into surrounding soils, while insect larvae and bacteria degrade the solid component. In time, the hole fills from regular use. The outhouse is then relocated above a new excavation and the process repeats itself. In warm weather, hydrated lime can be added to the waste to reduce odors.

### Cesspool – Structure and Function

The advent of indoor plumbing has added relatively large volumes of water to daily domestic wastewater production rates. For much of the world today, showers, flush toilets, laundry, and dishwashing needs result in a per capita production rate of approximately 200 l $day^{-1}$ of wastewater. To accommodate these increased volumes, outhouse holes in the ground have been increased in depth and diameter. Many of these excavations are lined with stone or cribbed with timbers. Properly constructed, the common cesspool can provide decades of service with little or no maintenance needs other than periodic removal of the solids. A secure cover over the top of concrete or wood is required for safety reasons.

## The Modern Septic System

For more technically oriented households, the use of strong soaps, bleaches, and cleaning products quickly reduces the ability of fecal bacteria to degrade the solids associated with the wastestream. Cesspools accumulate solids more quickly, minimizing the solid–water interface's ability to effectively conduct water out of the unit. For this reason a separate solids containment unit has evolved that is commonly called a septic tank.

### Septic Tank

With mean hydraulic retention times of a day or more, approximately 65% of the solid matter can be removed by simple settling in a septic tank. A baffle or 'T' at the inlet end helps reduce short-circuiting of incoming surges, and a similar structure at the outlet end prevents floating grease and oils from leaving the unit. Although some biological degradation does occur anaerobically, the modern septic tank usually requires cleaning every 2–5 years. Concrete, metal, fiberglass, and other synthetics are commonly used to build septic tanks, but wood structures are also fairly common. The tanks are normally watertight and buried 30–60 cm belowground to facilitate the gravity flow of wastewater from the dwelling. A gas vent from the building sewer is typically hidden in the outside wall of the dwelling, where it normally exits at the roof eve. Anaerobic gases from the septic tank migrate upwards through the building sewer and out the roof vent. During periods of thermal inversion, vent-stack gasses can sink to ground level, giving false signs of septic system failure by way of odor.

### Soil Absorption Systems

Once the liquid component of wastewater exits in a septic tank, it must be disposed of in some acceptable manner that is safe for the environment and public health. Soil absorption systems (SASs) satisfy this need in a very effective manner. A SAS can be built in a variety of shapes and configurations, but all SASs are generally designed to do the following:

1. Provide sufficient surface area to permit 100% water exfiltration for many years following the development of a typical soil interface 'scum layer';
2. Provide sufficient separation distance to insure that the wastewater does not enter the groundwater before treatment;
3. Provide sufficient retention time to insure adequate treatment of the liquid component in the SAS, the scum layer (usually 10 cm); and in the underlying unsaturated soils.

Typically a SAS is built in a shallow excavation 50–100 cm deep made in naturally occurring soils. Perforated 10-cm pipes are used to distribute liquid throughout the excavation. The pipes (usually plastic, concrete, or ceramic) are surrounded by clean coarse stone to further enhance the distribution process. A layer of fine stone, synthetic fabric, asphalt, paper, or hay covers the stone bed to prevent infiltration of natural soils that are replaced above. Grass is usually planted over a SAS so that the unit blends in naturally with the surrounding landscape.

From time to time an SAS must be elevated aboveground infill to provide adequate separation distance either to bedrock or the groundwater table. These aboveground units often require that a pump be used to lift wastewater from a 'pump tank' that follows the septic tank. Typical SAS configurations are often referred to as 'beds,' 'trenches,' 'galleries.' Gallery units are often prefabricated using plastic, concrete, or other structurally sound materials. They can also be built using concrete blocks, wood cribbing, or stone. In each case, the SAS is built on a level base to promote the uniform distribution of wastewater over the entire base area. When deeper SAS units can be constructed (i.e., greater than 50 cm), the sidewall area may be utilized to provide part of the required absorption surface area.

### SAS Size and Soil Infiltration Rates

Although heuristic methods are often used to size SAS units, it is more common to employ some type of soil-testing to evaluate infiltration rates. These methodologies include: percolation-testing, *in situ* infiltrometer methodologies to determine coefficients of permeability (meters per day), or laboratory hydraulic-conductivity measurements using permeameters. Where testing is required, regulations typically prescribe a loading rate (volume per area) for SAS units, as well as design-flow rates based on the number of dwelling bedrooms, bathrooms, or inhabitants. Flow-rate estimates are also available for categories of use (e.g., schools, cafeterias, hotels, restaurants).

Percolation testing is an *in situ* procedure requiring that a hole of specific diameter and depth (e.g., 30 cm × 45 cm) be hand dug into the soil strata being considered for a SAS. Once filled with clean water, the rate of water percolating is measured and reported as centimeters per minute or some arbitrary unit of depth versus time. Usually the slower the percolation rate, the greater the area required for loading, thereby increasing the total surface area needed for a SAS unit being designed for a given design-flow rate. In fine soils (silts, fine sands, and clays), a three-bedroom residential structure may require a SAS area several hundred square meters. In coarse-textured soils (sands and loamy sands), the required area for the same dwelling may be as little as 50 $m^2$. Regardless of the methodologies used, the objectives are to (1) contain 100% of the liquid fraction even during periods of peak loading; (2) provide adequate treatment time to reach acceptable infiltration concentration; and (3) provide appropriate separation distances to protect groundwater and surface-water resources.

## Advanced Treatment Systems

Because septic systems have proliferated in some areas in a relatively uncontrolled manner, groundwater and surface water problems have evolved. Usually these problems take the form of nitrate ($NO_3^{2}$) contamination in groundwater, or nutrient addition to surface waters that accelerate eutrophication processes. In some instances bacterial and viral contamination has also occurred in wells and surface-water resources. When these situations occur there are usually two paths of action considered. One is to sewer the area to provide municipal treatment. The other is to upgrade SAS units to provide appropriate levels of advanced treatment (e.g., disinfection, nitrate removal, nutrient removal). In the latter half of the twentieth century, many innovative alternative technologies have been developed. Unfortunately, most of these units require some degree of operation and maintenance that individual homeowners may not be able to provide. There has also been increased interest in decentralized sewerage systems where small community systems can be constructed.

### Artificial Wetlands

A new variety of SAS incorporates the treatment processes known to be occurring in natural wetlands. These artificial or manmade wetlands can be constructed with or without liners, depending on soil texture and groundwater-protection needs. Top-loading, vertical flow, submerged bed systems are constructed to provide a park-like environment, while simultaneously producing a near drinking-water effluent. Generally these operate with no chemicals, no moving parts, and little or no energy needs other than the sun and gravity. Wetland plants growing in a coarse-textured medium provide oxygen to soil bacteria that in turn function as a biofilm of microbes metabolizing both organic and inorganic wastewater components. A dense stand of grasses and reeds is generally grown in these systems.

## Summary

In the USA alone, it has been estimated that approximately 25% of all wastewater produced is treated with on-site septic systems. As rural lands continue to develop, the demand for individual treatment systems will surely rise. Considerable data emerging to show that conventional wastewater-treatment technology is not very environmentally friendly, nor is it always safe for public health.

Large quantities of fossil fuels are consumed each year, providing the energy needs that conventional technology demands. Disinfection processes using chlorine introduce carcinogens into surface waters. Conventional systems seldom are designed to remove nutrients, so that nitrogen, phosphorus, and potassium compounds are being directly discharged into lakes, rivers, and oceans. As these aging wastewater-treatment facilities become more costly to maintain, alternative technologies are likely to play a leading role in the industry. This is especially true for industrially emerging countries in Southeast Asia, Central and South America, and Africa. Properly designed, built, and operated, the conventional septic systems can help to meet the environmental and public health needs in many regions.

*See also:* **Microbial Processes:** Environmental Factors; Kinetics; **Pollutants:** Persistent Organic (POPs); **Waste Disposal on Land:** Liquid

## Further Reading

Salvato JA (1992) *Environmental Engineering and Sanitation*, 4th edn, pp. 489–578. New York: John Wiley & Sons, Inc.

USEPA (2002) *Onsite Wastewater Treatment Systems Manual*, EPA/625/R-00/008. Cincinnati, OH: National Risk Management Research Laboratory.

# SHIFTING CULTIVATION

**R Lal**, Ohio State University,
Columbus, OH, USA

## Introduction

Soil, the three-dimensional upper layer of the Earth's crust that directly or indirectly supports all terrestrial life, is a nonrenewable resource over the human time scale. Further, soil resources of the world are finite, unequally distributed among geographical regions, and prone to severe degradation by misuse and mismanagement. Consequently, soil resources are shrinking with continuous increase in human population, due to degradation and conversion to nonagricultural uses, including urban encroachment. Therefore, judicious management of soil is critical to global food security, the health and welfare of humans and animals dependent upon it, the overall economy, and regional social and political stability.

Soil management implies strategic planning of all inputs into and outputs from the soil ecosystem so that there is a favorable balance of essential components that constitute the basis of soil's life-support system. These components comprise the soil organic matter (SOM) pool, reserves of essential plant nutrients (e.g., N, P, K, Ca, Mg, Zn, Cu, Mo, etc.), water balance and plant-available water capacity (AWC), and effective rooting depth. It is the imbalance caused by drastic differences between outputs and inputs of one or more of these components that sets in motion soil degradative processes. For example, depletion of SOM or biological degradation is caused when removal of crop residue and agronomic output exceeds the input of compost and other biosolids. Soil fertility declines and nutrient depletion occurs when nutrients harvested in plant and animal produce exceed nutrients supplied as chemical fertilizers or organic amendments. Depletion of bases (e.g., Ca, Mg, K) from soil can lead to acidification and elemental imbalance, resulting in excess of Al, Mn, and Fe. Waterlogging and soil salinization are caused by excessive use of irrigation water without provisions for drainage. The problem is aggravated if irrigation water is of poor quality and the subsoil contains a high concentration of water-soluble salts, as is generally the case in soils of arid and semiarid regions. Exposure of a structurally fragile soil to climatic elements (e.g., intense rains, flowing water, and blowing wind) can lead to removal of topsoil, its transport into streams and reservoirs, truncation of the rooting depth, and decline in the soil's productivity and environment moderation capacity.

It is this imbalance in outputs and inputs, due to soil misuse and mismanagement, that causes soil degradation (**Figure 1**). The latter is defined as a decline in the soil's capacity to perform functions of interest to humans. These functions include biomass productivity; filtering pollutants out of water; biodegradation of contaminants and pathogens; and cycling/storage

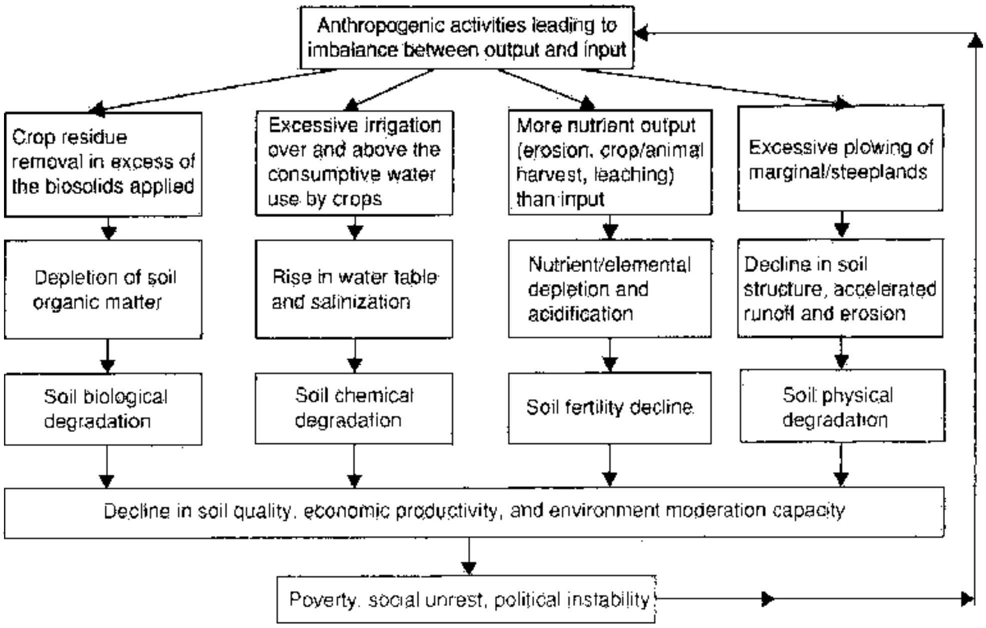


**Figure 1** Soil degradation caused by imbalance between outputs and inputs.

of carbon (C), nitrogen (N), and other elements. Degradation severely jeopardizes the soil's capacity to perform these functions.

Soil degradation has been a challenge to humanity since the dawn of agriculture some 10 000 years ago. Severe soil degradation caused some of the once-thriving ancient civilizations to vanish. Indeed, it was the severe erosion caused by deforestation of cedar forests in the Mediterranean Basin that toppled the Phoenicians and destroyed granaries of the Roman Empire. The fertile land of Mesopotamia, present-day Iraq, was converted to desert and shifting sand dunes by deforestation, severe erosion of the hills and watersheds that nurtured the ancient irrigation system, and salinization of agricultural lands. The extinction of the Harappan-Kalibangan civilization in the pre-Aryan Indus Valley has been attributed to erosion-caused desert encroachment and salinization. The collapse of the 1700-year-old Mayan culture in Central America about AD 900 was caused by severe erosion of a thin topsoil by excessive cultivation necessary to support a rapid increase in population. In most cases, soil degradation was caused by the 'tragedy of the commons,' which could have been avoided by rational management and conservation of soil and natural resources.

Therefore, 'conservation' of soil resources is vital to the survival of the human race. The term 'soil conservation' means 'a judicious use of soil so that its productivity and environment moderation capacity are perpetually maintained.' Soil conservation also means 'rehabilitation of degraded soils for restoring their productivity and environment moderation capacity.' In view of the high demographic pressure and conflicting demands on shrinking soil resources, soil management and conservation are more important now than ever before. In fact, the importance of soil management and conservation can never be overemphasized.

## Shifting Cultivation

Soil management and conservation have a special significance to regions where shifting cultivation is practiced. The term 'shifting cultivation' refers to farming or agricultural systems in which a short but variable cultivation phase on 'slash-and-burn' cleared land alternates with a long, equally variable fallow period. Shifting cultivation is the most ancient system of agriculture in which soil fertility is restored by long periods of fallowing rather than by off-farm inputs of fertilizers and amendments, nutrients are recycled between natural vegetation and crops, and ecological balance is maintained by adopting diverse and complex cropping systems rather than monoculture. Shifting cultivation is still being practiced in some parts of the tropics and subtropics, especially in the humid tropics of west and central Africa, Southeast Asia and South/Central America. It is an extensive system of agriculture, with a wide range of variation depending on soil type, vegetation, terrain, climate, and social/ethnic factors. The system is stable where the ratios of the fallow period to cultivation phases are up to 10 or more. The system breaks down, causing severe soil degradation and decline in productivity, when the fallow period is drastically reduced. An important reason for the reduction in fallow period is an increase in the demographic pressure, and an attendant decrease in available soil resources.

Shifting cultivation systems have the following characteristics:

1. They are highly diverse and complex, involving simultaneous growing of 10 or more crops in combination with trees and livestock. Mixed cropping is the general rule.
2. Shifting cultivation systems are labor-intensive with minimal use of purchased off-farm inputs. Restoration/management of soil fertility is based on nutrient cycling through lengthy fallow, biological nitrogen fixation (BNF), and choice of compatible species.
3. Shifting cultivators are generally small-size landholders with a normal size of <2 ha. Therefore, farm operations are performed by the farm family.

There are several attempts to classify shifting cultivation systems. A farming system is designated nomadic herding if the value of the land-use factor, $L \geq 10$; bush fallow or land rotation for $L$ between 5 and 10; rudimentary or sedentary agriculture for $L$ between 2 and 4; compound farming terrace farming or paddy rice cultivation for $L < 2$. Most classification systems are based on 'intensity' or 'frequency' of cultivation. The various categories designate different degrees of intensification as evaluated on the basis:

$$L = (C + F)/F \qquad [1]$$

where C is the number of years of cultivation and $F$ is the number of years fallow. Traditional systems are also classified on the basis of the degree of intensification:

Phase I: Hunting and gathering
Phase II: Subsistence agriculture
Phase III: Early consumer stage
Phase IV: Primary mechanization stage

The true shifting cultivation or the hunting and gathering stage ($L = \geq 10$) is no longer widespread in any region. Yet, subsistence agriculture including

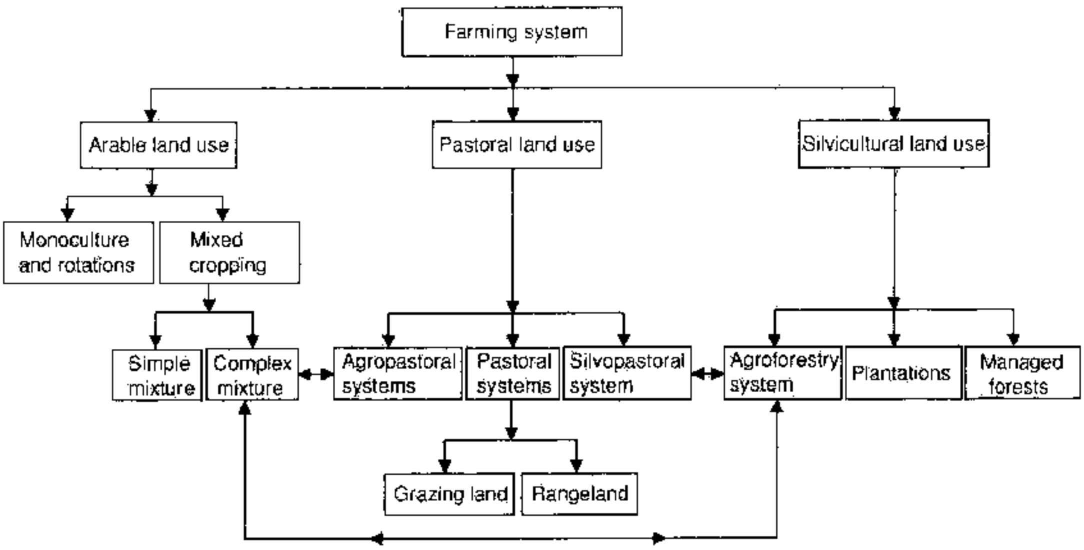


**Figure 2** Permanent farming system.

bush fallow rotation ($L = 5$–$10$) is practiced in some regions of Africa, Asia and South/Central America. Subsistence farmers grow a variety of crops (mixed cropping) in association with selected tree species and animals, and as much as 90% of the produce is consumed on the farm. A majority of the shifting cultivators in the tropics practice rudimentary sedentary agriculture at the early consumer stage ($L = 2$–$4$). These farmers market less than 50% of their produce, and practice mixed farming. Farmers may also use some machinery and low rates of fertilizers. A predominant type of mechanization use by phase-III farmers involves tractor-based tillage.

In contrast to shifting cultivation, farming systems involving permanent agriculture are outlined in Figure 2. These systems are based on three principal land uses and appropriate two-way or three-way combinations. For example, agropastoral systems are based on a combination of crops and livestock, and agrosilvopastoral systems are based on a combination of crops, livestock, and trees.

## Productivity of Complex Farming Systems

Because a large number of crops are grown simultaneously on the same piece of land, and often harvested at different times, it is a challenging task to assess the productivity of such complex systems. Some of the indicators commonly used to assess productivity include the following:

1. Productivity: Productivity can be assessed as biomass (agronomic) yield or income. It can be expressed either as a simple productivity or total factor productivity (TFP):

$$P = p/R \qquad [2]$$

$$\mathrm{TFP} = p/\sum_{i=1}^{n}(R_i \times C_i) \qquad [3]$$

where $P$ is productivity, $p$ is total production, $R$ is the resource consumed/used, and $C$ is the cost of the resource. Being truly resource-based, it may be difficult to assess the true cost of all resources used for phases I and II of shifting cultivation systems. Therefore, the following indices are used to assess productivity.

2. Land equivalent ratio: Productivity of mixed cropping systems is assessed by computing the land equivalent ratio (LER):

$$\mathrm{LER} = \sum_{i=1}^{n} y_i/(y_m) \qquad [4]$$

where $y_i$ and $y_m$ are yields of component crops in the intercrop and monoculture system, respectively, and $n$ is the number of crops involved.

3. Area × time equivalency ratio (ATER): In shifting cultivation systems (phases II and III), there is a wide range and a large number of crop/tree species involved. These species vary widely in their maturity period. The ATER index assesses productivity while considering the crop duration:

$$\mathrm{ATER} = 1/t\left(\sum_{i=1}^{n}(d.y_i)/y_m\right) \qquad [5]$$

where $d$ is the growth period of crop in days and $t$ is the total time in days for which the field remained occupied, i.e., the growth period of the longest-duration crop. The numerical value of ATER index approaches that of LER for a mixture consisting of crops of approximately identical growth periods, i.e., when $t = d$ in eqn [5].

4. Resource-use efficiency: When some off-farm inputs are used (phase IV), the productivity can also be assessed in terms of the resource-use efficiency. This is a relevant option in case of a system with the most limited resource such as P, N, water, soil organic carbon (SOC), or energy use.

5. Carbon dynamics: In view of the rapid increase in atmospheric concentration of $CO_2$ and other greenhouse gases, assessing C dynamics is an important indicator of the productivity.

$$P = \left(C_o/(\sum_{i=1}^{n} C_i)\right)_t \qquad [6]$$

where $P$ is productivity, $C_o$ is carbon produced in above- and belowground biomass, $C_i$ is carbon input in all components including fertilizers and pesticides etc., and $t$ is time. The higher the ratio, the more sustainable is the system.

## Ecosystem-Related Constraints to Agricultural Intensification

Shifting cultivation systems are ecologically viable as long as there is enough land for long (10–20 years) restorative fallow, and expectations of crop yield and the attendant standards of living are not too high. These systems are naturally suited for harsh environments and fragile ecosystems of the tropics. That is why attempts at finding viable alternatives to shifting cultivation have met with only limited success. There are several soil-related and climate-related constraints (**Table 1**) that must be alleviated. These constraints lead to severe problems of soil degradation upon conversion from shifting cultivation to mechanized intensive agriculture. Consequently, many large-scale mechanized farming schemes introduced in the humid tropics have been a failure.

**Table 1** Ecological constraints to finding viable alternatives to shifting cultivation

| *Soil-related constraints* | *Climate-related constraints* |
|---|---|
| 1. Low inherent soil fertility | 1. Low level of radiation during the growing season |
| 2. High susceptibility to soil erosion | 2. High relative humidity |
| 3. Severe soil physical degradation | 3. High storage losses |
| 4. Nutrient/elemental imbalance | 4. Problems with grain drying |
| | 5. High incidence of pests and disease |
| | 6. High-intensity rains |

The severe problem of soil degradation is caused by indiscriminate and intensive land use based on monoculture regardless of soil capability, introduction of pastures with high stocking rate and uncontrolled grazing, or intensive cropping without input of chemical fertilizers or compost at the required rates. The problem of soil degradation is exacerbated by harsh climate characterized by intense rains of high erosivity and structurally weak soils.

## Sustainable Management of Soils Under Shifting Cultivation

Any viable alternative to shifting cultivation system must successfully:

1. maintain soil structure at a favorable level so that risks of soil physical degradation (e.g., crusting, compaction, low infiltration) are minimized;
2. control excessive runoff and accelerated soil erosion so that effective rooting depth is maintained and the nutrient-rich topsoil is not eroded away;
3. strengthen nutrient-recycling mechanisms so that nutrients are not lost out of the ecosystem, and are effectively recycled through incorporation of woody perennials and legumes in the rotation cycle, and judicious use of chemical fertilizers and organic amendments;
4. manage pests through enhancing biodiversity and using a healthy balance between predators and parasites through strategies of integrated pest management; and
5. increase farm income through value-added products and off-farm employment.

While traditional shifting cultivation systems meet these requirements through low cropping intensity and lengthy fallows, new systems have to be identified that produce high yields per unit area of land and per unit time. Sustainable soil management technologies must address specific ecological constraints such as those outlined in **Table 1**. Any viable alternative must address the synchrony between food security and environment quality without jeopardizing the soil/natural resources, which are already under great stress. Therefore, sustainable soil management systems must:

1. assure food security for a high and increasing population;
2. enhance soil productivity through ecosystem restoration and judicious use of off-farm input;

3. improve environment quality, especially that of water and air, and reduce emissions of greenhouse gases into the atmosphere; and
4. transform subsistence farming into commercially viable agriculture.

## Characteristics of an Improved/ Sustainable Soil Management System

Important characteristics of an improved/sustainable soil management system are outlined in Figure 3. The new system must have high energy flux in terms of drastic increases in crop yield brought about by discriminant or judicious use of off-farm input. The objective is to increase nutrient input into the system through integrated nutrient management (INM). The latter involves elemental recycling (e.g., use of crop residue, BNF, use of compost and city sludge, etc.), supplemented by a judicious use of chemical fertilizers. The objective is to increase energy efficiency and flux, because a low-output system can be highly efficient but unproductive, just as a high-output system can be extremely inefficient. Therefore, the goal is to develop a high-output and highly efficient production system. The efficiency can be increased by decreasing the losses. To be sustainable, the system must of necessity be soil-restorative. Most soils under shifting cultivation are being depleted and degraded because of nutrient mining over a long period of time. Some components of an improved system comprise the following:

1. New land development: The emphasis should be toward intensification of existing land so that there is no need for additional deforestation and conversion of a natural to an agricultural ecosystem. If deforestation is absolutely necessary, it must be done in a way to cause as little perturbation in the ecosystem as possible, especially with regards to the following:

a. water and energy budgets, which are disturbed by deforestation, causing an increase in surface soil temperatures and acceleration of surface runoff with the attendant problem of severe erosion. Thus the objective is to cause a minimal disturbance to the hydrological cycle, especially the surface runoff component, which must be kept under control.
b. maintenance of soil structure, which is crucial to ensuring a high water infiltration rate, minimizing risks of crusting and compaction, and decreasing risks of accelerated soil erosion. Application of crop-residue mulch and establishment of cover crops are necessary to enhance activity of soil biota and improve soil structure.
c. increasing plant nutrient reserves, which is important to obtaining high crop yields. This can be achieved by replacing nutrients harvested in crops and those lost by erosion and leaching through use of INM strategies.

Therefore, removal of existing vegetation can be done either manually or by those machines that cause least disturbance to the surface soil (e.g., those with shear blade). Land-use capability assessment and proper planning are essential to any successful development of new land.

2. Intensive use of existing lands: Because of the shortage of potentially suitable new land and the undesirability of converting forest/savanna ecosystems into croplands, an increase in production in several countries must be achieved by more intensive cultivation of land already developed. It is equally important to restore productivity of degraded soils. Therefore intensification of existing arable land is a high priority.

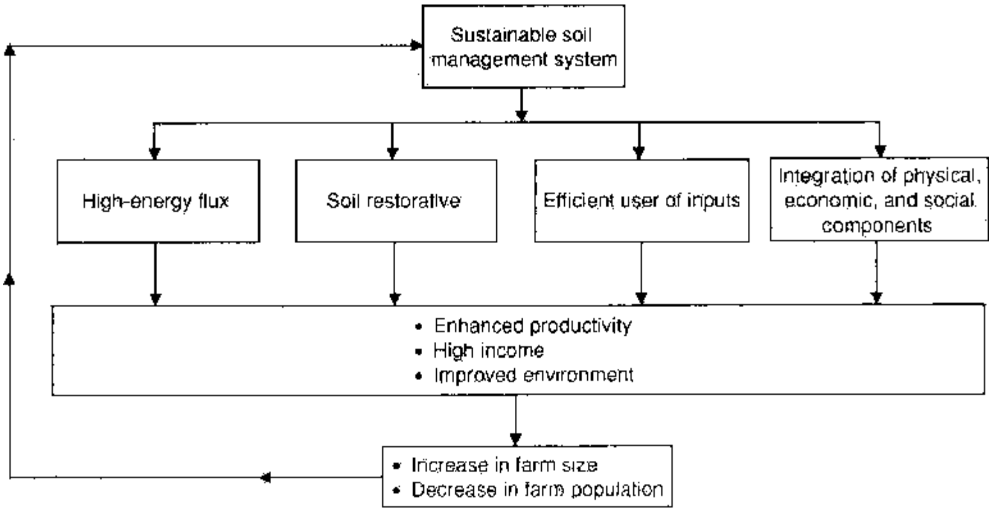


**Figure 3** Characteristics of a viable alternative to shifting cultivation systems.

## Examples of Sustainable Soil Management Systems

Farming systems, the right combination of several components and subsystems, are soil- and site-specific. Therefore, it is difficult to provide examples of generic farming systems that may be applicable to diverse soils and environments. A logical approach is to identify integral components that can be combined into site-specific farming systems to meet the site-specific needs. Some examples of improved components and subsystems to address specific issues include the following:

1. Erosion management: Conservation-effective techniques for reducing soil-erosion risks involve the following strategies:
   a. using crop residue mulch to minimize the raindrop impact;
   b. improving soil structure through bioturbation for reducing surface crust and soil compaction, and maintaining high infiltration rate;
   c. installing terraces for decreasing the slope length; and
   d. constructing grassed waterways to dispose of excess runoff safely.

Principal techniques of erosion management are outlined in **Figure 4**. Preventive measures are based on the principles of soil management that reduce the impact of high rainfall erosivity by reducing the raindrop impact and maintaining a high infiltration rate. These measures include mulch farming techniques involving conservation tillage, using cover crops in the rotation cycle, and providing a continuous ground cover through diverse (mixed) cropping systems. Two principal constraints to a widespread adoption of conservation tillage are high cost of herbicides for weed control and lack of appropriate seeding equipment to plant through wet crop-residue mulch. Soil compaction is another impediment to adopting conservation tillage in structurally inert soils of low organic matter content.

Erosion control measures are designed to reduce runoff amount and its velocity by using engineering structures that lead to safe disposal of excess runoff. These structures are grouped under two broad categories: structures to reduce runon (e.g., diversion channels, cut-off drains), and structures to dispose of runoff safely (e.g., terraces, waterways, drop structures). Cultivation of sloping lands (slope gradient > 5%) requires using a combination of erosion preventive and control measures, especially in situations where erosion cannot be effectively controlled by preventive measures alone. In such situations, engineering structures are needed as back-up measures. However, design and installation of these structures require engineering skills and proper planning. Further, such structures are costly to install and maintain. It is also important to realize that failure of engineering structures can lead to deep gullies, mass movement, and often irreversible damage to soil and terrain.

2. Nutrient management: Once soil erosion is controlled and the risks of drought are reduced by water conservation, management of plant nutrients or soil fertility is essential to obtaining high yields. Most soils under shifting cultivation have low inherent soil fertility, and nutrient supply from off-farm sources is essential. The goal is to use an INM strategy (**Figure 5**), not by reducing external inputs but by minimizing the use of chemical fertilizers through

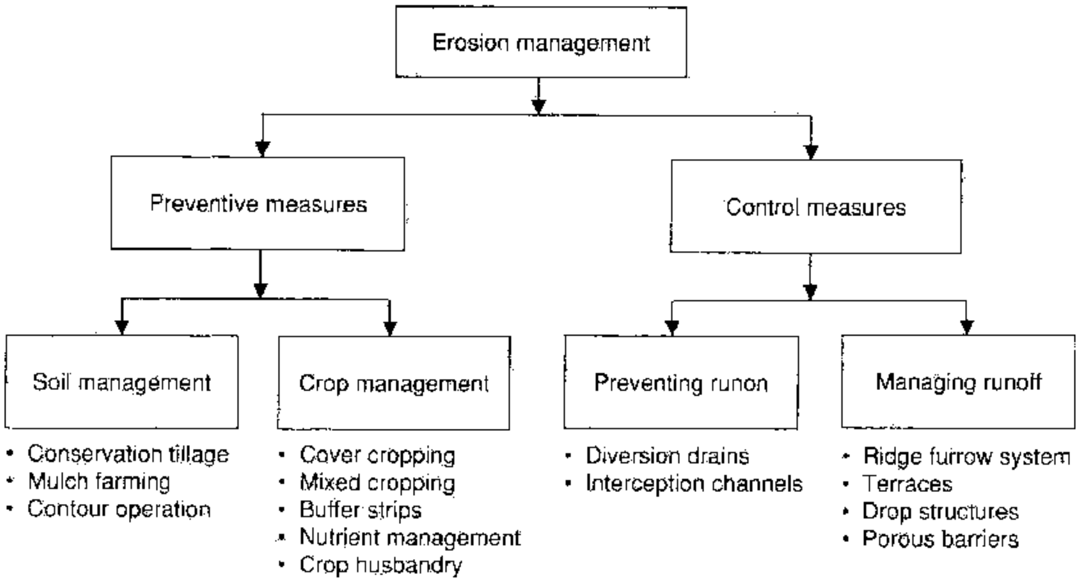


**Figure 4** Soil, crop, and runoff management techniques for erosion control.

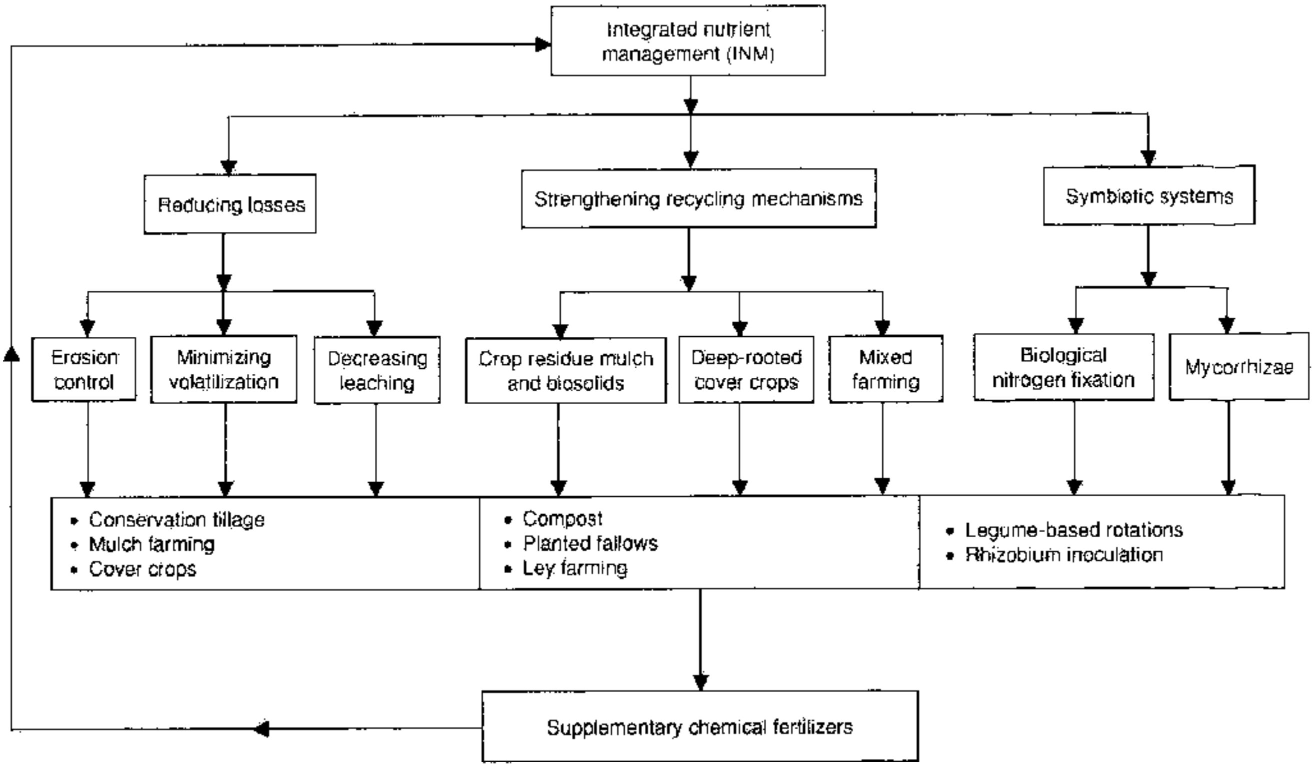


**Figure 5** Integrated nutrient-management strategies are designed to enhance nutrient supply without increasing dependence on chemical fertilizers.

decreasing losses, recycling nutrients and increasing BNF. Reducing losses of plant nutrients (in runoff, soil erosion, leaching, and volatilization) is crucial to enhancing nutrient-use efficiency. Equally important is recycling of plant nutrients through the use of biosolids as compost, sludge, and manure. Planting deep-rooted perennials, as in agroforestry systems, is also important to recycling nutrients. Soil N reserves can be enhanced through BNF by incorporating legumes in the rotation. All these measures can decrease the amount of chemical fertilizers required to obtain the desired yield. The objective is not to eliminate chemical fertilizers but to decrease the quantity required and increase the use efficiency of critical elements (e.g., N, P).

3. Water management: There is a strong link between nutrient management and water management. The nutrient-use efficiency cannot be increased without increasing the water-use efficiency. Similar to nutrient management, the overall goal is to decrease losses, strengthen water recycling, and provide supplemental irrigation (Figure 6). The latter may be needed to enhance crop yields in those regions of subhumid and semiarid tropics that are prone to short-term droughts during the growing season. In addition to losses by runoff, those by soil evaporation must be minimized, especially in the tropics, which have high evaporative demand throughout the year. Therefore, mulch farming is an important strategy. Recycling of water, similar to that of plant nutrients, also requires establishing compatible mixtures of diverse crops. Planting shallow-rooted annuals in association with deep-rooted perennials can enhance water-use efficiency. Excess runoff during the rainy season can be stored in small farm ponds and used for supplemental irrigation to prolong the growing season and alleviate drought. If supplemental irrigation involves use of canal (surface water) or tubewell (groundwater), it is absolutely necessary to make provisions for surface/subsurface drainage and reduce risks of a rise in the water table and build-up of excessive salts in the root zone.

## Ecological Approach to Management of Soils under Shifting Cultivation

The traditional shifting cultivation systems ($L = >5$) are ecologically compatible with harsh tropical environments because these systems do not cause drastic perturbations in ecosystem functions such as water balance, energy budget, elemental cycling, above- and belowground biodiversity. Therefore, any successful transformation of a resource-based (low input) and extensive (low productivity) system into viable (high energy flux) and commercial (profitable) farming must be based on similar principles of causing the least possible disturbance to ecosystem functions. Yet, the energy and biomass flow must be increased while strengthening mechanisms of recycling and

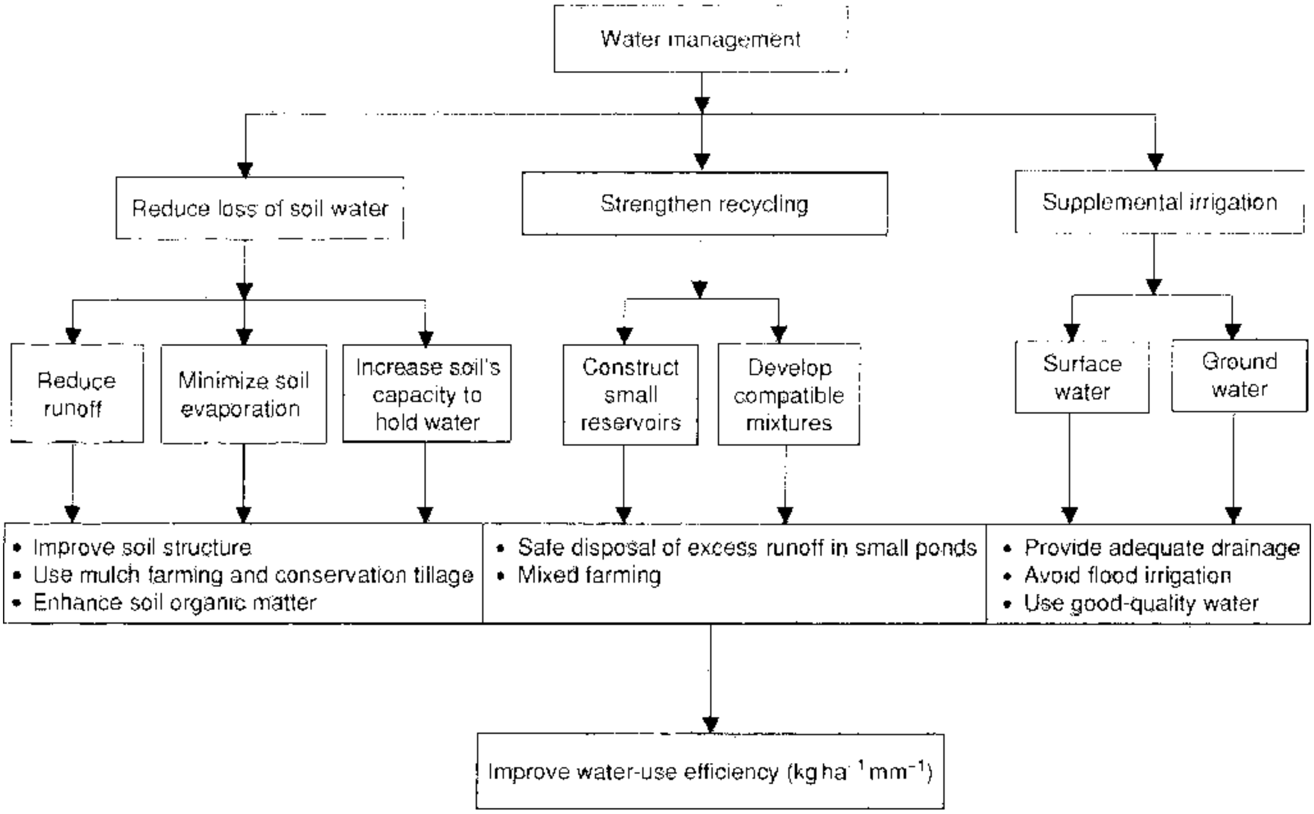


**Figure 6** Strategies for water management to enhance water-use efficiency.

minimizing loses through erosion. Essential components of an ecological approach are outlined in Figure 7, and involve science-based interventions. In addition to enhancing production, these technological interventions must minimize risks to the environment. The strategy is to increase productivity while sequestering C within soil and biota, and without releasing chemicals (fertilizers and pesticides) into the environment.

In this regard, the importance of improved germplasm cannot be overemphasized. The strategy is to introduce ecologically compatible crop/tree species and genetically improved varieties. Indeed, some of the soil-related constraints (e.g., toxicity of Al and Mn in the subsoil) may be alleviated by the introduction of improved varieties. Damage by pests can also be reduced through introduction of improved germplasm.

## Research and Development Priorities

There is a need to adopt proven technology through on-farm adaptive research with strong participation of the farming community. There is also a need for basic research in addressing new and evolving issues.

1. Terrestrial carbon sequestration: There is a strong link between the cycles of C, N, and $H_2O$. Harvesting N in the biomass for growing food crops by shifting cultivation systems has been responsible for emissions of $CO_2$ and other greenhouse gases ($CH_4$, $N_2O$) into the atmosphere. In fact, 1.5–2.0 Pg C is annually emitted into the atmosphere by deforestation and land-use change in the tropics. Rather than a source, tropical ecosystems can be a major sink for atmospheric C. Therefore, understanding the cycling of C, N, P, S, and other elements in natural and managed ecosystems, and enhancing their retention within soil and biota are crucial to reducing emissions of greenhouse gases into the atmosphere.

In this regard, restoration of degraded soils and ecosystems is extremely important. Degraded ecosystems are severely depleted of their C and N pools, and have been sources of greenhouse gases. Restoration of these ecosystems would enhance biomass productivity and soil quality, and increase the C pool within the soil and biota. It is important, therefore, to understand elemental cycling in natural and managed ecosystems and how recycling can be enhanced, losses reduced, and use-efficiency increased. Understanding the link between water and elemental cycling is also important to identifying ecologically compatible systems.

2. Soil quality and soil resilience: There is a close link between elemental cycling and soil quality. Rapid and severe soil degradation is linked to depletion of SOC reserves and elemental imbalance, but the exact mechanisms are not widely understood. An answer to the question of how can soil degradative trends be

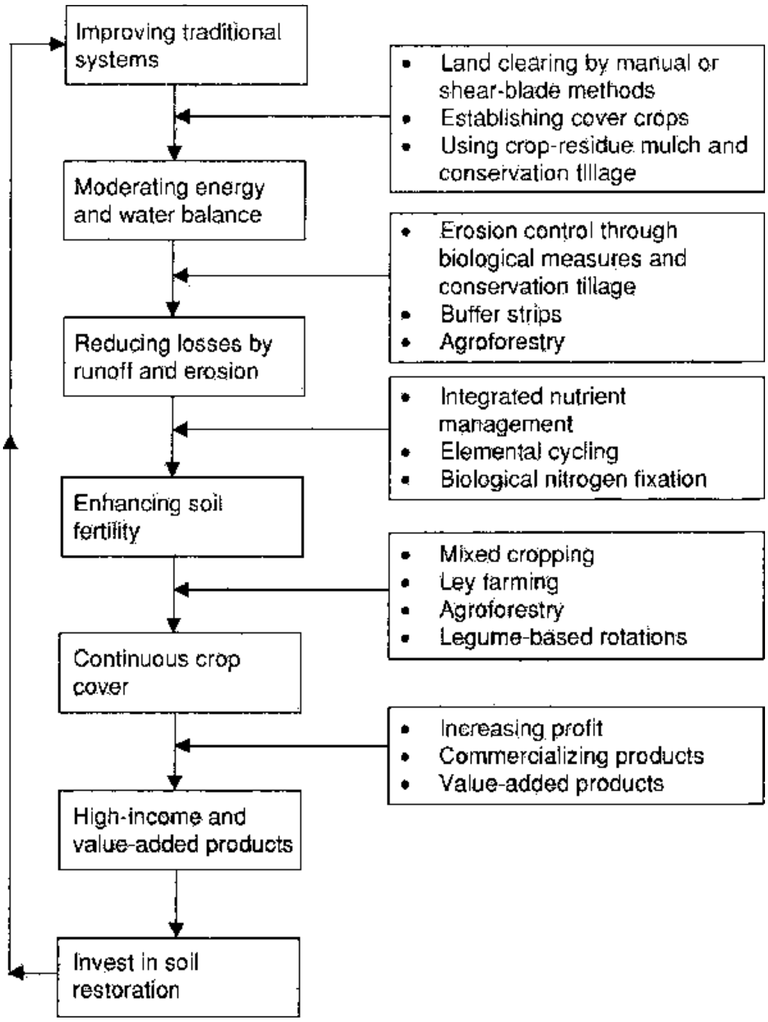


**Figure 7** An ecological approach to transforming subsistence agricultural systems.

reversed lies in understanding the relation between soil quality and soil resilience. The strategy is to enhance soil resilience and improve soil quality while improving productivity. Therefore, soil-specific research is needed to understand the mechanisms and processes responsible for enhancing soil resilience and improving soil quality. The maintenance of high levels of the SOC pool may be one such strategy.

3. Agricultural intensification: It is important to develop innovative systems of enhancing the productivity of existing agricultural lands by alleviating soil, biotic, and social/economic constraints. Restoration of productivity of depleted/degraded soils is crucial to achieving food security and improving the environment.

4. Indicators of sustainability: It is important to formulate quantitative criteria of assessing the sustainability of a farming system. Developing quantifiable and scientifically sound indicators of sustainability is a high priority.

5. Modeling: Predictive models are needed to assess the impact of land use and farming systems on soil quality, water balance, micro- and meso-climate, agronomic/biomass productivity, and sustainability. Validation of such models for specific soils and ecosystems necessitates strengthening the database on soil resources, hydrology, vegetation, and climate.

## Conclusions

Shifting cultivation and related bush fallow systems have served a useful purpose in preserving the ecological balance in regions of low population pressure. Although subsistence and low in productivity, such systems had minimal adverse impact on the environment. With increasing population pressure and concerns about environmental pollution and contamination by biomass burning and resource degradation, there is a need to identify viable alternatives to these systems. The available research indicates that productivity of these soils can be substantially increased by adopting improved and science-based technologies. Whereas soil and water conservation are achieved by biological measures based on mulch farming and cover crops, soil fertility can be enhanced by nutrient recycling, BNF, and supplemental fertilizer use. Principal attributes of sustainable systems of soil and water management are erosion control, maintenance of soil structure, increase in SOC pool and biotic activity of soil fauna, and judicious use of chemical fertilizers to create synergistic effects between fertilizers and biosolids.

Major characteristics of sustainable systems are high energy fluxes that increase productivity and use-efficiency of inputs, restorative effects on soil and environments, and increasing C stocks in soil and biota. There is a strong need to conduct basic research in understanding the principles and processes that govern soil degradation, soil quality, and soil resilience. Understanding the link between SOC dynamics on the one hand and elemental cycling (e.g., C, N, P, S) and water balance on the other is critical to developing sustainable systems that are viable alternatives to shifting cultivation.

*See also:* **Agroforestry**; **Erosion:** Irrigation-Induced; **Tropical Soils:** Arid and Semiarid; Humid Tropical

## Further Reading

Dregne HE (1982) Historical perspective of accelerated erosion and effect on world civilization. In: Schmidt BL, Allmaras RR, Mannering JV, and Papendick RI (eds) *Determinants of Soil Loss Tolerance.* ASA Special Publication 45, pp. 1–14. Madison, WI: Soil Science Society of America.

Eckholm EP (1976) *Losing Ground.* New York: Norton.

Hardin G (1968) The tragedy of the commons. *Science* 162: 1243–1248.

Harwood RR (1976) Multiple cropping in tropical Asia. In: Papendick RI, Sanchez PA, and Inplett GB (eds) *Multiple Cropping.* ASA special publication 27. Madison, WI: Soil Science Society of America.

Herdt RW (1993) Measuring sustainability using long-term experiments. Proceedings of the Conference at Rothamsted Experiment Station, 29–30 April 1993, Rothamsted, UK.

Hiebsch CK and McCollum CE (1987) Area × time equivalency ratio (ATER): a method for evaluating the productivity of intercrops. *Agronomy Journal* 79: 15–22.

IPCC (2000) *Land Use, Land Change and Forestry.* Special Report. Cambridge, UK: Cambridge University Press.

Lal R (1994) *Methods and Guidelines for Assessing Sustainable Use of Soil and Water Resources in the Tropics.* SMSS technical monograph 1. Columbus, OH: OSU.

Lal R (1995) *Sustainable Management of Soil Resources in the Humid Tropics.* Tokyo, Japan: United Nations University.

Lowdermilk WC (1953) *Conquest of the Land through Seven Thousand Years.* SCS Agricultural Information Bulletin 99.

Nye PH and Greenland DJ (1960) *The Soil Under Shifting Cultivation.* Farnham Royal, UK: Commonwealth Agricultural Bureau.

Okigbo BN (1984) Improved management production systems as an alternative to shifting intermittent cultivation. In: *Improved Production Systems as an Alternative to Shifting Cultivation.* FAO Soils Bulletin 53. Rome, Italy: FAO.

Oldeman LR (1994) The global extent of soil degradation. In: Greenland DJ and Szabolcs I (eds) *Soil Resilience and Sustainable Land Use.* Wallingford, UK: CAB International.

Olson GW (1981) Archaeology: lessons on future soil use. *Journal of Soil and Water Conservation* 36(5): 261–264.

Ruthenberg H (1971) *Farming Systems in the Tropics.* London: Oxford University Press.

Singh HP (1982) Management of desert soils. In: Kanwar JS (ed.) *Review of Soil Research in India*, pp. 679–699. New Delhi: ICAR.

# SITE-SPECIFIC SOIL MANAGEMENT

**C J Johannsen and P G Carter**, Purdue University, West Lafayette, IN, USA

## Introduction

A new approach to intensive agricultural production, called 'site-specific agricultural management,' has gained acceptance by many farmers and producers around the globe since the 1990s. Site-specific management makes use of new tools, known as 'spatial technologies,' which encourage intensive management of soil resources using tools such as global positioning systems (GPS), geographic information systems (GIS), and remote sensing. Other frequently used terms to describe this intensive agricultural system include 'precision agriculture,' 'prescription farming,' and 'variable-rate technology' (VRT). Site-specific management has been defined as:

> an information- and technology-based agricultural management system to identify, analyze, and manage a site relating to the spatial and temporal variability within fields for optimum profitability, sustainability, and protection of the environment

Another definition puts it more simply as:

> using information technologies to tailor soil and crop management to fit the specific conditions found within a field

Services are available to farmers that involve satellites for collecting data and transmitting geographic information, sensors on airplanes to provide measurements as an aid to crop and soil management, sensors on farm equipment to monitor soil and crop parameters, and computer systems for receiving or analyzing digital spatial information. The word 'spatial' is frequently used related to site-specific management and means the area represented around a data point or reference. Spatial data contain information on location, shape, and relationships among geographic features such as soil, vegetation, and hydrology. Therefore data may be collected to represent an entire field, where one sample of 220 t may be the total yield, or data may be collected to represent a small portion of the field, such as an element from a yield monitor of $4\,m^2$, where the specific yield might be $22\,t\,ha^{-1}$.

Successful application of spatial technologies at the farm scale is dependent upon having a defined management system that has the capability to handle large volumes of data in an organized way, and a solid base of experience. It arises at the confluence of three principal technologies: GPS, GIS, and remote sensing.

New technologies make data collection, analysis, and interpretation more beneficial for site-specific farming and are important in improving the soil environment. Some of these technologies include GIS for integrating data layers and data (soil properties, topography ('digital terrain'), slope, and slope aspect) to assist in the analysis of collected data; GPS for relating the location of ground observations and measurements; and remote sensing for defining soil and crop anomalies within a field and assessing their crop production impact. In addition to these technologies, improvements in communications, e.g., the internet, are providing rapid delivery of newly acquired, remotely sensed data and timely transfer of information.

## Technologies

GPS was first used during the 1980s by military forces. Composed of 24 satellites and generically identified as a GPS, Navistar blankets the globe with location-finding radio signals that can help determine a person's location to within less then 1-m accuracy. GPS assist farmers to locate areas within a field where application of fertilizer and chemicals is necessary; adjustment of application rates is based on soil conditions and yield data.

The Navistar satellite system produces a low-energy signal containing a 'data message' that includes information about its location, its atomic clock status, and its general condition. These data messages can be 'read' and interpreted by GPS receivers anywhere on the ground, at sea, or in the air. For higher accuracy, simultaneous readings are taken from four or more satellites.

The value of the GPS to a farmer includes: (1) knowing the locations of soil samples, so laboratory results can be compared with a digitized soil map; (2) knowing the locations of soil properties (clay and organic matter content) and soil conditions (relief and drainage) so that fertilizer and pesticides can be applied optimally; (3) the use of lightbar technology which provides guidance in the exact direction for accurately applying and documenting the rates of pesticides; and (4) being able to monitor and record yield data from harvesting equipment within the fields. All of these applications involve more precision in the collection of data and information, which means that the farmer or producer needs a system for keeping track of these data for future decision-making.

GIS technology is also essential to precision farming, because it is the system for keeping track of data and information. GIS is computer software that is used to collect, sort, analyze, map, graph, store, and move data. A GIS records many different types of data in stacked layers that are all precisely aligned geographically. For the farmer this means that information obtained from a field at different times can be easily integrated into a composite profile of each small unit or spatial area of the field. Not only can farmers store yield information by location, but they can also layer the yields on top of their soil map to correlate yield with soil type. A GIS also allows farmers to see costs of production for any part of the field. Accurate data are required and need to be accessed by the GIS software.

Without a GIS, data from soil tests, dates of planting, fertilizer application rates, weather, and yield are often stored as piles of paper on the farmer's desk to sort manually for making decisions. The key attribute of GIS is that all data must be referenced to a common geographic base. This is usually a base map; one of the most commonly used is digital orthophotography, which has geographic references for each pixel (picture data point) such as a longitude and latitude reference. There are more than 40 different reference systems; fortunately, software has been written to convert all of these to a selected base map. Once the base has been selected, all other data that are collected are referenced to that base. GPS is the most common means of providing the reference points for the collected data.

Remote sensing, the third technology available to precision farming, entails the use of aircraft or satellites as platforms, from which farmers capture images of their fields during the growing season. The images are used to assess whether something unusual has occurred; this information can be used to help improve yield and profit. Remote sensing technology has improved, especially since the late 1990s, to include finer spatial resolution, more wavelengths of recorded light energy, and more frequent coverage over a specific location. Because of improvements in sensors, computer chips, software, and services, intensive agricultural systems may now reap benefits from these technologies.

Farmers can either analyze this remotely sensed information themselves or rely on commercial companies to perform this service for them for a fee. Satellite images from Landsat and SPOT, US and French satellites, respectively, have been used to distinguish locations affected by such problems as weeds, excessive moisture, low soil pH, and imbalance of plant nutrients. Satellites such as the IKONOS, Quickbird, and Orbview satellites have allowed farmers to obtain the equivalent of aerial images of their fields and have already shown benefits for management of intensive crops such as vegetables, orchard fruits, and grapes.

The cost and timeliness of obtaining the images have been the biggest deterrent to date to regular use of such remotely sensed data. Satellites launched during the last few years by Digital Globe, Space Imaging, and others are competing for the agricultural market; however, there are not enough satellites available to provide cloud-free images for critical periods. These multispectral satellites provide images in three to six different frequencies plus black-and-white images at 0.6- to 4-m spatial resolutions.

Remote sensing needs to prove itself in the grain crop areas, since the profit margin of these crops has been low. It has, however, been tested successfully on potato, sugar beet, and cotton fields in some parts of the world. Commercial remote-sensing companies have used weekly images to map the acreage of these crops and monitor their conditions throughout the growing season. The images and interpretations provide growers with an improved bargaining position when contracting with food processors, since they have a prediction of harvest dates and knowledge of how their competitor's crops are performing.

The value of precision farming is that the farmer can perform more timely operations related to better soil management such as tillage, plan more accurate crop-protection programs, adjust seeding rates, and know the yield variations commonly caused by soil differences within a field. The potential value of this system therefore translates into increased net return per acre; improved efficiency of such inputs as seeds, fertilizers, and chemicals; improvement of the soil environment; reduced threat to water quality; and increased managerial skills. This potential, while promised in many farm advertisements, still needs to be proven in a consistent manner and depends on the grower's management style. Farmers need to learn how to use these technologies to improve soil and crop quality based upon conditions found within their fields.

## Soil Management

Soil management is a key to the success of site-specific soil management. It starts with a farmer's ability to vary the depth of tillage according to soil conditions and is important in proper seedbed preparation, control of weeds, and fuel consumption, with the potential to lower production costs within an individual field. With a GPS as their guide, farmers using conservation tillage (the practice of leaving residues on the soil surface for erosion control) can adjust tillage depth as they traverse the different soil types. With conservation and reduced tillage, the amount of soil disturbance is minimal, but adjustments according to soil conditions such as texture, moisture content, and pH are important to seed depth and fertilizer placement. The adjustments may contribute to higher yields and safer production at lower cost. This part of site-specific farming is in its infancy. Equipment companies are developing tillage equipment with GPS and selected controls tailored to site-specific farming that will be evolving over the next decade.

The application of chemicals in proper proportions is of environmental and economic concern to farmers. Environmental regulations call for the discontinuance of certain pesticide applications within 30 m of a stream, water body, or well, or within 20 m of an intermittent stream. Using a GPS along with a digital drainage map, a farmer is able to apply these pesticides with VRT in a safer manner (**Figure 1**). Spraying equipment can be preprogrammed to turn off automatically when it reaches the distance limitation or buffer along the drainage feature. Additionally, farmers can preprogram the amount of pesticide to be applied according to the soil characteristics in each area of the field. This saves money and allows for safer use of these materials.

Some companies are marketing sprayers utilizing a sensor positioned ahead of each spray nozzle for applying herbicides in chemical-fallow conditions

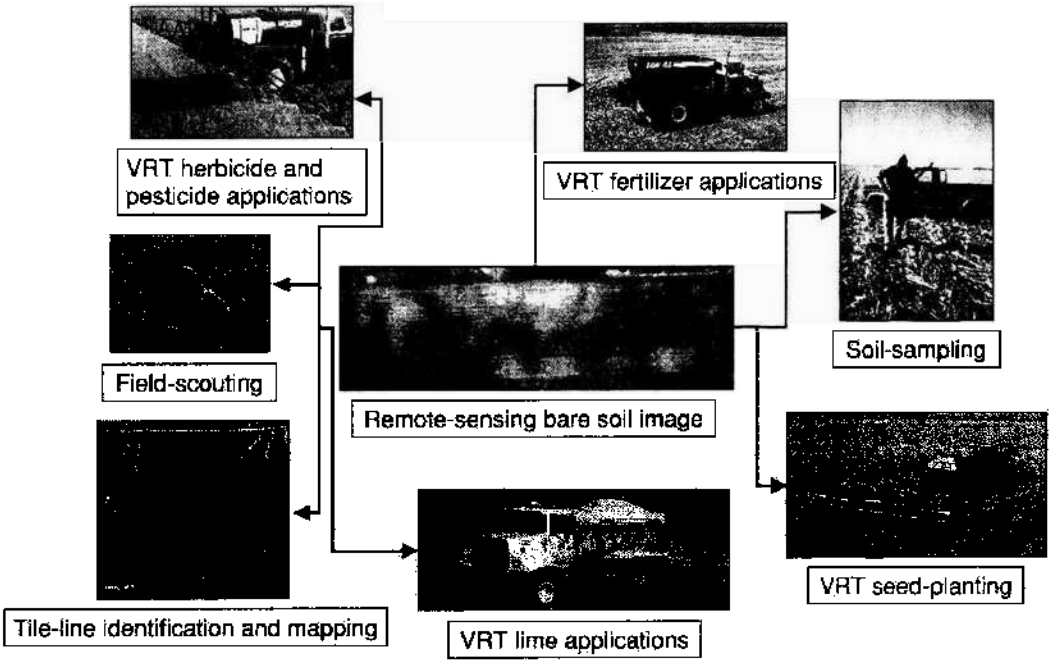


**Figure 1** Agricultural global positioning system (GPS) applications as related to a remotely sensed image. The image provides the information to develop variable-rate technology (VRT) application maps.

where a crop is grown every other year. Incident light sensors detect and measure green vegetation, and, when the vegetation, e.g., unwanted plants or weeds, takes up more than 5% of a detector's field of view, a controller automatically turns on the individual spray nozzle. This type of sprayer, which greatly reduces the area sprayed and expenditure on chemicals, is a good conservation tool.

Similarly, nutrient applications of nitrogen and other materials are gaining acceptance owing to effective soil testing, such as taking samples according to soil maps or soil management zones. The application of lime to correct soil acidity by development of application maps based on soil samples has gained acceptance among many farmers.

## Proper Seeding

Crops grow best when seeds are spaced such that plants obtain maximum sunlight and moisture and are placed at a proper depth within the soil. This is best accomplished by varying the seeding rate according to soil conditions such as texture, organic matter, and available soil moisture. For example, fewer seeds would be planted in sandy soil than in silt loam soils due to the amount of available moisture. To maximize yield per plant, seeds would be sown at a lower density, which would usually produce larger heads (ears) of harvested seeds because of more soil nutrients and sunlight available per plant.

However, planting fewer seeds runs counter to the practice of increasing seed density to increase yields, a trajectory that when overdone can yield a biomass so dense that some sunlight is blocked and the plants receive insufficient sunlight to produce seeds. The farmer's challenge is to find the right seed density to match all the localized factors affecting plant growth. Since soils can vary significantly across one field, the farmer can change seeding rates while crossing the field and thereby maximize the seeding rate according to the soil conditions.

Equipment modifications needed for VRTs have been designed for corn planters, grain drills, and air seeders (**Figure 1**). The desired rates are usually determined by reviewing previous yield images and/or soil productivity estimates. A computerized soil map of a specific field on a computer fitted on the tractor along with a GPS can provide farmers with their exact position in the field and guide them in adjusting the seeding rate accordingly. Seed companies are beginning to make recommendations on various seeding rates for individual hybrids, which will help the farmer with the settings on the planter.

## Harvesting

A key benefit in the use of site-specific management techniques such as the adjustment of seeds, pesticides, fertilizers, and tillage as a field is crossed is in knowing the precise yields at every location. Harvesting equipment can be equipped with yield-monitoring devices that are coupled to a GPS, which provide a yield measurement for specific areas. A yield amount and location can be obtained every 1–5 s as the field is traversed. With appropriate software, a yield map

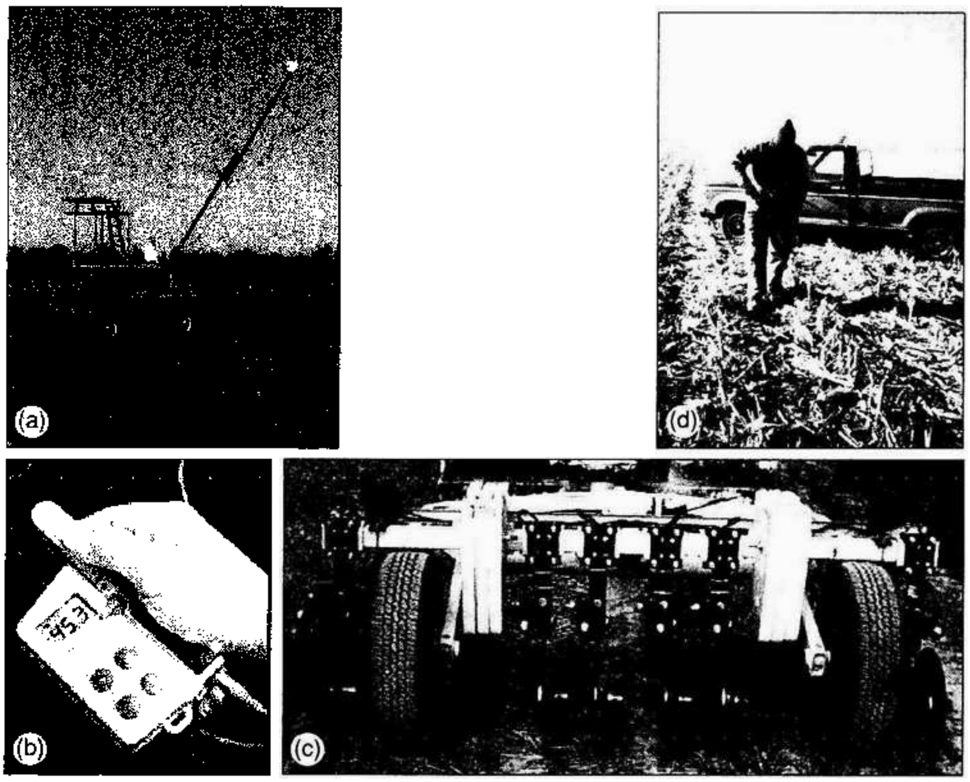


**Figure 2** Collection of soil and plant measurements in the field for site-specific management: (a) field spectrometer; (b) chlorophyll meter; (c) electrical conductivity; (d) soil-sampling.

can be produced showing the yield variation throughout the field. This allows farmers to inspect the precise location of the highest- and the lowest-yielding areas of the field and determine why the differences exist. It allows the analysis of costs and yield to determine the most profitable practices and rates that apply to each field location. This provides farmers with their own research information that can be used in a practical and profitable way.

For farmers wanting to adopt precision farming, yield monitors are a good place to start. Yield monitors are already in use for measuring corn, soybeans, cotton, and other small grains such as oats and wheat, and monitors for other crops such as potatoes and peanuts are being developed. Farmers are also encouraged to monitor their fields closely by scouting (walking the field) and using remote-sensing techniques so that they have proof of problems or changes. By using a GPS unit they can go to a specific low-yield location and soil sample to determine whether lack of specific nutrients is the cause (**Figure 2**). With this information, farmers can assess the cost-effectiveness of the additional input. Some low-yield areas might best be ignored for growing crops.

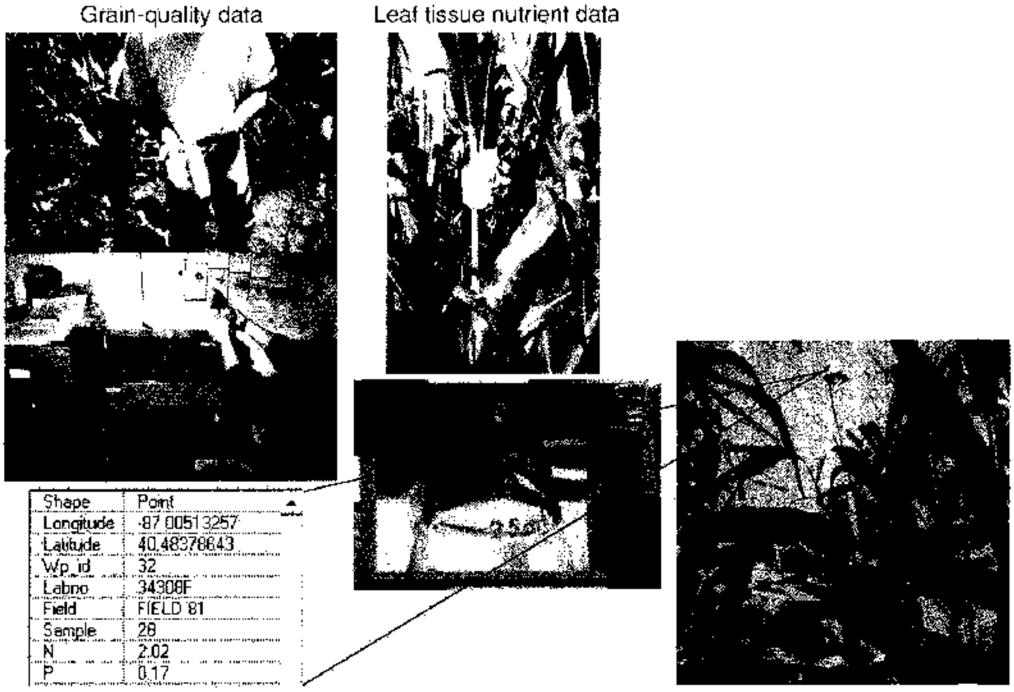
## Adoption Considerations

Farmers must know how the technology operates and be computer-literate. If they are working with a crop consultant who can analyze their data and make recommendations for them, that is a start, but being computer-literate is an advantage (**Figures** 3 and 4).

The number of farmers involved in precision or site-specific management practices around the globe is not known. It is likely that approximately 50% of the farmers in developed countries have had exposure to the use of VRT applications of soil-nutrient amendments and pesticides. Many new combines have yield monitors as standard equipment, which means that the opportunity to learn how to use yield images and/or maps increases with each combine that is purchased. The use of the yield images has caused farmers to request remotely sensed images during the growing season for comparison. Most are hoping to determine the crop yield prior to harvest and in the process are learning that this is difficult to achieve; however, the knowledge of the locations of crop and soil anomalies is beneficial to improving their management skills. Current trends in the use of remote sensing data in agriculture are driven by goals that are part of site-specific crop management. Farmers can adopt precision-farming techniques through cooperatives and through consultants who use GPS technologies (**Figures** 3 and 4).

## The Future

Several companies are starting to market GIS recordkeeping systems so farmers can record all of the field operations, including planting, spraying, cultivation,



| | |
|---|---|
| Shape | Point |
| Longitude | -87 00513257 |
| Latitude | 40.48378643 |
| Wp_id | 32 |
| Labno | 34308F |
| Field | FIELD 81 |
| Sample | 28 |
| N | 2.02 |
| P | 0.17 |

**Figure 3** Global positioning system (GPS) technology is used to record the field locations where ground reference plant samples are collected. GPS locations make it possible to relate sample results to remote sensing image data.

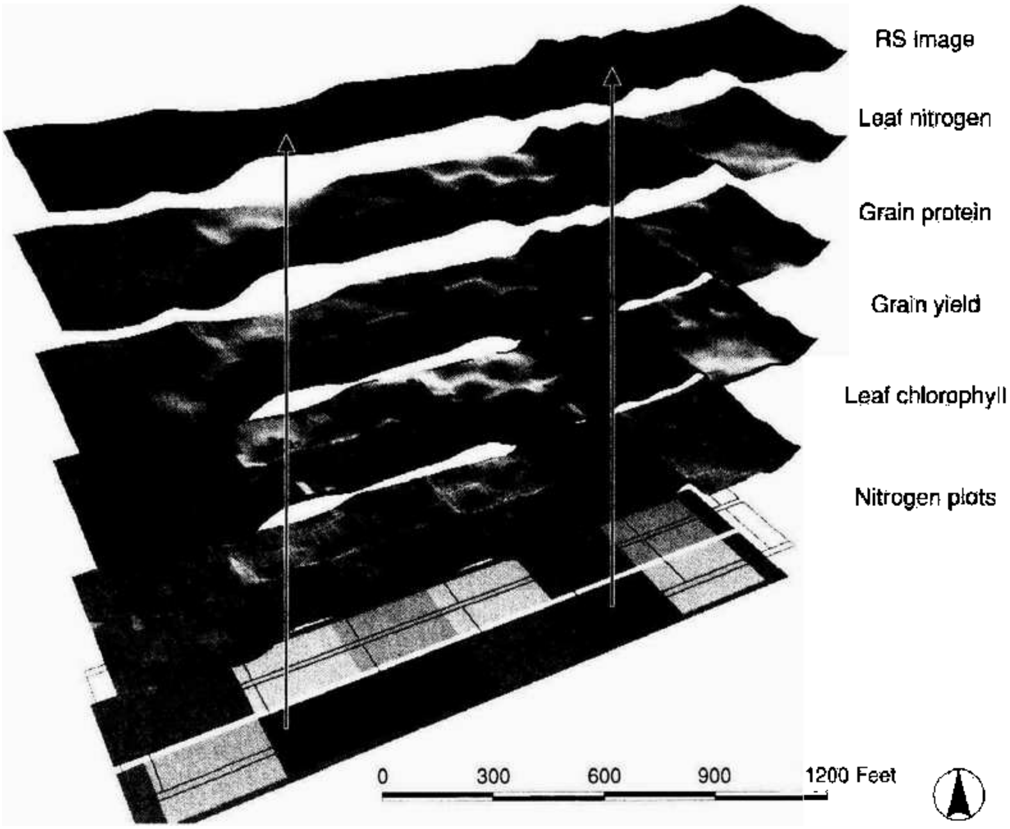


**Figure 4** GIS layers in a digital terrain format showing a remotely sensed (RS) image and measured plant properties for a field experiment testing four rates of nitrogen that would provide two inadequate rates, an adequate and an excessive rate within three random replicated corn plots.

and harvest (along with specific information such as type of equipment used, usage rates, weather information, time of day performed, and weather data). Additionally, farmers can record observations such as weed growth, unusual plant stress, coloring, and growth conditions throughout the season. All GPS-collected data can be automatically recorded and used in conjunction with a GIS program.

Remotely sensed data can be analyzed and added to the GIS using soil maps and digital-terrain and field-operations information as ground references. This can be used to guide further field operations such as spraying, fertilizing, and irrigating; plus it is part of a permanent record. A number of new satellites provide data with increased spatial resolution such as 1, 5, and 10 m, which is similar to digital aerial images. This means that remotely sensed data would complement the yield data collected.

Additionally, the spectral resolution will be increased through hyperspectral image data, meaning that, instead of a few broad bands of 100 nm bandwidth, there will be more than 100 narrow bands of 5 and 10 nm bandwidth. This helps to define the colors more sharply, which leads to more detail about leaf pigments, soil organic matter, and other plant and soil properties. Much research is being done to determine precisely what is being measured.

On the horizon are sensors operated from the tractor that will measure soil properties such as organic matter, clay content, and soil moisture in real time. This will have a significant impact on planting operations. It means that, based on measurements from sensors mounted on planters, farmers will be able to adjust seeding, fertilizer application, and chemical application rates as they plant. By recording these inputs and tracking rates by location, they will be able to overlay the yield information and calculate the precise cost of production for any portion of their fields. Precision farming is making a strong impact on the way farmers in highly mechanized systems are able to protect their soil environment and manage their farm operations. The steady growth of remote sensing, GIS, and GPS technologies will continue to encourage site-specific soil management.

*See also:* **Conservation Tillage**; **Factors of Soil Formation:** Biota; **Geographical Information Systems**; **Remote Sensing:** Organic Matter; Soil Moisture

## Further Reading

Johannsen CJ (2001) Glossary of terms for precision farming. In: Lowenberg-DeBoer JM and Erickson KM (eds) *Precision Farming Profitability*, pp. 125–132. Lafayette, IN: Agricultural Research Programs, Purdue University.

Johannsen CJ, Baumgardner MF, Willis PR, and Carter PG (1998) *Advances in Remote Sensing Technologies and their Potential Impact on Agriculture.* 1st International Conference on Geospatial Information in Agriculture and Forestry. Proceeding vol. 1, pp. 6–11. Ann Arbor, MI: Verdian International.

Johannsen CJ, Petersen GW, Carter PG, and Morgan MT (2003) Remote sensing changing natural resource management. *Journal of Soil and Water Conservation* 58: 42–45.

Lowenberg-DeBoer JM and Erickson KM (eds) (2001) *Precision Farming Profitability.* Publication SSM-3, Agricultural Research Programs. West Lafayette, IN: Purdue University Press.

Medlin CR, Hawkins SE, Morris DK, and Parsons S (2001) Variable-rate technologies. In: Lowenberg-DeBoer JM and Erickson KM (eds) *Precision Farming Profitability.* Publication SSM-3, Agricultural Research Programs. West Lafayette, IN: Purdue University Press.

Moran MS, Inoue Y, and Barnes EM (1997) Opportunities and limitations for image-based remote sensing in precision crop management. *Remote Sensing of the Environment* 61: 319–346.

Plant RE (2001) Site-specific management: the application of information technology to crop production. *Computers and Electronics in Agriculture* 30: 9–29.

Robert PC and Stafford J (2003) *Precision Agriculture: An International Journal on Advances in Precision Agriculture* 4: 1–444.

Strickland RM and Parsons S (2001) The basics of geographic information systems. In: Lowenberg-DeBoer JM and Erickson KM (eds) *Precision Farming Profitability*, pp. 97–102. Publication SSM-3, Agricultural Research Programs, West Lafayette, IN: Purdue University Press.

## Slash and Burn Agriculture *See* Shifting Cultivation

## Sludge *See* Waste Disposal on Land: Liquid; Municipal

# SODIC SOILS

**G J Levy and I Shainberg**, The Volcani Center, Bet Dagan, Israel

## Introduction

Presence of sodium (Na) as an exchangeable cation is found in many soils worldwide. Sodic soils exhibit poor soil–water and soil–air relations, which in turn adversely affect root growth and plant production, increase soil erosion, and make the soils difficult to farm. Problems associated with sodic soils are expected to increase. The need to provide food to an ever-expanding population, coupled with the increasing demand for good-quality water from the urban and industrial sectors, results in poorer-quality water (e.g., treated effluents) and soils being used for food production. Understanding the behavior of sodic soils and its effects on agriculture and the environment is, therefore, essential in order to manage sodic soils properly for securing crop production and maintaining sustainable agriculture worldwide.

## Definition of Sodic Soils

Two important parameters are used to define sodic soils. The first is a soil property – the exchangeable sodium percentage (ESP), which describes the level of adsorbed Na in soil:

$$\text{ESP} = (\text{exchangeable Na}/\text{cation exchange capacity}) \cdot 100 \qquad [1]$$

Instead of cation exchange capacity it is also possible to use the sum of exchangeable cations (Ca + Mg + Na + K). The second parameter which indicates the sodicity level of the irrigation water or soil solution is termed sodium adsorption ratio (SAR):

$$\text{SAR} = (\text{Na})/((\text{Ca}+\text{Mg})/2)^{0.5} \qquad [2]$$

where parentheses reflect cation concentration in $\text{mmol}_c\ \text{l}^{-1}$. Thus, SAR has units of $(\text{mmol}_c\ \text{l}^{-1})^{0.5}$. It must be emphasized that the sodicity hazard of solutions is related to the ratio of Na to the divalent cations present in the water and not to Na concentration solely.

Exchange reactions take place between the soil solution and the exchangeable phase. Thus, soil ESP can be estimated from the SAR of saturated paste extracts using the following empirical relationship:

$$\text{ESP} = 100(-0.0126 + 0.01475 \cdot \text{SAR})/(1 + [-0.0126 + 0.01475 \cdot \text{SAR}]) \qquad [3]$$

or from the nomograph presented in **Figure 1**.

To date there is no widely accepted critical ESP value which separates sodic from nonsodic soils. This is because the effects of ESP cannot be separated from that of the electrolyte concentration (EC) of the soil solution. The interrelationship between ESP of the soil or SAR of its equilibrium solution and the EC of the soil solution may help estimate whether a sodic or nonsodic behavior will occur.

## Origin and Distribution of Sodic Soils

In many sodic soils sodicity originates from the nature of the parent material and pedogenic processes. There are also sodic soils where sodicity arises from anthropogenic processes (human activity), and this is termed secondary sodification. Irrigation with sodic water or without proper drainage, forest clearing, and other land management practices that lead to water logging

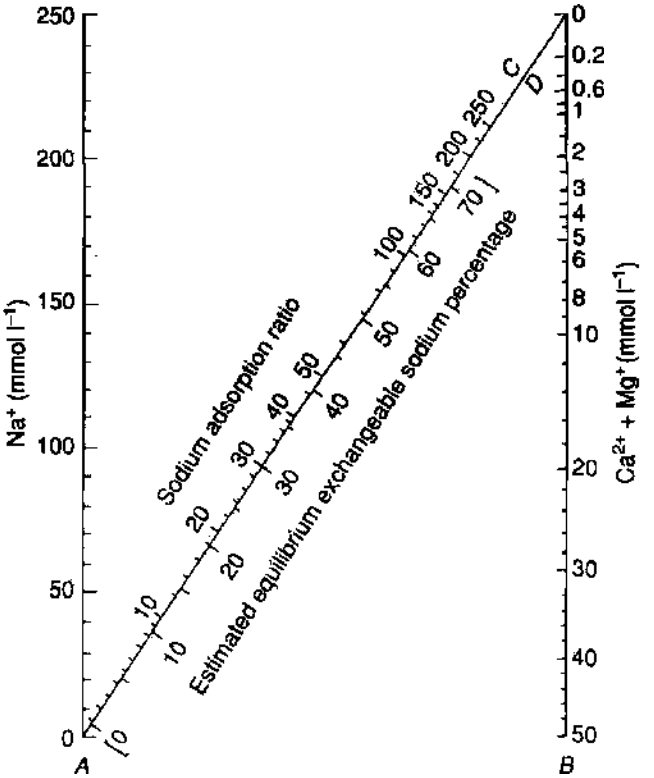


**Figure 1** Nomograph for determining the sodium absorption ratio of a saturation extract and estimating the corresponding exchangeable sodium percentage value of soil at equilibrium with the extract. Reproduced from US Salinity Laboratory Staff (1954) *Diagnosis and Improvement of Saline and Alkali Soils.* Agricultural Handbook 60. Washington, DC: USDA.

**Table 1** World distribution of sodic soils

| *Continent* | *Country* | *Area of sodic soils (thousand hectares)* |
|---|---|---|
| North America | Canada | 6974 |
| | USA | 2590 |
| South America | Argentina | 53 139 |
| | Bolivia | 716 |
| | Brazil | 362 |
| | Chile | 3642 |
| Africa | Botswana | 670 |
| | Cameroon | 671 |
| | Chad | 5950 |
| | Ethiopia | 425 |
| | Kenya | 448 |
| | Madagascar | 1287 |
| | Namibia | 1751 |
| | Niger | 1389 |
| | Nigeria | 5837 |
| | Somalia | 4033 |
| | Sudan | 2736 |
| | Tanzania | 583 |
| | Zambia | 863 |
| South Asia | Bangladesh | 538 |
| | India | 574 |
| | Iran | 686 |
| North and Central Asia | China | 437 |
| | USSR | 119 628 |
| Australia | Australia | 339 971 |

Reproduced from Massoud FI (1977) Basic principles for prognosis and monitoring of salinity and sodicity. *Proceedings of an International Management of Saline Water for Irrigation*, pp. 432–454. Lubbock, TX: Texas Technical University.

are key activities that may yield rapid secondary sodification.

The total area worldwide of sodic soils is estimated at 210 million ha. The area of sodic soils in each continent and in some countries where they form an important proportion of the soils is presented in Table 1. The spatial distribution of sodic soils indicates that they occur in a range of climates. However, the outlined distribution of sodic soils and the conditions promoting their formation are based on a pedological approach. Following the increase in the contribution of human activity to soil sodification, the definition of sodic soils is now based on soil behavior. Many agricultural soils that exhibit sodic behavior may not fit into the classical sodic soil group. Consequently, the sizeable area these soils cover remains unaccounted for.

## Mechanisms Responsible for Sodic Behavior

Among the various soil constituents, the fraction that considerably determines the physical behavior of soils is the colloidal clay. Compared with other soil constituents, the clay fraction possesses a large specific surface area (surface area per gram of clay), which, combined with its charge, makes the clay a very reactive soil constituent in physicochemical processes, mainly swelling and dispersion. These two processes largely determine the microstructure of the soil, thus significantly affecting many of its physical properties.

The intensity of swelling and dispersion processes and the magnitude of their adverse effects on soil properties become more severe with the increase in soil sodicity. However, the swelling and dispersion behavior of soil clays differs significantly from that of pure clay systems, possibly because soil clays usually occur as mixtures and because of their association with other minerals, oxides, and organic matter present in the soil. It has been demonstrated that the EC required for preventing dispersion of soil clays is two- to 10-fold higher than that for pure clays. On the other hand, the presence of hydrous oxides or sparingly soluble minerals such as $CaCO_3$ has resulted in lower levels of soil clay dispersivity. Thus, extrapolation from pure clay systems to soil systems may often prove problematic.

## Structural Stability of Sodic Soils

Soil structure is commonly referred to as the size, shape, and arrangement of particles (micro- and macroaggregates) and pores in soil. Soil structural stability is therefore the ability of the soil to retain its arrangement of solids and pores when exposed to different stresses. The presence of high levels of sodicity aggravates aggregate susceptibility to breakdown and slaking. However, these effects of sodicity are intricate and should not be considered independently of the EC of the soil solution and soil texture (**Figure 2**).

Sodicity-related degradation of soil structure leads to deterioration in soil water transmission properties, and increased susceptibility to crusting, runoff, and erosion. Water movement in the soil profile is commonly characterized by its hydraulic conductivity (HC). In cases where a thin seal is formed at the soil surface, water penetration into the soil is characterized by the infiltration rate (IR) rather than HC. Because the properties of the seal differ from those of the underlying soil, and its thickness is difficult to measure, this makes the determination of its HC extremely difficult.

### Hydraulic Conductivity

The HC of a soil depends on its properties and those of the percolating fluid. Existing models have failed in adequately predicting soil HC, largely because they do not take into account the effect of fluid properties

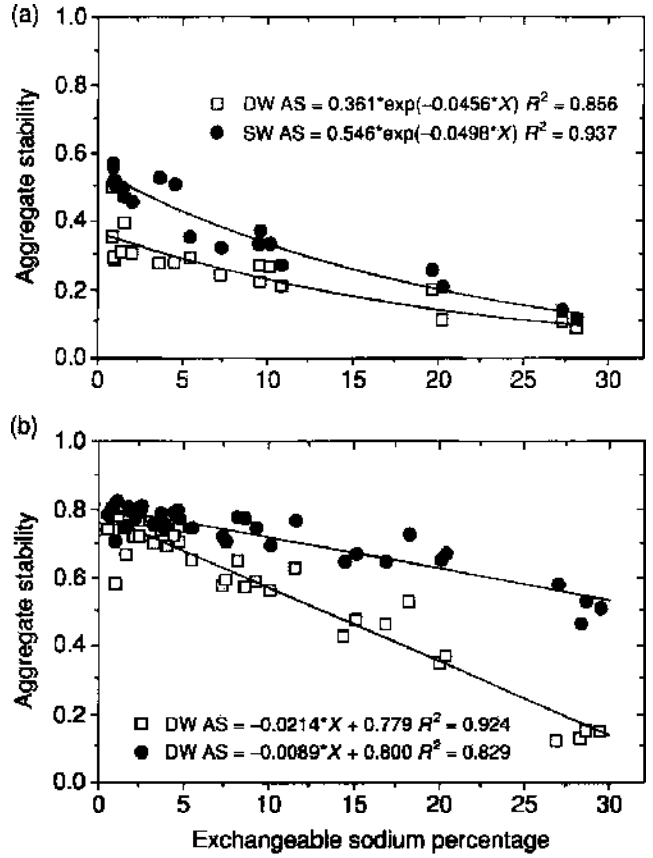


**Figure 2** Aggregate stability (AS, expressed as a fraction) as a function of exchangeable sodium percentage and electrolyte concentration for (a) soils with clay content ≤25% and (b) soils with clay content ≥35%. DW, distilled water; SW, saline water. Modified from Levy GJ, Mamedov AI, and Goldstein D (2003) Sodicity and water quality effects on slaking of aggregates from semi-acid soils. *Soil Science* 168: 552–562.

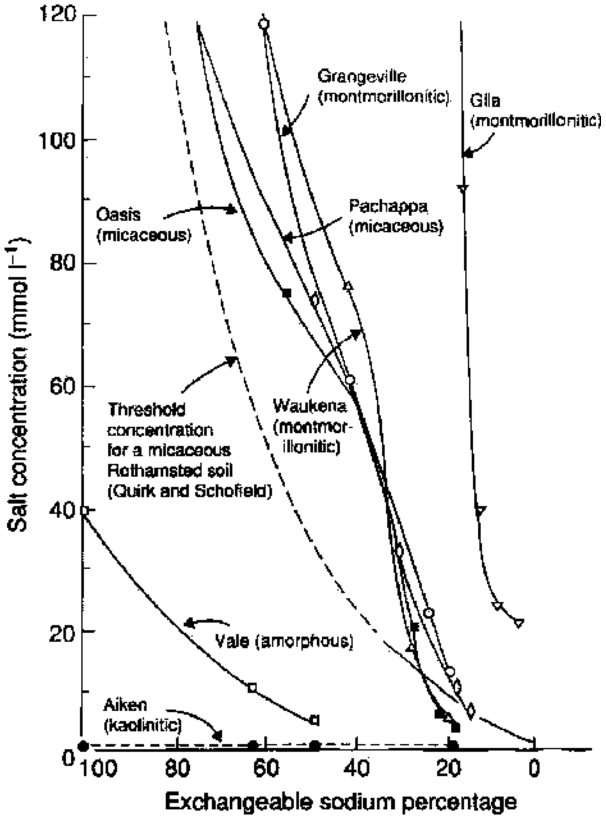


**Figure 3** Combinations of salt concentration and exchangeable sodium percentage required to produce a 25% reduction in hydraulic conductivity for selected soils. Reproduced with permission from McNeal BL and Coleman NT (1966) Effect of solution composition on soil hydraulic conductivity. *Soil Science Society of America Proceedings* 30: 308–312.

on the soil matrix geometry. Swelling and dispersion of clay in response to sodic conditions and the level of EC in the percolating solution change the size of the conducting pores and hence affect the HC.

**Sodicity and electrolyte concentration** The realization that soil HC depends on both the SAR and the EC of the percolating solution led to the development of the 'threshold concentration' concept, which was defined as the EC required to prevent a given decrease in the HC for a given soil ESP or SAR of the percolating solution (Figure 3). This concept has been tested and validated on a large variety of soil types. It explains observations of decreased HC in soils, even Ca-soil and soils of low ESP, exposed to water of low salinity (e.g., rain or snow water). The decrease in HC is because salt concentration in the soil solution is not sufficient to prevent swelling and clay dispersion. For instance, in Australia under rain-fed conditions, soils (especially clayey ones) often waterlog during the rainy season because what little salt is present is leached in the first rain, causing the clay to disperse and the HC to drop sharply. Conversely, it is possible to maintain high HC in an Na-saturated soil provided the EC is sufficiently large. In Israel, experience from commercial fields showed that, for Na-affected soils, irrigation with water having an SAR value as high as 26 with an EC of 4.6 dS $m^{-1}$ exhibited no permeability problems, as the electrolyte concentration in the irrigation water was sufficient to prevent the dispersive effect of Na. However, upon applying electrolyte-free water (to simulate rainwater in winter), soil HC dropped to a small fraction of its initial high value.

A distinction should be made between the effects of clay swelling and those of clay dispersion on the HC of soils. Soils begin to swell as the EC of the soil solution is reduced, but the effect of swelling on HC becomes apparent only at SAR values above 10 in medium- and heavy-textured soils. In light-textured soils the effect of swelling on reducing the conducting pores and hence HC is hardly noticeable. Furthermore, the swelling process is reversible; thus an increase in HC can be noted following an increase in EC of the soil solution. Clay dispersion, on the other hand, can take place at already low ESP levels provided the EC is below a critical level. When clay dispersion occurs, the dispersed clay particles move down the soil profile and may even cause a complete blockage of the conducting pores, and hence an irreversible change in the HC. Consequently, the effects of reducing EC on the HC are likely to be more severe when clay dispersion, rather than swelling, is the predominant cause.

**Soil properties affecting sodic behavior** The response of soils to sodic conditions varies among soils, mainly because soil properties have a significant effect on the dependence of the HC on sodicity. A major factor causing differences among various sodic soils in their susceptibility to hydraulic failure when leached with low-EC water is their rate of salt release from mineral dissolution. Mineral dissolution determines the electrolyte concentrations of percolating solutions and hence soil HC. Sodic soils containing minerals ($CaCO_3$ and a few primary minerals) that readily release soluble electrolytes will not readily disperse when leached with distilled water (DW) at moderate ESP values, because a sufficiently high electrolyte concentration ($\sim$3 mmol$_c$ l$^{-1}$) can be maintained to prevent clay dispersion. In addition, the ESP level will decrease, because most of the cations released are divalent (Ca and Mg).

Clay content, through its effect on pore size distribution, contributes significantly to the response of soils to sodic conditions. The sensitivity of HC to sodic conditions is generally greater for soils having medium and high clay content. When clay content is high, clay swelling may significantly reduce the size of the water-conducting pores and lead to low HC. Similarly, under dispersive conditions the dispersed clay particles clog the pores and cause a reduction in soil HC. Conversely, in sandy soils, due to the low clay content, clay swelling has only a limited effect on the HC. Conditions supporting clay dispersion may cause the HC first to decline and subsequently to recover. This phenomenon is explained by the change in flow pattern, from a solution flowing through a matrix of sand covered with clay which starts to disperse and reduce HC, to a flow of a clay suspension in a pure sandy matrix having large pores resulting in subsequent increase in HC.

Other soil properties that may affect the response of soils to sodic conditions include:

1. Presence of aluminum and iron oxides: These oxides, commonly found in highly weathered soils, enhance soil stability, via a cementing action, and substantially decrease its susceptibility to sodic conditions.
2. Clay mineralogy: Soils with a high content of 2:1 layer silicates (e.g., montmorillonite) are the most labile ones while those high in kaolinite and sesquioxides are the least labile.
3. Exchangeable Mg: Although in many soils the exchangeable Ca/Mg molar ratio is high, in some soils the reverse is true. The effect of adsorbed Mg on soil hydraulic properties seems to depend on soil type. In calcareous soils, the presence of Mg enhanced the dissolution of $CaCO_3$, producing electrolytes which prevent clay dispersion and HC decay, while in noncalcareous soils Mg causes a decrease in the HC of the soils beyond that of the Na/Ca system. There is the direct ('specific') effect of exchangeable Mg in causing a decrease in HC, and there is the inability of Mg in irrigation water to counter the accumulation of exchangeable Na in the soil.

**Sodicity and conditions prevailing in the soil** It has recently been recognized that the soil response to sodic conditions depends not only on soil properties but also on the conditions prevailing in the soil (e.g., antecedent moisture content and wetting rate). Soil conditions affect aggregate stability/breakdown and thus the pore size distribution and consequently the HC. Aggregate breakdown results from dispersion and from slaking during wetting, whose rate determines the degree of slaking. It has been observed that the HC of sodic soils leached with DW decreased more steeply and to lower values with the increase in the rate at which the soil was wetted. Fast wetting, which increases aggregate slaking, increases the susceptibility of the soils to sodic conditions. Conversely, slow wetting decreases significantly the effects of sodicity on soil HC, especially in fine-textured soils (**Figure 4**).

### Infiltration Rate

Soil infiltration rate (IR) is defined as the volume flux of water flowing into the profile per unit of soil surface area under any set of circumstances. Often when water is supplied to soil, the IR decreases from its initial high rate due to formation of a thin layer (<2 mm) at the soil surface, termed seal. The discussion to follow will focus on structural seals usually formed by the impact energy of water drops.

Sensitivity of seal formation and IR of a smectitic loamy sand to ESP when exposed to distilled water under laboratory-simulated rain is presented in **Figure 5**. Even at the lowest sodicity (ESP 1.0), a seal was formed and the IR dropped from an initial value >100 to a final infiltration rate (FIR) of 7.0 mm h$^{-1}$. An ESP value of 2.2 was sufficient to cause a further drop in the final IR of the sandy loam to 2.4 mm$^{-1}$. The amount of rain required to approach the FIR (i.e., the rate of seal formation) was

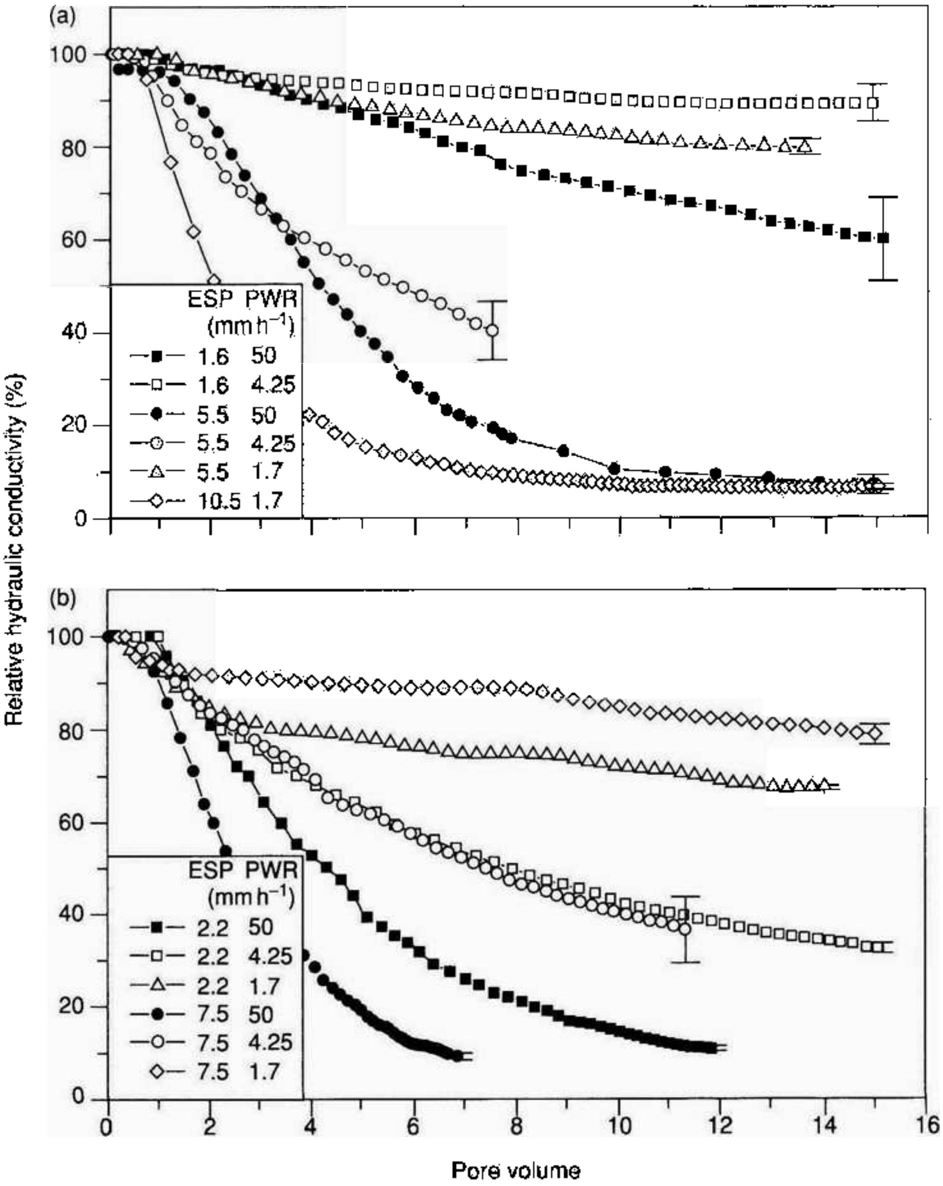


**Figure 4** Relative hydraulic conductivity of (a) sandy clay with exchangeable sodium percentage (ESP) of 1.6, 5.5, and 10.5, and (b) clay Y with ESP 2.2 and 7.5 as a function of three prewetting rates (PWR: 50, 4.25, and 1.7 mm h$^{-1}$) and leaching with deionized water. Reproduced with permission from Shainberg I, Levy GJ, Goldstein D *et al.* (2001) *Australian Journal of Soil Research* 39: 1279–1291. CSIRO Publishing, Melbourne, Australia.

also affected by sodicity (**Figure 5**). As raindrop impact energy was the same in all the experiments, the differences in IR curves for the various treatments were the result of chemical dispersion of the soil clay caused by sodicity. The high sensitivity of the soil surface to sodicity (even at low ESP values) is explained by the mechanical impact of the raindrops, which enhances chemical dispersion, and the almost total absence of electrolytes in the applied distilled water.

In cases where soil is exposed to overhead irrigation (e.g., sprinklers), IR is sensitive to both the EC and SAR of the applied water. The lower the concentration of the electrolyte solution, the faster the rate at which the IR decreases (**Figure 6**). For similar electrolyte concentrations, increasing soil ESP leads to a sharper decrease in and a lower final IR (**Figure 6**). Increasing EC greatly affected IR even at low sodicity. It was reported for Californian soils that final IR increased from 2 to 28 mm h$^{-1}$ as cation concentration in the applied water having SAR <5 increased from 5 to 28 cmmol$_c$ l$^{-1}$. These results, and others, indicate that IR is far more sensitive than HC to both SAR and electrolyte concentration of the applied water. Furthermore, when Mg is the complementary cation to Na, then IR in Mg-Na soil may often be lower than that in the Ca-Na soil, especially under conditions where chemical dispersion is dominant in controlling IR (i.e., water drops with low to medium kinetic energy).

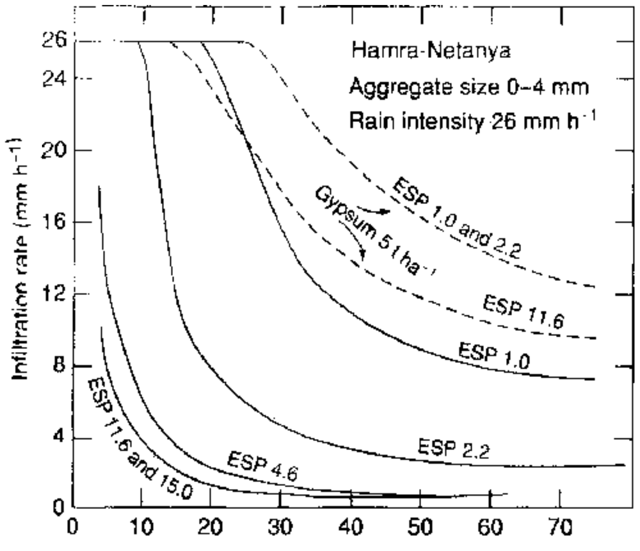


**Figure 5** The effects of the soil exchangeable sodium percentage (ESP) and phosphogypsum application on the infiltration rate of the Hamra (Netanya) soil as a function of the cumulative rain. Reproduced with permission from Kazman Z, Shainberg I, and Gal M (1983). Effect of low levels of exchangeable applied phosphogypsum on the infiltration rate of various soils. *Soil Science* 35: 184–192.

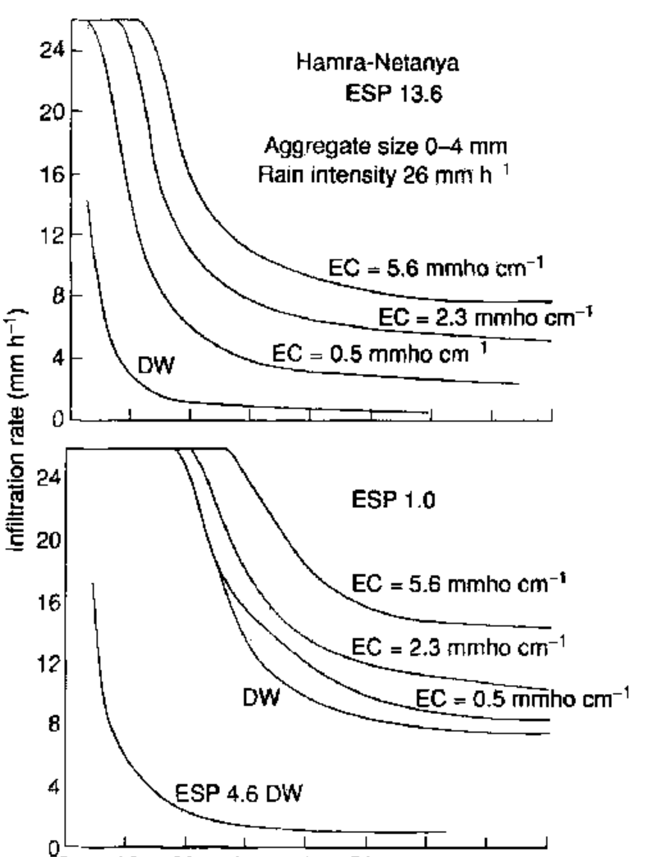


**Figure 6** Effect of electrolyte concentration in rain simulation experiments on the infiltration rate of a loess soil. ESP, exchangeable sodium percentage; DW, distilled water; EC, electrical conductivity. Reproduced from Agassi M, Shainberg I, and Morin J (1981) *Soil Science Society of America Journal* 45: 848–851.

Similarly to HC, it has recently been noted that the IR of sodic soils depends not only on soil properties but also on the conditions prevailing in the soil (e.g., antecedent moisture content and wetting rate) because IR is greatly affected by aggregate stability/breakdown. The dependence of IR of sodic soils on the rate of aggregate wetting is closely associated with soil texture. In coarse- and medium-textured soils, sodicity effects on IR are only slightly altered by aggregate wetting rate. In fine-textured soils, decreasing the rate of aggregate wetting prior to exposing them to rain significantly decreases the sensitivity of the IR to sodic conditions (Figure 7).

## Sodicity and Soil Productivity

Sodic conditions are expected to affect soil productivity adversely because such conditions can lead to poor aeration, restricted root development, and enhanced root disease. Moreover, sodic conditions can lead to nutritional imbalances and micronutrient deficiencies since they could alter soil redox potential, pH, and dissolved organic carbon concentration.

Nitrogen, phosphorus, and potassium are the most important nutritional elements. Phosphorus and potassium do not pose a nutritional problem in sodic soils. Conversely, sodic soils may suffer from inadequate nitrogen availability. The reasons for this may stem from low rates of nitrogen mineralization, denitrification under restricted aeration conditions, volatilization of ammonia under alkaline conditions, and inhibition of nitrate ($NO_3$) uptake by chloride (Cl) or sulfate ($SO_4$) ions which are usually in abundance in sodic soils. Sodic soils may also suffer from impaired uptake of Ca. The ratio of Ca to total cation concentration (TCC) is considered a better indicator of Ca availability to plants than actual Ca concentration. Saturated paste extracts of sodic soils often exhibit Ca/TCC $< 0.1$, which for many crops indicates Ca deficiency. In addition, high Mg/Ca ratios, often found in sodic soils, can also cause Ca deficiency.

Two approaches exist with respect to studying crop tolerance to sodicity, i.e., in the presence or absence of sodicity-related adverse soil physical conditions. Studies on crop tolerance to sodicity based merely on nutritional factors have indicated that, already at relatively low ESP levels (2–10), some crops show signs of sodium toxicity; however, most field crops are classified as moderately tolerant or tolerant to sodicity (Table 2). On the other hand, studies that have been carried out under conditions where soil structure had been allowed to be affected by sodicity indicated that grasses were tolerant, wheat and barley were moderately tolerant, and pea and beans were

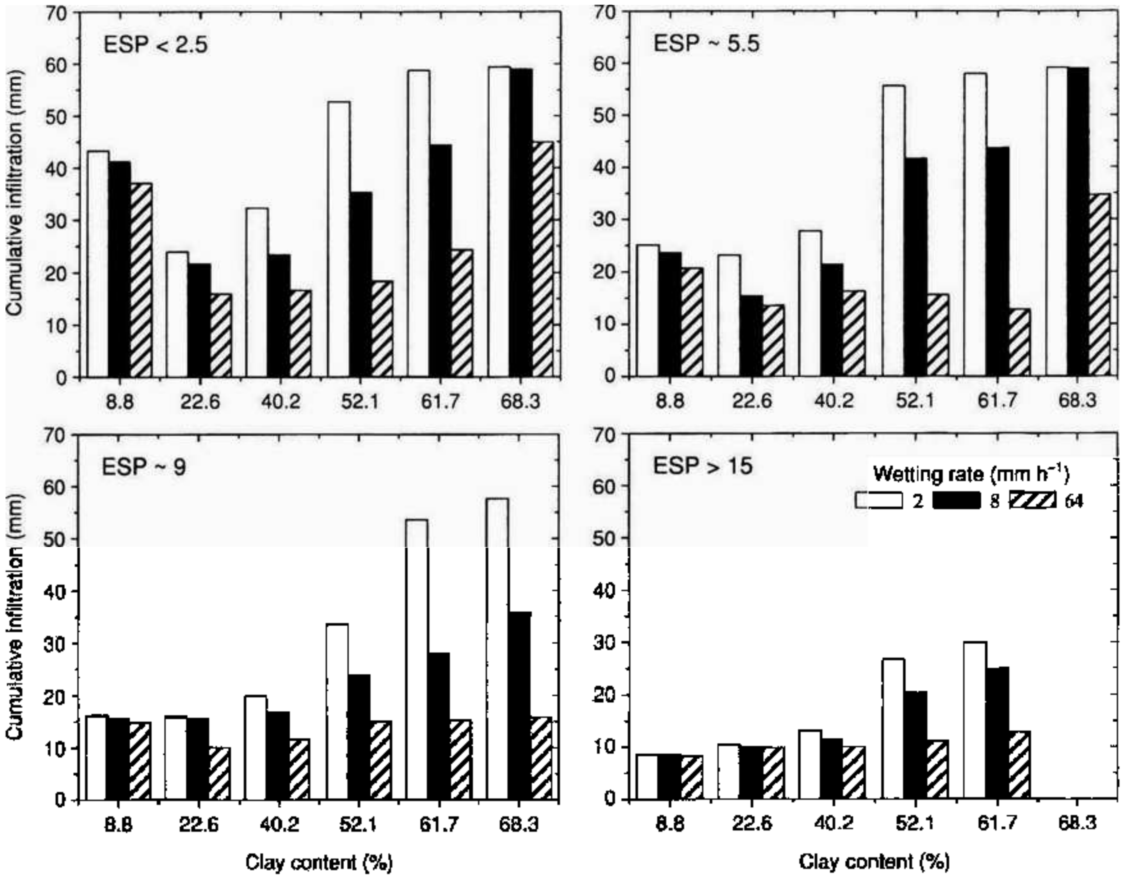


**Figure 7** Effects of wetting rate on the cumulative infiltration from a 60-mm rain storm for six soils varying in clay content and for four exchangeable sodium percentage (ESP) levels.

**Table 2** Tolerance of various crops to exchangeable sodium

| *Tolerance to ESP* | *Crop* |
|---|---|
| Extremely sensitive (ESP 2–10) | Deciduous fruits |
| | Nuts |
| | Citrus |
| | Avocado |
| Sensitive (ESP 10–20) | Beans |
| Moderately tolerant (ESP 20–40) | Clover |
| | Oats tall fescue |
| | Rice |
| | Dallis grass |
| Tolerant (ESP 40–60) | Wheat |
| | Cotton |
| | Alfalfa |
| | Barley |
| | Tomatoes |
| | Beets |
| Most tolerant (ESP >60) | Crested and fairway wheat grass |
| | Tall wheat grass |
| | Rhodes grass |

ESP, exchangeable sodium percentage. Reproduced from Pearson GA (1960) Tolerance of crops to exchangeable sodium. *US Department of Agriculture Information Bulletin* 216.

sensitive to sodicity. It should be borne in mind that the results of such studies could be site-specific, and may thus hinder their applicability to other soil types.

## Reclamation of Sodic Soils

Reclamation of sodic soils, which is the mending of soil structure, is essential if productivity is to be maintained. The main concept for reclamation of sodic soils advocates the addition of a source of Ca for replacing exchangeable Na coupled with excess water to leach the Na from the root zone deep into the soil profile. Less common methods include the addition of a variety of organic ameliorants to maintain high levels of organic matter in the soil and thus stabilize its structure.

### Chemical Reclamation

Reclaiming sodic soils via the addition of divalent cations can be achieved by using chemical amendments that release Ca ions to the soil, or by applying successive dilutions of high-salt water containing divalent cations.

**Chemical amendments** Typical amendments are those containing a source of soluble Ca or that dissolve Ca on reaction in the soil. Common amendments include gypsum, lime, $CaCl_2$, sulfuric acid, and sulfur.

Gypsum ($CaSO_4$), from both geological and nongeological sources (industries that produce $CaSO_4$ by-products), is the most commonly used amendment for sodic soil reclamation, primarily because of its low cost, reasonable solubility, and availability. Gypsum added to a sodic soil can increase its permeability by means of both EC and cation-exchange effects. The electrolyte effect is important during application of water with extremely low EC.

The cation exchange effect is important in soils where treating high ESP of the soil profile is important. The amount of gypsum required depends on the amount of exchangeable Na in a selected depth of soil. The amount of exchangeable Na to be replaced per unit land area ($Q_{na}$, $mol_c\,ha^{-1}$) during reclamation depends on the initial and desired final exchangeable Na fraction, the cation exchange capacity of the soil, its bulk density, and the depth of soil to be reclaimed. Thus, the amount of gypsum (in $Mg\,ha^{-1}$) needed to reclaim the soil, termed 'gypsum requirement' (GR), is calculated as follows:

$$GR = (8.61 \times 10^{-5})Q_{na} \qquad [4]$$

The efficiency and rate of exchange, namely the percentage of Ca that exchanges for adsorbed Na, vary with ESP, being much greater at initially high ESP values. A number of GR models exist, and provide a powerful tool for quantitative predictions of water and gypsum required to reclaim a soil to a predetermined level of sodicity.

The use of $CaCO_3$ to reclaim sodic soils is considered to be of limited value, mainly because of its low solubility and low dissolution rate. Naturally occurring $CaCO_3$ may help soil reclamation by maintaining EC in the soil solution at a level high enough to prevent clay dispersion and a subsequent reduction in soil HC. However, because of its low solubility, $CaCO_3$ is not effective in maintaining high permeability at the soil surface during rain.

**Successive dilutions of high-salt water** In areas where water is not a limiting factor, reclamation of sodic soils can be obtained by leaching the soil with successive dilutions of high-salt water containing divalent cations. In the early stage of the leaching, the high EC in the applied water prevents clay dispersion and induces flocculation of the soil colloids. Simultaneously, the Ca ions in the water decrease sodicity by replacing exchangeable Na. Upon dilution of the saline water with high-quality water, the SAR of the water is reduced. To ensure successful soil reclamation, the depth of water added should be at least 9–10 times the depth of soil to be reclaimed.

## Management of Sodic Soils

In order to maintain sustainable agriculture in sodic soils, and keep their productivity, proper management practices must be applied, including the use of soil amendments, tillage, and cropping.

### Soil Amendments

Common soil amendments used to maintain high permeability in sodic soils include gypsum that provides electrolytes to prevent clay dispersion and organic polymers that stabilize soil aggregates. Phosphogypsum (PG) and flue gas desulfurization gypsum (FGDG), which are by-products of the phosphate and power generation industries, are readily available and therefore frequently used. Spreading powdered PG at rates of $5\,Mg\,ha^{-1}$ on the soil surface can be effective in slowing down the rate of IR decline and maintaining a higher final IR compared with the control, even in soils of ESP $< 2.5$ (**Figure 5**), indicating that some chemical dispersion takes place even at very low ESP values, and possibly even in Ca-saturated soils. The use of PG is also very effective in maintaining high IR in soils with high ESP (**Figure 5**). The favorable effect of PG on IR should be attributed not only to its effect on the EC of the percolating water but also to: (1) the physical interference with the continuity of the seal; and (2) the partial mulching of the soil surface which protects the soil from the beating action of the raindrops.

Organic polymers used for improving aggregate stability, maintaining high IR, and reducing seal formation are mainly polysaccharides (PSD) and polyacrylamides (PAM). The addition of small amounts of these polymers ($10$–$20\,kg\,ha^{-1}$), either sprayed directly on to the soil surface or added to the applied water, was noted to be effective in stabilizing and cementing aggregates together at the soil surface, and hence maintaining high IR values in soils with ESP $< 5$. On the other hand, when added to soils with high ESP ($>20$), polymers were ineffective in controlling seal formation and maintaining high IR values; in soils with moderate ESP (10–15) the effects of polymers were inconsistent and depended on water quality and the amount of polymer added. The efficacy of anionic polymers in preventing seal formation is enhanced when the soil clay is maintained in a flocculated state and the resulting final IR could be 10 times higher than the control. It is therefore recommended that in soils exposed to rain, polymer application should be accompanied by spreading a source of electrolytes (e.g., gypsum) at the soil surface. Of the polymers currently available and under study, anionic polyacrylamide with high molecular weight ($>10^6$ Da) has been found to be the most effective in controlling seal formation and maintaining high IR.

### Tillage

Tilling sodic soils may result in the improvement of soil physical and hydraulic properties, especially when accompanied by some reclamation measures. Tillage reduces bulk density and increases total porosity (especially macroporosity), and improves internal drainage and saturated and unsaturated HC. Deep tillage (1–2 m) or ripping breaks up hardpans, and cemented layers mix soil layers and generally cause a permanent improvement in soil structure and physical properties. However, deep tillage can sometimes lead to waterlogging and infiltration problems by bringing sodic subsoil to the surface. An alternative to deep tillage in the case of sodic subsoil is the use of ridges instead of bed farming techniques. With ridges, the plants are located about 0.25 m above the neighboring furrows and thus aeration is improved, especially in heavy-textured soils suffering from water logging in the subsoil. Shallow tillage (disking or chiseling) is effective in reducing soil permeability problems associated with surface crusts, hardsetting, and compaction. Shallow tillage requires repetition from time to time.

## Sodicity-Related Environmental Hazards

Sodic soils are more susceptible to clay dispersion and surface sealing, which generate runoff flow and subsequently lead to elevated levels of soil erosion, a serious problem from the crop production point of view. In addition, off-site damages of runoff and erosion from cropped fields are severe; in the USA alone the cost of these damages is estimated to be in excess of $1 billion. Sediments can clog drainage ditches and irrigation canals and decrease the capacity of water reservoirs, thus increasing their maintenance costs. In addition, sediments can damage or entirely destroy fish-spawning areas. Sedimentation of rivers and streams reduces their capacity to carry water and increases the danger of flooding and consequent damage to property and to human health.

The presence of suspended sediments in water increases its turbidity (a measure of water clarity). Under certain conditions, common in sodic soils (e.g., high pH and Na levels), organic matter can dissolve in runoff water and enhance clay dispersion and mobilization, and hence increase water turbidity. Turbid water is less attractive for recreational activities and thus this reduces the value of the water. Turbidity also impedes the passage of sunlight through the water, thus negatively affecting biological productivity of the water. High turbidity in water bodies used for domestic consumption requires cleaning the water, which in turn causes an increase in their price.

An additional environmental hazard stemming from sodic soils is the transport of various contaminants by soil colloidal material, whether to groundwater or to surface water bodies via runoff. Metals which are usually fairly immobile in soil have a strong affinity to soil colloids and thus have been shown to move together with the mobile colloids. Similarly, insoluble pesticides and herbicides, and some nutrients (e.g., phosphate), have a high affinity for soil colloids and can thus be transported with the suspended solids to adjacent water bodies and groundwater. The increased concentration of suspended colloids in runoff water due to sodic conditions greatly increases the danger of higher than permitted loadings of these pollutants in groundwater and surface water bodies.

## Conclusions

Many soils worldwide contain sodium at levels that may negatively affect their behavior, and should thus be considered sodic soils. The area occupied by sodium-affected soils is expected to grow due to increased use of poor-quality soils and water for agricultural production. The adverse effects of sodic conditions on soil physical and hydraulic properties are aggravated when the soil is exposed to good (i.e., low EC) water and aggregates are slaked. Productivity of sodic soils is likely to decline, not only because of physical soil degradation but also because of nutritional imbalances, especially with respect to nitrogen and Ca. Sodicity enhances runoff and soil erosion, which lead to both on-site (cultivated fields) and off-site (environmental) damages. However, sodium-related soil degradation and loss of productivity are, to a certain degree, reversible as sodic soils can be reclaimed by the addition of amendments or by excessive leaching. Realization of sodicity hazards and consequent adoption of proper management tools allow for a safe and sustainable use of sodium-affected soils, without the danger of them being abandoned.

*See also:* **Aggregation**: Physical Aspects; **Calcium and Magnesium in Soils**; **Cation Exchange**; **Conditioners**; **Crusts**: Structural; **Flocculation and Dispersion**; **Liming**; **pH**; **Redox Potential**; **Salination Processes**; **Structure**

## Further Reading

*Australian Journal of Soil Research* (1993) Special issue, vol. 31, No. 6.

*Australian Journal of Soil Research* (2001) Special issue, vol. 39, No. 6.

Bresler E, McNeal BL, and Carter DL (1982) *Saline and Sodic Soils. Principles – Dynamics – Modeling*. Berlin: Springer-Verlag.

Keren R (1995) Reclamation of sodic-affected soils. In: Agassi M (ed.) *Soil Erosion Conservation and Rehabilitation*, pp. 353–374. New York: Marcel Dekker.

Levy GJ (2000) Sodicity. In: Sumner E (ed.) *Handbook of Soil Science*, pp. G27–G63. Boca Raton, FL: CRC Press.

Levy GJ, Mamedov AI, and Goldstein D (2003) Sodicity and water quality effects on slaking of aggregates from semi-arid soils. *Soil Science* 168: 552–562.

Shainberg I and Letey J (1984) Response of soils to saline and sodic conditions. *Hilgardia* 52: 1–57.

Sumner ME and Naidu R (1998) *Sodic Soils*. New York: Oxford University Press.

US Salinity Laboratory Staff (1954) *Diagnosis and Improvement of Saline and Alkali Soils*. Agricultural Handbook 60. Washington, DC: USDA.

---

# SOIL–PLANT–ATMOSPHERE CONTINUUM

**J M Norman and M C Anderson**, University of Wisconsin, Madison, WI, USA

## Introduction

The dominant distinguishing feature of the Earth, in comparison with all other celestial bodies we observe, is the presence of life. Humans, as a part of life on Earth, attempt to understand all of the life system, albeit from an anthropocentric perspective. James Lovelock proposed a perspective that was unsettling to some and appealing to others when he suggested that the entire Earth may be functioning as a living organism. As our integrative understanding of living systems within the soil–plant–atmosphere continuum (SPAC) grows, the idea that a whole ecosystem can function as a living entity becomes more acceptable.

Consider a terrestrial ecosystem on land: The most essential ingredient for life on Earth is water and most of the water is not on the land. However, a rapid, long-distance system for transporting water from oceans to land is available through the atmosphere, much like the xylem system in a tree, which transports water long distances (relative to cell sizes) from the soil to the leaves high above the ground. This long-distance atmospheric transport system is driven unceasingly by the uneven, solar heating of the Earth. On the land, where many organisms reside, water moves much more slowly through the soil than through the atmosphere and this provides an essential storage reservoir that maintains the critical continuous flow of water to plants in the presence of fluctuations in the atmospheric supply. The rapid atmospheric transport system that is so critical to moving water from the ocean to the land also fluctuates wildly over the life span of most organisms. The combination of a low-storage, highly fluctuating system for rapid global transport in the atmosphere and a large-storage, stable, low-flow buffer in the soil redistributes water over the Earth to sustain life on land. This is a highly structured system like an organism. Furthermore, the living systems exert a strong influence on this larger system to enhance their functioning, just as we humans modify our environment to improve our lot.

Water is the most obvious component of the SPAC, and the interaction of this water with energy is the cornerstone of our understanding of how life interacts with its environment. Every schoolchild learns the water cycle: evaporation of liquid water from surfaces supplies water vapor to the atmosphere, where air currents lift this water to heights that cool it and cause condensation of the water vapor back to the liquid form of droplets, which fall back to the Earth to be recycled. The extraction of energy from the surface as a result of evaporation and the release of energy to the atmosphere by condensation represent the primary means by which the energy from the sun ultimately warms the atmosphere. This 'latent heat,' which is associated with the change of phase of water between liquid and vapor states, and not a temperature change, provides the key feedback between water and energy cycles that stabilizes the soil–plant–atmosphere system, permitting it to sustain life.

Materials such as water flow through the ecosystem seamlessly, moving from place to place, serving innumerable functions, and consuming or releasing energy to achieve a phase suitable to the medium that contains it. This is the SPAC; 'continuum' here does not refer to the material storage forms so apparent to our eyes such as plant leaves or roots, lakes, raindrops, the soil itself, or even the air we breathe; rather it refers to the unimpeded flow of many forms of material and energy throughout an elegant system for sustaining life on land. Animals are as much a part of this system as plants, but in this article plants are our emphasis. This ceaseless flow of material and energy sustains life at every level; from the molecular interactions that sustain and mutate the genetic 'memory,' to cellular processes that make up whole organisms, to the global level of a living, breathing planet.

Living systems cycle matter through the SPAC and themselves to create and maintain their structure and extract energy for powering their life functions. All material necessary for sustaining life is cycled on all spatial levels from subcellular to global. The connectedness and functioning of the SPAC can be illustrated by reference to water, the most critical mass constituent to life on Earth. However, understanding the SPAC through water will necessarily involve some considerations of at least carbon and energy.

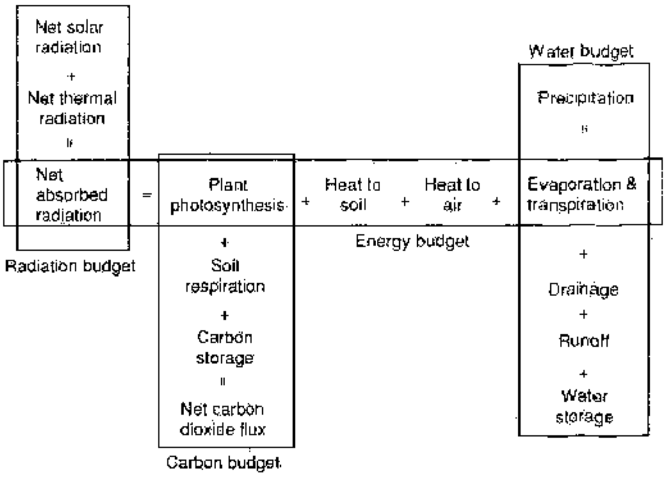


**Figure 1** The interlocking energy and mass budgets for water, carbon, and energy in the soil–plant–atmosphere continuum. Altering individual components can cascade through the system and affect many seemingly unrelated factors.

## Fundamental Principles

Two general principles are indispensable for studying the SPAC: (1) the conservation principles that take the form of mass and energy 'budgets,' and (2) the transport principles that relate the flow of some quantity to the difference or gradient of other quantities that influence or 'force' the flow and describe the 'state' of the exchange process.

The principles of conservation of mass and energy are the backbone of integrative studies of the SPAC. Because mass and energy can take many forms as they move throughout the SPAC, and even interact with each other, budgets are constructed to quantify the important stores and flows of important life-enabling constituents such as water, carbon, or energy. A budget is simply the application of the conservation principles (mass or energy) to a specific system that must be carefully defined. This system may be defined as a leaf, a community of plants, or even the entire Earth. Like balancing a bank account, a budget is a formal statement of the following: Incoming quantity minus outgoing quantity equals the increase in storage of the quantity in the system. In the case of water ($W$), carbon ($C$), and energy ($E$), the budgets become the following (Figure 1):

$$W_{\mathrm{IN}} - W_{\mathrm{OUT}} = \Delta W$$

$$C_{\mathrm{IN}} - C_{\mathrm{OUT}} = \Delta C$$

$$E_{\mathrm{IN}} - E_{\mathrm{OUT}} = \Delta E$$

The intertwined budgets of water, carbon and energy inflow, outflow and storage within the soil–plant–atmosphere system are shown in Figure 1. Understanding the SPAC through the role of water will involve simultaneous consideration of carbon and energy budgets on a spatial scale appropriate for living systems.

Consider the energy budget: Energy can exist in many forms as it flows through the soil–plant–atmosphere system, and the energy budget represents a way to keep track of the energy, no matter what its form. The rate of energy flow (joules per second = watts) per unit of area (meters squared) is a convenient way to describe the energy budget for a particular land area, because the size of the area does not matter and the emphasis can be on the energy exchanges. The radiant energy from the sun and surroundings can be absorbed by a leaf (this is referred to as the net radiation) and be transformed to the following: (1) sensible heat energy that is convected away from the leaf by moving air that is cooler than the leaf; (2) latent energy in molecules of water that are converted from liquid to vapor inside the leaf as they are transpired (evaporated) from the leaf; and (3) biochemical energy in the form of organic compounds such as sugars and starches created through photosynthesis and other physiological processes to sustain life. These components of the leaf energy budget are shown in Figure 1.

The transport principle simply states that the flow of some quantity such as water is equal to an appropriate 'driving force' divided by a resistance to transport exerted by the various media through which the quantity moves. The transport principle is more difficult to use than the conservation principle because the 'driving forces' and transport resistances depend on the characteristics of the system and mechanism of transport. For example, the diffusion of liquid water depends on the water potential difference across a hydraulic transport resistance, whereas the transport of water vapor depends on the water vapor pressure difference across a diffusive transport resistance. In addition, movement of liquids or vapor by mass flow is 'driven' by pressure differences across transport resistances. In general, diffusion is an effective mode of transport over small distances, but mass flow can move materials and energy over longer distances. Heat transport is driven by temperature differences. The relevant thermal resistance depends on whether the heat is moving through a solid such as soil (conduction, which is analogous to diffusion), or through a moving fluid such as air (convection, which is analogous to mass flow and is a much more rapid transport mechanism than diffusion). Heat transport by radiation depends on still-different formulations. Thus much research on the SPAC relates to determining appropriate transport resistances and state variables for characterizing the flow of important quantities throughout the system. A complete understanding of this soil–plant–atmosphere system is a formidable task, but such an understanding will improve the ability of humans to make choices that will sustain a diverse and resilient ecosystem, upon which our survival depends. The fundamental concepts of energy and mass conservation and transport relations constitute the theoretical backbone of environmental biophysics (the study of the SPAC).

## Water in the Soil–Plant–Atmosphere-Continuum

The smallest spatial scale that is usually associated with the SPAC is a field plot (square meters to thousands of square meters) from the bottom of a root zone (a few meters deep) to a few tens of meters above the top of the vegetation. Although individual studies may use smaller systems such as individual leaves or potted plants to reveal mechanisms, the scale of at least a small field is required to encompass most of the essential components of the natural SPAC and yet be accessible to direct measurement. Researchers generally use formulations nearly identical to those measured in field plots to represent idealized patches of the Earth's surface thousands of square kilometers in size; then they combine these patches to understand better the human influence on global climate processes. Typical time scales for small-scale studies vary from minutes to months or even years. On these spatial and temporal scales, researchers attempt to measure all the important flow and state variables, which are used to validate models that can be applied in systems that cannot be measured directly. After all, society cannot afford to measure everything everywhere, but will support measuring some things in many places and many things in a few places.

The transport and storage of water in the SPAC are shown in **Figure 2**. The flow of liquid water from the soil into the plant depends on the plant having a lower water potential than the soil; thus the soil water potential determines the availability of soil water for diffusion into the plant. The water potential is the amount of energy (joules) that can be obtained from moving a mass of water (kilograms) to a pool of pure water at atmospheric pressure and specified elevation; and in the soil and plant it is usually negative, indicating that energy must be expended to remove the water. As soils dry, the water potential decreases, making it more difficult for plants to obtain the water they need to grow. The soil water potential is related to the water content of the soil through the 'moisture-release curve,' which varies from soil to soil. Moisture-release curves have been measured for many different soil types. Typically soils can provide an amount of water to plants from storage that is equivalent to 5% (sands) to 20% (silt loams) of the plants' rooting depth. Thus for a plant rooted to a depth of 1 m, soils can provide approximately 50–200 mm of water to the plant. With evapotranspiration rates of 2–10 mm day$^{-1}$, soils can store enough water to sustain plants for a month or more without rainfall or irrigation.

Once water enters the plant through its roots, it moves through the plant vascular system to the leaves if the leaves have lower water potentials than the roots. Since the transport resistance to mass flow in the stem xylem is usually less than the transport resistance to diffusion through the root, and the flows through both root and xylem must be equal according to conservation principles, the water potential difference from the soil to the root xylem will usually be greater than the water potential difference from the root xylem to the leaf.

Water is lost from plants through pores in the leaf surfaces, called stomata. Evaporation from cells just below the stomatal pores draws water from the leaf cells, which in turn are hydraulically connected to a continuous water column down to the roots. As water is transpired through the stomata, additional water is drawn through the xylem to replace it, very much like suction applied to a straw in a soda. Of course, this evaporation that occurs just below the leaf surface consumes most of the Sun's energy that is absorbed by the leaf, because water vapor molecules carry away latent energy that was not contained in the liquid water that came up from the roots. While the flow of liquid water from the soil to the leaf is proportional to water potential differences, the flow of water

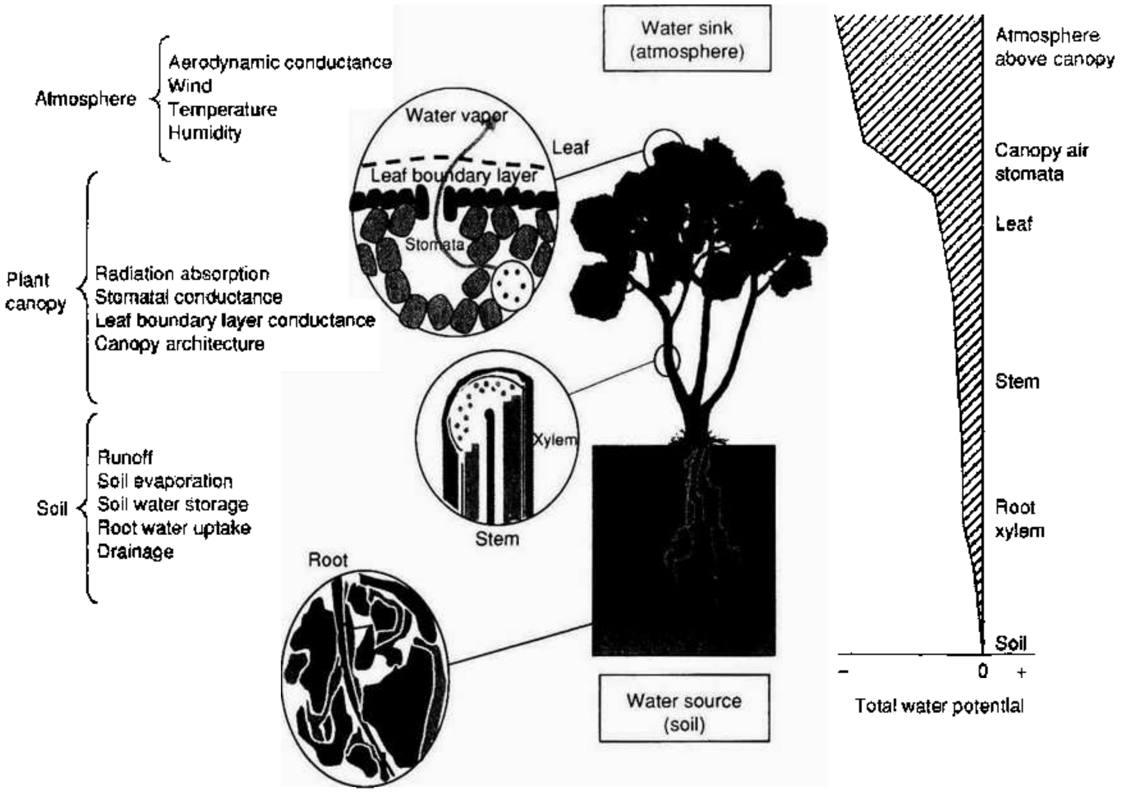


**Figure 2** Water movement and storage in the soil–plant–atmosphere continuum. The graph on the right indicates the decrease in water potential that occurs as water moves from the soil to the atmosphere. On the left the three components of the soil–plant–atmosphere system are listed with the major factors for each that should be considered in a model of intermediate complexity.

vapor from inside the leaf to the leaf surface is proportional to water vapor pressure differences. At the interface, a near-equilibrium exists between the liquid water and the water vapor in contact with it. Although different processes are used to describe water movement in different parts of the system, the flow of water is continuous across these discontinuities of processes.

A very small amount of the water that passes from the soil to the atmosphere through the plant remains in the plant as storage as the plant grows (less than 5%). An even smaller amount of the hydrogen and oxygen in the water taken up from the soil remains in the plant in the form of organic molecules synthesized by the plant.

Water that exits the stomata must pass through a thin, still-air layer adjacent to the leaf, called the leaf boundary layer, and be transported through the canopy space by turbulent mixing along the path of decreasing water vapor pressure. Ultimately the water vapor exits the plant canopy, passes through the planetary boundary layer in the lower few thousand meters of the atmosphere, and is lifted high into the atmosphere, where it condenses into clouds and eventually falls back to the Earth as precipitation to repeat the cycle.

## Small-Scale Observations and Model Building

In plant-environment studies, research derives and validates the transport principles on small, manageable systems in the laboratory or in field plots where direct measurements are possible. These principles are then applied at the ecosystem and global levels, where direct measurements are not possible, to obtain estimates of the important flow or state variables to quantify human impacts.

Consider the measurements that are necessary to understand the cycling of water in the SPAC. Precipitation falls on a field and either runs off the soil surface to streams or infiltrates into the soil itself (studied in the discipline of hydrology). The infiltrated water may be stored in the soil matrix (soil physics), evaporated from the soil surface (micrometeorology), transpired from plants (plant physiology), or drain below the plant root zone. Water that drains below the plant roots becomes part of the groundwater system (hydrogeology) and may reenter rivers (hydrology). Plants are sustained by the water stored in the soil, and the quantity of this storage depends on the type of soil. Nearly all the water taken up by plants is transpired to the atmosphere at a rate that depends on solar radiation absorbed, weather variables, and the amount of vegetation. The amount of vegetation is estimated from the fraction of ground area covered by plants (vegetative cover). By keeping track of the daily water budget (**Figure 1** and **Table 1**), the amount of water stored in the soil can be estimated, and in turn the influence of this stored water on direct evaporation from the soil and plant transpiration (together referred to as evapotranspiration) can be quantified. The maximum amount of water available to plants depends not only on soil properties, but also on plant rooting depth. Plant physiologists have measured rooting depths for the major classes of vegetation so that, given vegetation type and soil maps, the amount of water available to the plants for growth and transpiration can be estimated.

**Table 1** contains an example of a 40-day-long water budget for a corn crop that begins about a month after emergence, during the period when the crop is growing rapidly. The incoming solar radiation column represents the maximum amount of evapotranspiration that could occur if all the solar radiation were used to evaporate water; of course some

**Table 1** Simulation of corn growth and water use from 31 to 70 days after emergence, during a rapid growth phase. The solar radiation is in units of millimeters per day of evaporation that could occur if all solar radiation were to evaporate water. Vegetative cover increases as the crop grows and crop biomass increases. Soil evaporation tends to decrease as the soil surface dries out after rainfall and as vegetative cover increases; evapotranspiration (transpiration plus soil evaporation) exceeds rainfall so soil water storage is gradually depleted

| *Days after emergence* | *Rainfall (mm)* | *Radiation (mm day $^{-1}$)* | *Vegetative-cover fraction* | *Biomass (kg m $^{-2}$)* | *Transpiration (mm day $^{-1}$)* | *Soil evaporation (mm day $^{-1}$)* | *Water stored in soil (mm)* |
|---|---|---|---|---|---|---|---|
| 31 | 0 | 11.02 | 0.17 | 0.047 | 1.16 | 2.91 | 177 |
| 32 | 0 | 10.10 | 0.21 | 0.061 | 1.26 | 2.53 | 173 |
| 33 | 0 | 6.76 | 0.24 | 0.073 | 0.97 | 1.13 | 171 |
| 34 | 0 | 11.70 | 0.26 | 0.083 | 1.83 | 1.53 | 168 |
| 35 | 0 | 6.93 | 0.31 | 0.101 | 1.23 | 0.72 | 166 |
| 36 | 0 | 11.17 | 0.34 | 0.113 | 2.13 | 0.99 | 163 |
| 37 | 0 | 7.62 | 0.38 | 0.135 | 1.61 | 0.57 | 161 |
| 38 | 0 | 6.04 | 0.41 | 0.151 | 1.35 | 0.39 | 159 |
| 39 | 0 | 11.63 | 0.43 | 0.164 | 2.72 | 0.67 | 155 |
| 40 | 0 | 10.94 | 0.47 | 0.191 | 2.75 | 0.53 | 152 |
| 41 | 20 | 5.60 | 0.51 | 0.219 | 1.49 | 0.74 | 170 |
| 42 | 10 | 6.53 | 0.53 | 0.234 | 1.78 | 0.82 | 177 |
| 43 | 0 | 11.41 | 0.55 | 0.252 | 3.20 | 1.45 | 173 |
| 44 | 11 | 5.82 | 0.59 | 0.284 | 1.70 | 0.62 | 181 |
| 45 | 0 | 9.43 | 0.60 | 0.301 | 2.81 | 0.95 | 178 |
| 46 | 0 | 11.61 | 0.63 | 0.329 | 3.55 | 1.07 | 173 |
| 47 | 0 | 10.70 | 0.66 | 0.364 | 3.36 | 0.89 | 169 |
| 48 | 0 | 11.03 | 0.69 | 0.398 | 3.54 | 0.83 | 164 |
| 49 | 0 | 11.49 | 0.71 | 0.433 | 3.76 | 0.78 | 160 |
| 50 | 0 | 11.31 | 0.73 | 0.471 | 3.76 | 0.69 | 155 |
| 51 | 0 | 11.13 | 0.75 | 0.509 | 3.75 | 0.62 | 151 |
| 52 | 0 | 10.16 | 0.77 | 0.546 | 3.45 | 0.52 | 147 |
| 53 | 0 | 9.88 | 0.78 | 0.581 | 3.39 | 0.46 | 143 |
| 54 | 0 | 12.13 | 0.79 | 0.614 | 4.19 | 0.38 | 139 |
| 55 | 0 | 10.51 | 0.81 | 0.656 | 3.65 | 0.24 | 135 |
| 56 | 0 | 10.83 | 0.82 | 0.693 | 3.79 | 0.20 | 131 |
| 57 | 0 | 12.19 | 0.83 | 0.731 | 4.28 | 0.19 | 126 |
| 58 | 0 | 12.76 | 0.84 | 0.773 | 4.50 | 0.17 | 122 |
| 59 | 0 | 9.08 | 0.85 | 0.818 | 3.21 | 0.10 | 118 |
| 60 | 20 | 6.70 | 0.85 | 0.851 | 2.38 | 0.19 | 136 |
| 61 | 0 | 11.61 | 0.86 | 0.874 | 4.13 | 0.32 | 131 |
| 62 | 0 | 11.98 | 0.86 | 0.916 | 4.27 | 0.31 | 127 |
| 63 | 0 | 11.04 | 0.87 | 0.958 | 3.94 | 0.27 | 123 |
| 64 | 0 | 11.12 | 0.87 | 0.998 | 3.97 | 0.26 | 118 |
| 65 | 0 | 11.05 | 0.87 | 1.037 | 3.95 | 0.25 | 114 |
| 66 | 0 | 12.14 | 0.88 | 1.077 | 4.35 | 0.27 | 109 |
| 67 | 0 | 10.83 | 0.88 | 1.120 | 3.89 | 0.23 | 105 |
| 68 | 0 | 10.71 | 0.88 | 1.159 | 3.84 | 0.23 | 101 |
| 69 | 0 | 9.87 | 0.88 | 1.198 | 3.54 | 0.20 | 98 |
| 70 | 0 | 8.32 | 0.88 | 1.233 | 2.99 | 0.17 | 94 |

radiation is reflected so the actual evapotranspiration is the sum of the transpiration and direct evaporation from the soil, which is approximately one-third of the maximum possible evapotranspiration. The maximum amount of water that can be stored in the rooting zone of the soil is 200 mm for this soil, and clearly the corn crop is using water faster than the rainfall is replenishing it, so that at the end of 70 days of growth approximately half of the maximum soil water storage has been used up. At this point the water depletion in the soil will begin to cause the growth rate of the corn to decrease unless there is more rainfall. This buffering capacity of soil is crucial to growing crops in continental climates all over the world.

Measurements of water transport and storage are difficult and expensive to make, so they can only be done at a few sites under a few conditions. Models that capture the fundamental conservation and transport principles are therefore essential to applying results of limited SPAC studies to a wider range of conditions than could possibly be measured directly. A wide array (hundreds) of computer models are available for quantifying the role of water in the SPAC and the range of applicability of such models depends on having a minimal, faithful representation of critical factors in each component of the SPAC; namely in the soil, the plant, and the atmosphere. SPAC models can be divided into three classes: simple, intermediate, and complex.

Simple SPAC models are the easiest to create, generally emphasize the knowledge base of one discipline with a narrowly focused purpose, and are the most limited in applicability. For example, the Simple Inverse Yield Model, for precision agriculture, and the CERES Maize model, for crop yield and water-use prediction, neglect the effect of vegetation roughness and wind on evapotranspiration, reflecting their agricultural roots and limited applicability. There is also a simple micrometeorological model of evapotranspiration that omits soil processes, reflecting its atmospheric origins and limited applicability.

Intermediate SPAC models incorporate at least a minimal consideration of key soil, plant, and atmospheric components to provide wide applicability without undue complexity. These intermediate models incorporate atmospheric factors (radiation, wind, temperature, humidity, and precipitation), vegetation factors (plant type, vegetative-cover fraction and height, rooting depth), and soil properties (soil evaporation, root-zone water transport and storage). Examples of Intermediate SPAC models include the Two-Source Model, for remote sensing applications, and the ALEX (atmosphere–land exchange) model, for local site-specific studies.

Complex SPAC models incorporate most of the processes known to be important at small (leaf to field) scales, and can be applied to a wide range of cover types and climatic conditions. Because they are very detailed, they generally involve a large array of coefficients related to specific plant and soil types. For this reason, complex models are often not well-suited to regional scale applications, where these myriad of parameters need to be defined at all locations. The greatest value of complex SPAC models lies in improving our understanding of key processes and using this knowledge to derive intermediate models with the fewer parameters and wider applicability.

An example of this use of a complex SPAC model is the derivation of the Two-Source and ALEX models from the complex Cupid model. The Cupid model predicts plant–atmosphere exchanges of carbon, water, and energy in a very detailed mechanistic way, and it requires dozens of species-specific coefficients related to plant physiology. By studying the behavior of the Cupid model, the ALEX model could be created using only a few carefully chosen physiological coefficients and yet achieve a surprising level of generality compared with Cupid. Not only does ALEX link growth and water use in a way nearly as fundamental as Cupid, but it does so with a widely used 'light-use efficiency' coefficient that is already tabulated for most vegetation types around the world. The result of this process is a robust model suitable for continental-scale applications.

Complex SPAC models have also been used to study the relation between the architecture of vegetation canopies and their remotely sensed signatures from aircraft and satellites. Since remote sensing is the only way to characterize vegetative cover on the regional and continental scales, we must understand how to interpret remotely sensed land observations to accommodate the profound influence of vegetation density on energy and material exchanges in the SPAC. For example, in **Table 1**, the increased importance of transpiration (and extraction of soil water to great depths by plant roots) over evaporation from the soil (water extraction to only a few centimeters of soil depth) is apparent as the vegetation cover increases from 17% to almost 90% during the 40 days of growth.

Determining the level of complexity appropriate for an application depends on the objective of the endeavor; typically, the simplest model that provides adequate results is the model of choice. Our understanding of energy and water flows in the SPAC is sufficiently well developed to extend this knowledge to larger scales, where direct measurements are only rarely, if ever, possible and evaluate implications of human activities on ecosystem health.

## Application to Regional and Global Scales

The impact of humans on the global ecosystem has become so profound that extending our understanding of biophysical processes from small fields to whole regions is becoming critical for our survival. The extension of process-level understanding of energy and material flows in the SPAC to larger scales imposes a severe constraint that is not present in plot-scale studies; namely, that the allowable input quantities are limited to state variables that can be monitored over large areas. Only a few state variables are routinely measured widely in networks across the world and thus can be considered generally available for monitoring the SPAC. For example, near-surface weather variables (air temperature, humidity, wind speed and direction, precipitation, and cloud cover), solar radiation, atmospheric carbon dioxide, soil maps, vegetation-type maps, river and stream monitoring, and satellite images are routinely obtained at many locations in many countries of the world. This means that these are the main quantities that can be assumed to be generally available everywhere (not always true, but a reasonable assumption). All other inputs required for SPAC models must be computed from fundamental principles or from quantitative relations derived from validation experiments at small scales. This pursuit has been exceedingly challenging, but systematic research over decades in numerous disciplines has identified many of the critical processes, and integrative computer modeling has synthesized much of this information so that monitoring water transport and storage on continental scales may now be possible.

Consider an example of a suite of models that has been assembled to explore systematically the energy and material flows in the SPAC over a range of scales from small fields to the continental USA (**Figure 3**).

At the heart of this suite of models is a pair of Intermediate SPAC models; namely ALEX for computing in the forward direction and the Two-Source Model for inverse computation using satellite observations as inputs. When the sky is cloudless for a few hours in the morning, operational satellites can be used to quantify vegetative cover and monitor surface temperature to estimate the surface flows of

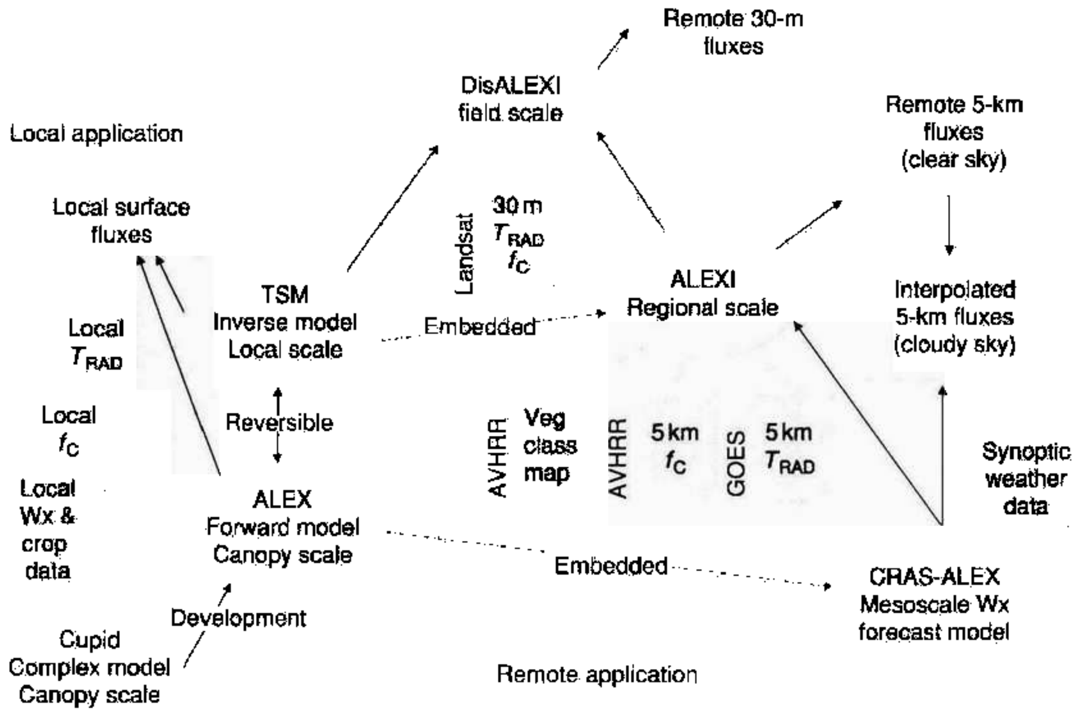


**Figure 3** Suite of models for estimating evapotranspiration on large spatial scales (remote application) based on measurements from studies on small fields (local application). Cupid, the Atmosphere–Land Exchange model (ALEX), and the Two-Source Model (TSM) are used on the field scale and require inputs that can be directly measured in field studies. ALEX Inverse (ALEXI) disaggregation ALEXI (DisALEXI), and CRAS-ALEX (Cooperative Institute for Meteorological Satellite Studies (University of Wisconsin) weather forecast model) are used on the 5- to 10-km-grid scale for continental-sized applications, and they require inputs from satellites (Landsat, AVHRR, or GOES) or national weather measurement networks (synoptic weather data). The TSM is used at the local scale, embedded in ALEXI for use at the continental scale, and used in DisALEXI for remote estimates at the large field scale; likewise, ALEX is used at the local scale and also embedded in the weather forecast model for use on the continental scale. GOES, geosynchronous operational environmental satellite; AVHRR, advanced very-high-resolution radiometer satellite; $T_{RAD}$ radiometric temperature; $f_C$, fraction of vegetative cover on the ground; Wx, weather.

energy and water at the 5- to 10-km scale and larger using the atmosphere–land exchange Inverse ALEXI model (**Figure 3**). ALEXI is a combination of the Two-Source Model and a meteorological model of the behavior of the lower 2000 m of the atmosphere (the planetary boundary layer). The ALEXI model can be used to update a mesoscale weather forecast model wherever a cloudless morning occurs. Likewise, the weather forecast model can be used to interpolate quantitites used in ALEXI whenever the satellite view is obscured by clouds. Mesoscale weather forecast models capture important feedbacks between land and atmospheric processes so that, as humans change the land characteristics, these models consider how these changes may affect subsequent weather. Meteorologists have refined these mesoscale models for decades and can now capture the transport of quantities from spatial scales of approximately 10 km and time scales of minutes to continental spatial scales over days, months, or even years of time.

The combination of the mesoscale model and ALEXI can yield time-continuous maps of the flow of water and energy in the SPAC at scales of 5–10 km and larger. How can the accuracy of such estimates be validated? Two methods have been used: (1) the use of research aircraft in intensive field experiments, and (2) the application of a disaggregation technique (DisALEXI), which uses data from a high-spatial-resolution satellite (such as Landsat, with a 30-m spatial resolution) to disaggregate 5- to 10-km ALEXI flux estimates to the spatial scale of in situ ground measurements (hundreds of meters) based on micrometeorological methods such as eddy covariance. Not only can the DisALEXI model assist with validation of the ALEXI model, but it can be used to update crop growth and water-use models in agriculture by providing occasional independent estimates of the flow of energy and water on the fine spatial scale of 30 m.

This process of using fundamental principles to infer global implications of plot-scale measurements reminds one of a statement Mark Twain once made: “There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact.” Clearly the power that fundamental principles give scientists to extend the influence of humans is difficult for nonscientists to comprehend.

## Institutional Issues

Studies of the SPAC are interdisciplinary (involving research in areas that are not claimed by existing disciplines) and multidisciplinary (requiring collaboration among scientists from numerous disciplines). For example, to understand the influence of vegetation on weather forecasts involves the following activities: meteorologists study the effect of the surface energy and water exchanges on the lower few thousand meters of the atmosphere called the planetary boundary layer; physiologists characterize the dependence of photosynthesis and stomatal processes on atmospheric and soil environmental factors; soil scientists measure the heat- and water-holding characteristics of the soil as well as carbon-exchange processes; micrometeorologists relate vegetation-canopy characteristics and atmospheric variables to the exchange of heat, mass, and momentum with the atmosphere; and hydrologists consider the runoff characteristics of the soil surface and the recharge of groundwater by percolation. The study of vegetative-canopy architecture, within-canopy processes, and the interaction between remote sensing and vegetative cover are key interdisciplinary areas that are not associated with any particular discipline.

Most scientific research is funded by government agencies and conducted within a disciplinary structure of research establishments or universities; therefore, competition among disciplines for funding and the limited perspective of individual researchers mitigate against developing an integrated view of the SPAC. Although much lip service has been paid to the value of interdisciplinary and multidisciplinary research, no sustaining structure in the research establishment is apparent for such integrative activities, and progress continues to be dominated by ad hoc efforts from committed individuals or teams with determination and vision. One strategy that has been most valuable in recent decades is the organization and execution of large, international, multidisciplinary, intensive field experiments, with broad objectives that appeal to scientists in many disciplines. The wide distribution of and open access to data from such experiments have been responsible for tremendous gains in understanding of the SPAC in recent years. Two examples of such experiments are the First ISLSCP (International Satellite Land Surface Climatology Project) Field Experiment (FIFE) in the grasslands of Kansas, USA, and the Boreal Ecosystem Atmosphere Study (BOREAS) in the Canadian boreal forest.

The time-honored reductionist perspective of ‘changing a single variable and holding all other factors constant’ is not possible in holistic studies of the profoundly interconnected SPAC. Instead the challenge of exploring important processes without the constraints of limiting paradigms and parochial disciplines must be confronted, a formidable challenge for environmental scientists. Efforts to synthesize knowledge from various components of the SPAC into a holistic view continue to be challenging in a scientific world dominated by technological reductionism. The view of the Earth as a complex, self-organized system gains wider acceptance as our knowledge of the interconnectedness among components of the SPAC improves.

*See also:* **Energy Balance; Evapotranspiration; Water Cycle; Water-Use Efficiency**

## Further Reading

Anderson MC, Norman JM, Diak GR, Kustas WP, and Mecikalski JR (1997) A two-source time-integrated model for estimating surface fluxes using thermal infrared remote sensing. *Remote Sensing Environment* 60: 195–216. [Description of the ALEXI model.]

Anderson MC, Norman JM, Meyers TP, and Diak GR (2000) An analytical model for estimating canopy transpiration and carbon assimilation fluxes based on canopy light-use efficiency. *Agricultural and Forest Meteorology* 101: 265–289. [Description of the ALEX model.]

Campbell GS and Norman JM (1998) *Introduction to Environmental Bio-physics*. New York: Springer-Verlag.

Choudhury BJ, Ahmed NU, Idso SB, Reginato RJ, and Daughtry CST (1994) Relations between evaporation coefficients and vegetation indices studied by model simulations. *Remote Sensing Environment* 50: 1–17.

Foley JA, Prentice IC, Ramankutty N, Levis S, Pollard D, Sitch S, and Haxeltine A (1996) An integrated biosphere model of land surface processes, terrestrial carbon balance, and vegetation dynamics. *Global Biogeochemical Cycles* 10: 603–628.

Gower ST, Kucharik CJ, and Norman JM (1999) Direct and indirect estimation of leaf area index, fAPAR, and net primary production of terrestrial eco-systems. *Remote Sensing Environment* 70: 29–51.

Hillel D (1998) *Environmental Soil Physics*. New York: Academic Press.

Jones CA and Kiniry JR (1986) *CERES-Maize: A Simulation Model of Maize Growth and Development*. College Station, TX: Texas A & M University Press.

Lovelock JE (1987) *Gaia: A New Look at Life on Earth*. New York: Oxford University Press.

Mecikalski JR, Diak GR, Anderson MC, and Norman JM (1999) Estimating fluxes on continental scales using remotely-sensed data in an atmospheric-land exchange model. *Journal of Applied Meteorology* 38: 1352–1369.

Morgan CS, Norman JM, and Lowery B (2003) Estimating plant-available water across a field with an inverse yield model. *Soil Science Society of America Journal* 67: 620–629.

Norman JM and Arkebauer TJ (1991) Predicting canopy light-use efficiency from leaf characteristics. In: Ritchie JT and Hanks RJ (eds) *Modeling Plant and Soil Systems*, pp. 125–143. Madison, WI: American Society of Agronomy.

Norman JM, Kustas WP, and Humes KS (1995) A two-source approach for estimating soil and vegetation energy fluxes from observations of directional radiometric surface temperature. *Agricultural and Forest Meteorology* 77: 263–293.

Sellers PJ, Hall FG, Asrar G, Strebel DE, and Murphy RE (1992) An overview of the first international satellite land surface climatology project (ISLSCP) field experiment (FIFE). *Journal of Geophysical Research* 97(18): 345–371.

Sellers PJ, Hall FG, Kelly RD *et al.* (1997) BOREAS in 1997: experiment overview, scientific results, and future directions. *Journal of Geophysical Research* 102 (D24): 28 731–28 769.

Stewart JB, Kustas WP, Humes KS, Nichols WD, Moran MS, and De Bruin AAR (1994) Sensible heat flux-radiometric surface temperature relationship for eight semiarid areas. *Journal of Applied Meteorology* 33: 1110–1117.

# SOLUTE TRANSPORT

**M C Sukop**, Florida International University, Miami, FL, USA
**E Perfect**, University of Tennessee, Knoxville, TN, USA

## Introduction

The movement of chemicals in soils and groundwater aquifers (solute transport) plays a central role in global biogeochemical cycling. Human activities have placed many new chemicals into the environment and greatly enhanced the local loading of wastes to the subsurface. Obvious applications of an understanding of solute transport in porous media include the efficacy and fate of agricultural chemicals (fertilizers and/or pesticides); the treatment and disposal of wastewaters and wastewater treatment products from human, animal, and industrial sources; and contamination resulting from landfills, mining operations, industrial activities, high-level radioactive waste disposal, etc. Given its great importance in the maintenance of safe water supplies, the nourishment of plants, regulation of the fate of soil-applied chemicals, and many other critical hydrologic processes, solute transport has long been the focus of intense research in soil science, groundwater hydrology, and chemical engineering. Nevertheless, our *a priori* capability to predict solute transport remains limited.

Pores may contain water, soil gases, and/or introduced materials such as nonaqueous-phase liquids. Chemicals can be transported in any or all of these three fluid phases. The general aspects of transport are similar in each potential phase so the emphasis here and in much of the solute transport literature is on the transport of chemicals in aqueous solution. A solute is a chemical that is dissolved in a solvent to form a solution.

Three mechanisms operate to transport solutes in soils and aquifers. 'Advection' (or synonymously 'convection' in the solute transport literature) refers to the movement of chemicals with soil fluids and groundwater. Advection often dominates the solute transport process. Diffusion is the omnidirectional spreading of a chemical that results from thermal (Brownian) motion of the fluid and solute molecules. Dispersion results primarily from variations in fluid velocity that occur at all scales. Dispersion is anisotropic and enhanced in the overall direction of flow.

Breakthrough curves display the variation in solute concentration as a function of time or distance. These measurements and the advection–dispersion equation are the primary tools used in the evaluation of solute transport.

Solutes can interact strongly with soil surfaces by physical or chemical adsorption and their transport can be appreciably slowed in a process known as retardation. Repulsion or exclusion of solutes from soil surfaces can result in transport more rapid than that of the carrier fluid.

## Transport Processes

### Advection

In the most simplistic sense, dissolved solutes are carried along with moving soil fluids. Here, 'fluid' refers to either a gas or a liquid phase; simultaneous treatment of interacting gas and liquid phases requires coupling of the treatment described below. A fluid flowing through a porous medium at a flux $q$ (volume of fluid per unit area of soil per unit time) with a solute concentration $C$ (mass of solute per unit volume of fluid) carries with it a solute flux $J_c = qC$ (mass of solute per unit area of soil per unit time). The mean velocity of the flow $v$ (and the solute, assuming no losses, production, or interaction with the solid phase) can be computed from knowledge of the flux $\mathbf{q}$ and the volumetric fluid content $\theta$ via $\mathbf{v} = \mathbf{q}/\theta$. Under steady flow, $x = vt$ (where $v$ is the magnitude of $\mathbf{v}$), and the distance $x$ traveled by an average solute molecule in time $t$, or the time $t$ needed for the solute to traverse a given distance $x$, can be calculated. The advective solute flux $J_c$ can also be written in terms of the mean velocity, the concentration, and the fluid content: $J_c = v\theta C$. In some instances, for example when the solute source is continuous and the solute concentration is uniform in the flowing fluid, advection dominates the solute transport process and may be the only process that need be considered.

### Diffusion

Diffusion is the spreading of a solute due to thermally induced Brownian motion. It is generally considered to be omnidirectional and driven by concentration gradients according to concentration-independent coefficients, though close study reveals that it is more complicated when solute concentrations are high. Under the typical assumptions, diffusion can be treated with Ficks first law, which gives the diffusive solute flux $J_s$ as a function of the concentration gradient $\nabla C$ and the bulk diffusion coefficient $D_b$, i.e., $J_s = -D_b \nabla C$.

In soils, the diffusion paths are obstructed by solid particles and solutes are constrained to follow the tortuous, connected pore network. The connected network changes as the fluid content changes. The effective diffusion coefficient $D_e$ must therefore be computed from the bulk diffusion coefficient and the soil water content and porosity. A commonly used empirical expression for the estimation of the effective diffusion coefficient is $D_e = D_b \theta^{10/3}/\phi^2$, where $\phi$ is the porosity.

### Dispersion

Consideration of saturated water flow in perhaps the simplest possible pore geometry – a straight capillary tube with circular cross section – provides insight into the processes that take place in the more tortuous and convoluted pore networks that characterize natural porous media. The velocity of fluid in such a tube depends on the pressure drop between the ends of the tube (a distance $L$ apart), the tube's radius $r$, and the dynamic viscosity $\eta$. In general, the velocity is not uniform across the tube's cross section. In the laminar flow regime, generally thought to apply to water movement in most soils, the velocity profile in a capillary tube is parabolic:

$$v(y) = 2v\left(1 - \frac{y^2}{r^2}\right) \quad [1]$$

where $v(y)$ is the velocity as a function of radial distance $y$ from the tube center.

The mean velocity $v$ can readily be determined from Stokes law in the form:

$$v = \frac{r^2 \Delta P}{8\eta L} \quad [2]$$

According to this model, the maximum velocity at the center of the tube ($y = 0$) is $2v$, and there is zero velocity at the tube's inner surface ($y = r$).

A solute introduced at concentration $C_0$ at the upstream end of the tube at time zero ($t = 0$) will be convected rapidly in the center of the tube and more slowly near the tube's surface (**Figure 1**). Taking integrated samples from various cross sections along the tube at a later time ($t > 0$) results in the collection of fluid containing no solute and fluid containing solute at the introduced concentration $C_0$ in various proportions. Dilution of the solute with solute-free fluid in the sample occurs according to the relative cross-sectional areas of the solute paraboloid and the solute-free fluid. Thus, the solute front is spread out or dispersed.

In addition to the dispersion introduced by velocity variations within individual pores, the variable sizes and path lengths of pores in a natural porous medium and mixing at pore intersections also contribute to dispersion. Buoyancy factors may cause fingering and attendant mixing of a dense solution into a less-dense one. If the dispersion arises from such velocity and path-length variations, and hydrodynamic mixing alone, it is referred to as 'hydrodynamic dispersion.' The impact of variable velocities, path lengths, and mixing accumulates through the entire hierarchy of scales pertinent to the particular porous medium under consideration. Unsaturated conditions found in most soils generally add additional complexity to flow fields within individual pores and increase the tortuosity. Hence, the degree of dispersion is usually greater under partially saturated conditions than in the fully saturated case.

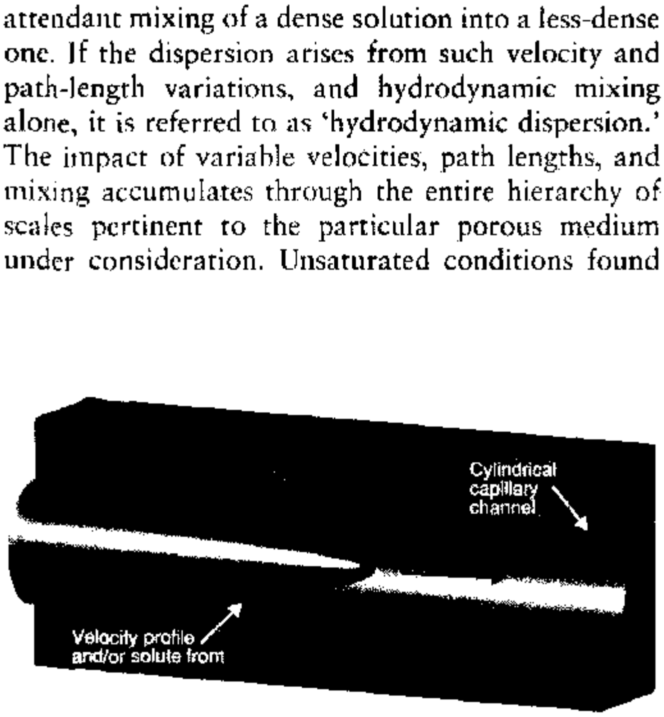


**Figure 1** Schematic velocity profile and/or solute front position in an idealized cylindrical capillary tube. Flow from left to right

Diffusion results in a spreading of solute fronts similar to that induced by dispersion. It is generally not possible to separate out the effects of diffusion and dispersion. They are typically considered to be additive with a lumped diffusion–dispersion coefficient $D_d$ given by $D_d = D_h + D_e$, where $D_h$ is the coefficient of hydrodynamic dispersion. The diffusive–dispersive solute flux $J_d$ is then treated assuming a Ficks law-like formulation: $J_d = -D_d \nabla C$.

The diffusion–dispersion coefficient has been notoriously difficult to estimate *a priori* and is generally determined via inverse modeling. Excellent computer programs for its determination from transport measurements are available. However, this reliance on measurements of the actual transport process for parameter estimation is unfortunate, because the measurements are difficult and time-consuming, and can rarely be carried out at the scale of interest.

Inside a pore, the dimensionless Peclet number ($Pe \equiv vl/D_h$, with $l$ a characteristic length) indicates the relative importance of diffusion and advection; large values of $Pe$ indicate an advection-dominated process, while small values of $Pe$ indicate the dominance of diffusion. Similarly, the dimensionless diffusion–dispersion coefficient $D^* \equiv D_d/D_h$ reflects the relative importance of hydrodynamic dispersion and diffusion. For porous media with well-defined characteristic lengths (i.e., bead diameter in packed beds of uniformly sized glass beads), $D^*$ can be estimated from $Pe$. Different classes of behavior have been proposed based on the observed relationship between $Pe$ and $D^*$ (**Figure 2**). Class I is characterized by very slow flow and the dominance of diffusion. Class II is transitional, with approximately equal and additive hydrodynamic dispersion and diffusion. Hydrodynamic dispersion dominates the dispersion process in classes III and above. In class III, the role of diffusion is still nonnegligible, but it ceases to be significant in class IV. Finally, in class V, the velocity is so high that the flow of many fluids is turbulent.

## Dimensionality

The dimensionality of the solute transport process is an important factor. In many soil science applications, for example the uniform application of a pesticide to the soil surface, the process can be treated as one-dimensional. However, in other instances two- or

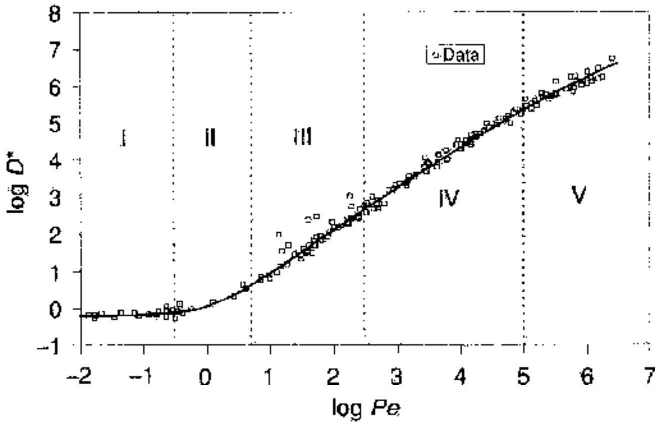


**Figure 2** Empirical relationship between dimensionless dispersion coefficient ($D^*$) and Peclet number (*Pe*) with data for uniformly sized particle beds. Roman numerals indicate behavioral class (see text). (Adapted from Fried JJ and Combarnous MA (1971) Dispersion in porous media. *Advances in Hydroscience* 7: 169–282, with permission.)

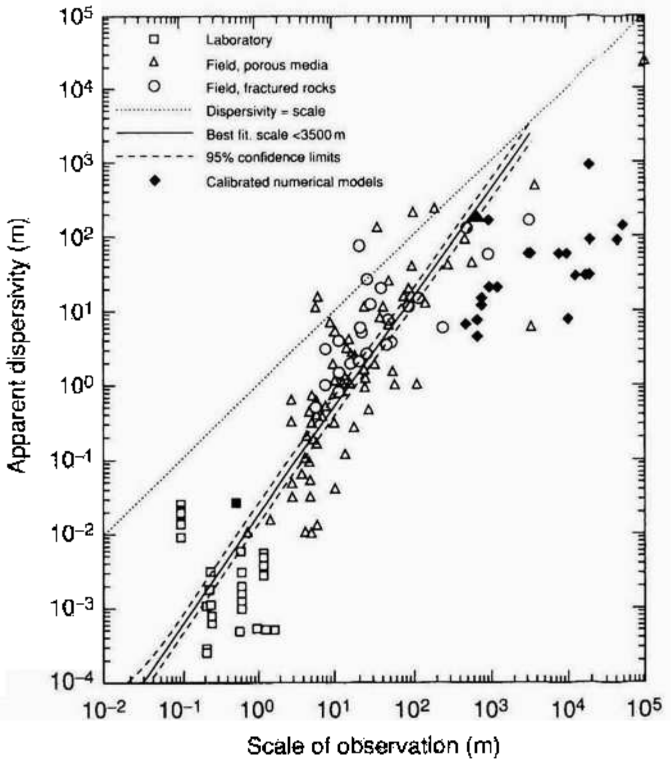


**Figure 3** Apparent dispersivity versus scale for laboratory and field measurements. (Adapted from Neuman SP (1995) On advective transport in fractal permeability and velocity fields. *Water Resources Research* 31: 1455–1460, American Geophysical Union, with permission.)

three-dimensional aspects of the transport process cannot be neglected. The problem can be viewed in terms of the geometrical configurations of the source and sink of interest. In the case of uniform soil application, the source is considered planar and the sink may be a planar water table. In contrast, a chemical spill limited to a small area overlying a thick vadose zone may need to be considered a point source in a fully three-dimensional system where solute may be dispersed both horizontally and vertically.

Even in a flow field that is macroscopically one-dimensional, dispersion is an anisotropic process and is enhanced in the direction of flow relative to transverse directions. A fully three-dimensional treatment of dispersion therefore requires a tensorial representation. For simplicity, we will restrict our discussion of solute transport to one-dimensional flow, where the effects of dispersion are embodied in a single scalar, the longitudinal dispersion coefficient, and the related longitudinal dispersivity.

## Longitudinal Dispersivity

The relationship between the one-dimensional diffusion–dispersion coefficient and mean pore-water velocity is frequently modeled with the empirical equation $D_d = m + \alpha v^n$, where $m$ and $n$ are fitting parameters, and $\alpha$ is the longitudinal dispersivity. In the hydrodynamic dispersion-dominated regime, the magnitude of $m$ is negligible and $n$ approaches unity. Under these conditions, the longitudinal dispersivity has dimensions of length and can be defined as $\alpha \equiv D_d/v$.

The dispersivity ideally would be a constant for a particular porous medium, and efforts have been made to relate it to other medium properties. In general, it has been observed that under saturated conditions the magnitude of the dispersivity depends upon the scale of the domain under consideration (**Figure 3**). The reasons for this scale dependency are unclear, but, at a macroscopic scale, the phenomenon may be related in part to spatial variations in the underlying saturated hydraulic conductivity field. New approaches for treating the relationship between $\alpha$ and scale have recently appeared and are described further below.

Another important influence on dispersivity under saturated conditions is soil structure, as influenced by differences in texture. In general, sandy soils have a narrower range of pore sizes and a smaller dispersivity than finer-grained soils. The combined effects of measurement scale and soil texture on $\alpha$ for soil columns (based on empirical relationships between dispersivity and soil hydraulic parameters and power law scaling of dispersivity with length) are shown in Figure 4.

## Breakthrough Curves

Plotting the downstream changes in solute concentration against time at a fixed position or against distance from the source at a fixed time results in a

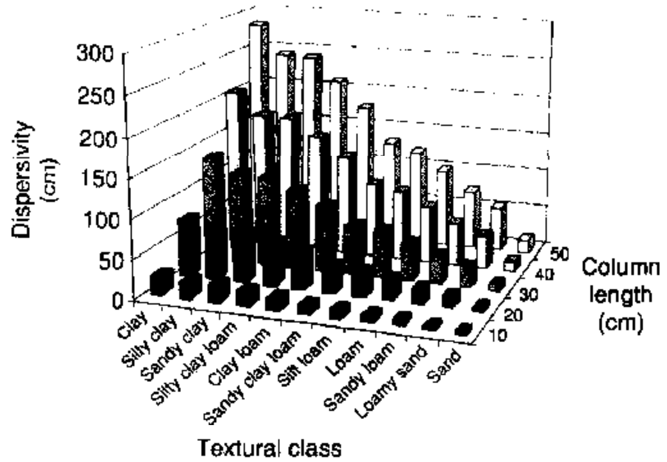


**Figure 4** Approximate effects of measurement scale and soil texture on dispersivity for soil columns based on empirical relationships between dispersivity and soil hydraulic parameters and power-law scaling of dispersivity with length.

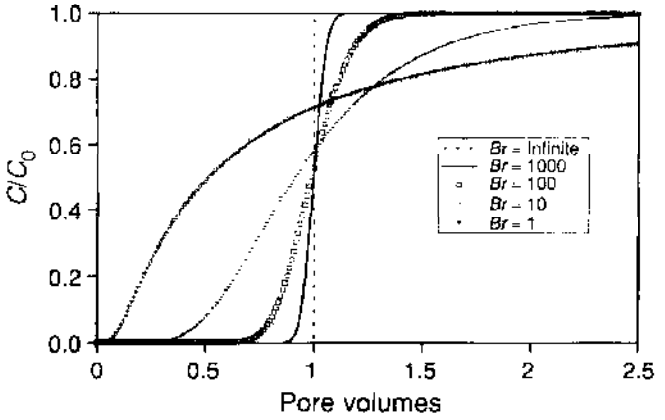


**Figure 5** Effect of Brenner number (*Br*) on relative concentration ($C/C_0$) in column effluent for a step-change application of solute.

breakthrough curve. Breakthrough curves generated from a continuous source will be approximately sigmoidal (resembling a cumulative Gaussian distribution function) provided the Brenner (or column Peclet) number ($Br \equiv vL/D_d$, where $L$ is the column length) is large enough. Complete breakthrough occurs when the effluent concentration ($C$) is equal to the influent concentration ($C_0$), giving a relative concentration $C/C_0 = 1$. If the solute is introduced as a short pulse, complete breakthrough may not be attained and the solute concentrations will eventually fall and return to background levels. In this case, a breakthrough curve will resemble a Gaussian density function for adequately large Brenner numbers. Small Brenner numbers (usually due to short columns or high dispersion coefficients) result in asymmetrical breakthrough curves.

It is common to plot breakthrough curves from soil columns as relative concentration against the number of pore volumes eluted from the column. Because the computation of eluted pore volume incorporates time and column length, it is a means of non-dimensionalizing the $x$-axis. The pore volume in a column of length $L$ and cross-sectional area $A$ is $LA\theta$. The volume of fluid eluted over time $t$ is $Atv\theta$. Thus, the number of pore volumes $T$ eluted at any given time $t$ is $(Atv\theta)/(LA\theta)$ or $T = tv/L$.

This representation has instructive and conceptual benefits because, during piston-type flow where the solute front is not dispersed, it is fairly obvious that, after the input of one pore volume of solution containing solute at concentration $C_0$ into an initially solute-free column, the concentration of the effluent will immediately jump from $C = 0$ to $C = C_0$. This is shown in Figure 5 where it corresponds to the curve with infinite Brenner number. Decreasing the Brenner number (for example by increasing the dispersion coefficient or shortening the column) causes spreading of the solute elution. Some solute travels faster than the mean velocity and elutes before one pore volume of fluid has entered the soil, while some travels slower than the mean and elutes after one pore volume of fluid is injected. When the Brenner number is relatively high (**Figure** 5; $Br = 1000$ and $Br = 100$), the early- and late-arriving solute masses are approximately equal, and the breakthrough curve is approximately sigmoidal and well described by a cumulative Gaussian distribution. The value of $C/C_0$ at one pore volume on this curve is approximately 0.5; this is considered a hallmark of an 'ideal' solute transport process by many practitioners, with deviations calling for elaborate modifications to standard theory. Inspection of the gross asymmetry introduced by decreases in the Brenner number (**Figure** 5; $Br = 10$ and $Br = 1$) indicates that such deviations should be expected under many conditions. This effect is pronounced for solute pulses, and is perhaps somewhat counterintuitive in that smaller Brenner numbers (more dispersion) can actually lead to higher peak concentrations than larger ones (**Figure** 6; $Br = 1$ vs. $Br = 10$).

## Adsorption and Exclusion

Adsorption of solutes on to soil solids generally causes them to move at a fraction of the velocity of the fluid. The most simplistic (and common) treatment of adsorption involves the assumption of a linear adsorption isotherm that relates the adsorbed concentration on soil solids ($C_s$, mass of solute per unit mass of soil) to the concentration in solution at equilibrium via $C_s = K_d C$, where $K_d$ is the soil/water partitioning or

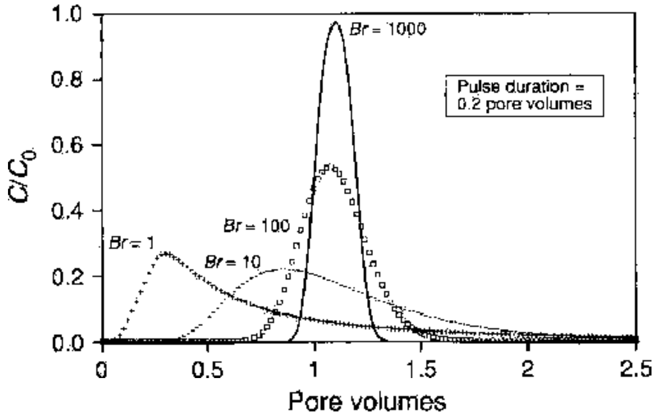


**Figure 6** Effect of Brenner number on concentration in column effluent for a pulse application of solute. Pulse duration is 0.2 pore volumes.

distribution coefficient with dimensions $L^3 M^{-1}$. The linear isotherm is primarily applicable to low-solute concentrations and adsorption of a nonspecific physical nature.

$K_d$ can be measured independently in the laboratory by mixing soil and solution and determining the equilibrium partitioning of the solute between the solid and fluid. In practice, however, the $K_d$ of organic solutes is often estimated from the organic carbon partitioning coefficient $K_{oc}(L^3M^{-1})$ using $K_d = K_{oc}f_{oc}$, where $f_{oc}(MM^{-1})$ is the mass fraction of organic carbon in a soil. There are many tables of $K_{oc}$ values available (e.g., **Table 1**) and $f_{oc}$ is a commonly measured soil property. Solutes with low $K_d$ values have a low affinity for soil surfaces and, for a conservative solute, $K_d = 0$. Other solutes can be strongly bound to soil and have very high $K_{oc}$ and $K_d$ values.

Exclusion, in which a solute is repelled from solid surfaces and hence concentrated in more rapidly flowing fluid away from pore walls, has also been observed. In particular, negatively charged anions are often excluded from negatively charged soil mineral surfaces. Thus, they may elute earlier from a column than the carrier fluid with which they were introduced.

Adsorption and exclusion phenomena are often incorporated into solute transport models through the use of a retardation factor, $R$, given by:

$$R = 1 + \frac{\rho_b K_d}{\theta} \qquad [3]$$

where $\rho_b$ is the soil bulk density. For conservative solutes, $R = 1$, for sorbing solutes, $R > 1$, and, for excluded solutes, $R < 1$.

Figure 7 shows the effects of solute adsorption and exclusion relative to the breakthrough curve for a

**Table 1** Organic carbon partitioning coefficients for nonionizable organic compounds

| *Compound* | *Mean* $K_{oc}$ *(l kg$^{-1}$)* |
|---|---|
| Acenaphthene | 5028 |
| Aldrin | 48 686 |
| Anthracene | 24 362 |
| Benz(*a*)anthracene | 459 882 |
| Benzene | 66 |
| Benzo(*a*)pyrene | 1 166 733 |
| Bis(2-chloroethyl)ether | 76 |
| Bis(2-ethylhexyl)phthalate | 114 337 |
| Bromoform | 126 |
| Butylbenzylphthalate | 14 055 |
| Carbon tetrachloride | 158 |
| Chlordane | 51 798 |
| Chlorobenzene | 260 |
| Chloroform | 57 |
| DDD | 45 800 |
| DDE | 86 405 |
| DDT | 792 158 |
| Dibenz(*a,h*)anthracene | 2 029 435 |
| 1,2-Dichlorobenzene (*o*) | 390 |
| 1,4-Dichlorobenzene (*p*) | 687 |
| 1,1-Dichloroethane | 54 |
| 1,2-Dichloroethane | 44 |
| 1,1-Dichloroethylene | 65 |
| *trans*-1,2-Dichloroethylene | 38 |
| 1,2-Dichloropropane | 47 |
| 1,3-Dichloropropene | 27 |
| Dieldrin | 25 604 |
| Diethylphthalate | 84 |
| Di-*n*-butylphthalate | 1580 |
| Endosulfan | 2040 |
| Endrin | 11 422 |
| Ethylbenzene | 207 |
| Fluoranthene | 49 433 |
| Fluorene | 8906 |
| Heptachlor | 10 070 |
| Hexachlorobenzene | 80 000 |
| $\alpha$-HCH ($\alpha$-BHC) | 1835 |
| $\beta$-HCH ($\beta$-BHC) | 2241 |
| $\gamma$-HCH (Lindane) | 1477 |
| Methoxychlor | 80 000 |
| Methylbromide | 9 |
| Methylchloride | 6 |
| Methylene chloride | 10 |
| Naphthalene | 1231 |
| Nitrobenzene | 141 |
| Pentachlorobenzene | 36 114 |
| Pyrene | 70 808 |
| Styrene | 912 |
| 1,1,2,2-Tetrachloroethane | 79 |
| Tetrachloroethylene | 272 |
| Toluene | 145 |
| Toxaphene | 95 816 |
| 1,2,4-Trichlorobenzene | 1783 |
| 1,1,1-Trichloroethane | 139 |
| 1,1,2-Trichloroethane | 77 |
| Trichloroethylene | 97 |
| *o*-Xylene | 241 |
| *m*-Xylene | 204 |
| *p*-Xylene | 313 |

Adapted from US Environmental Protection Agency. Soil Screening Guidance: Technical Background Document.

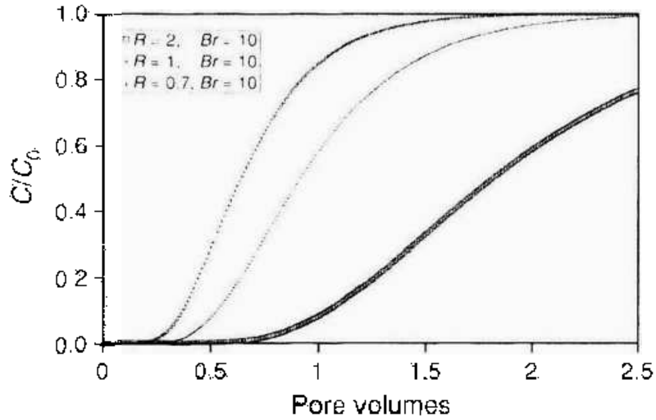


**Figure 7** Effects of retardation (*R*) and exclusion on breakthrough curve at *Br* = 10.

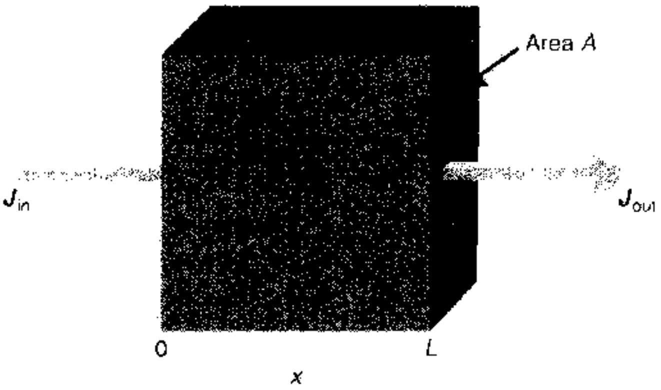


**Figure 8** Elemental volume, with one-dimensional solute fluxes. *J*, flux; *C*, concentration; *L*, length; *θ*, volumetric fluid content.

conservative solute ($R = 1$, $Br = 10$, in **Figure 7**). The retarded curve ($R = 2$, $Br = 10$, in **Figure 7**) shows a considerable delay in solute elution. Note also that the degree of solute dispersion is enhanced, even though the dispersion coefficient itself is unchanged. Conversely, the curve representing anion exclusion ($R = 0.7$, $Br = 10$, in **Figure 7**) shows early elution of the solute and less dispersion of the solute front.

For a continuous source under piston flow conditions, the solute front advances without dispersion and the velocity of the front ($v_s$) relative to the mean velocity of the fluid is $v_s = v/R$. With moderate dispersion and adequate transport length, the front assumes a sigmoidal shape, and $v_s$ represents the position of the front defined by $C/C_0 = 0.5$. For a pulse source, $v_s$ characterizes the velocity of the solute center of mass. In some instances, these simple relations may be adequate for estimating the solute front position or peak arrival. However, dispersion can play a critical role in determining the peak concentration and its duration. Evidence suggests that brief exposures to high concentrations may be exceptionally detrimental to human health in some instances.

## Advection–Dispersion Equation

The advection–dispersion equation (ADE, or convection–dispersion equation, CDE) has long been used as a model for solute transport. It can be derived from the flux terms described above (here taken as one-dimentional in the $x$ direction) and a solute mass balance. The flux of solute due to convection is $J_c = v\theta C$, while the diffusive–dispersive flux is $J_d = -D_d \partial C/\partial x$. Combining these two fluxes, we have $J = J_c + J_d = v\theta C - D_d \partial C/\partial x$. Consider a volume of soil (**Figure 8**) and steady one-dimensional flow. If the inflow of solute mass at $x = 0$ and the outflow at $x = L$ are not equal, then mass must accumulate in or be lost from the soil volume. Assuming for the moment that the solute can only be stored in the fluid, then either the volume of fluid at a constant concentration – given by $AL\theta$ – must change, or the solute concentration ($C$) in a fixed volume of fluid must change. The dimensions of the volume $AL$ are constant and, when the flow is steady, the fluid content $\theta$ is constant. Therefore, only the concentration can change. The time rate of change in solute mass inside the volume is $AL\theta\, \mathrm{d}C/\mathrm{d}t$.

The difference in the inflow and outflow of mass per unit time can be written as $(\text{mass}_{\text{out}} - \text{mass}_{\text{in}})/\Delta t$. The flux $J$ is the mass per unit area and time, so an equivalent expression is $A(J_{x=L} - J_{x=0})$. The spatial rate of change in solute flux is $\partial J/\partial x$, and the difference between the outflow and inflow is $A[J_{x=0} - (J_{x=0} + L\partial J/\partial x)]$. Equating this difference to the change in the stored mass gives:

$$AL\theta \frac{\partial C}{\partial t} = A\left[J_{x=0} - \left(J_{x=0} + L\frac{\partial J}{\partial x}\right)\right] \qquad [4]$$

which readily simplifies to:

$$\theta \frac{\partial C}{\partial t} = -\frac{\partial J}{\partial x} \qquad [5]$$

Incorporating our earlier expression for the solute flux $J$, we have:

$$\theta \frac{\partial C}{\partial t} = -\frac{\partial\left(v\theta C - D_d \frac{\partial C}{\partial x}\right)}{dx} \qquad [6]$$

Finally, with $D_d$, $v$, and $\theta$ constant, the ADE is:

$$\frac{\partial C}{\partial t} = -v\frac{\partial C}{\partial x} + \frac{D_d}{\theta}\frac{\partial^2 C}{\partial x^2} \qquad [7]$$

Eqn [7] is often written with $D_d/\theta$ replaced by an apparent diffusion–dispersion coefficient $D$.

The ADE can readily be modified to incorporate adsorption by accommodating solute storage on the soil solids in addition to storage in the fluid phase. Thus, the storage term becomes $AL\theta C + AL\rho_b C_s$ or, in terms of the solute concentration in the fluid and assuming a linear isotherm, $AL(\theta C + \rho_b K_d C)$. Then Eqn [6] becomes:

$$(\theta + \rho_b K_d)\frac{\partial C}{\partial t} = -\frac{\partial\left(v\theta C - D_d \frac{\partial C}{\partial x}\right)}{\partial x} \qquad [8]$$

Note that this formulation also assumes that the adsorption equilibrium is attained instantaneously. Alternative derivations that remove this restriction are available. Removing the constants from the differentials and dividing by $\theta$ yields:

$$R\frac{\partial C}{\partial t} = -v\frac{\partial C}{\partial x} + \frac{D_d}{\theta}\frac{\partial^2 C}{\partial x^2} \qquad [9]$$

where $R$ is the retardation factor given by Eqn [3].

Other isotherms, including the Langmuir (which introduces an adsorption maximum) and the Freundlich (which introduces a nonlinear concentration dependence), have been used to describe adsorption during solute transport. Realistic transport simulations of reactive species must often rely on considerably more complex, coupled reaction–transport models.

Many chemical species are susceptible to production or decay in the soil. Zero-order decay proceeds at a rate that is independent of the concentration of the substance where $\partial C/\partial t$ is a constant. First-order production and decay proceed at rates that depend on the concentration and follow $\partial C/\partial t = kC$, where $k$ is the rate constant ($T^{-1}$). Such potential sinks and sources, as well as physical sinks (e.g., root uptake) can readily be incorporated into the ADE.

Estimates of the parameters in the ADE can be obtained by moment analysis of solute breakthrough curves. A number of analytical solutions for the ADE are also available. Depending on the concentration detection mode (column effluent or 'resident' pore water) the following solutions have been recommended for a step change in concentration from $C_0$ to $C$ at time $t = 0$:

$$\text{Flux mode}: \frac{C}{C_0} = \frac{1}{2}\operatorname{erfc}\left[\frac{Rx - vt}{2(DRt)^{\frac{1}{2}}}\right] + \frac{1}{2}\exp\left(\frac{vx}{D}\right) \operatorname{erfc}\left[\frac{Rx - vt}{2(DRt)^{\frac{1}{2}}}\right] \qquad [10]$$

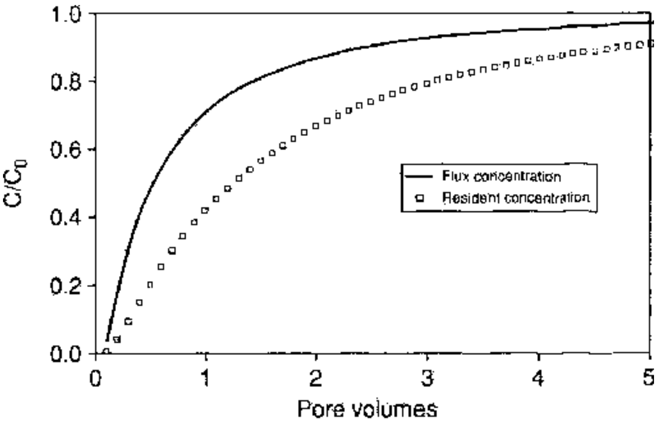


**Figure 9** Effect of concentration detection mode on breakthrough curve at $Br = 1$.

$$\text{Resident mode}: \frac{C}{C_0} = \frac{1}{2}\operatorname{erfc}\left[\frac{Rx - vt}{2(DRt)^{\frac{1}{2}}}\right] + \left(\frac{v^2 t}{\pi D r}\right)^{\frac{1}{2}} \exp\left(-\frac{(Rx - vt)^2}{4DRt}\right) - \frac{1}{2}\left(1 + \frac{vx}{D} + \frac{v^2 t}{DR}\right)\exp\left(\frac{vx}{D}\right) \operatorname{erfc}\left[\frac{Rx + vt}{2(DRt)^{\frac{1}{2}}}\right] \qquad [11]$$

Here 'exp' and 'erfc' are the exponential function and complementary error function, respectively. The differences between the two solutions become appreciable in highly dispersive systems. **Figure 9** shows the solutions at $Br = 1$. More complex and multidimensional formulations of the ADE generally require numerical solutions.

## Two-Region Advection–Dispersion Equation

This equation, also referred to as the 'mobile–immobile water model' (MIM), is a special case of the ADE, in which the liquid phase is divided into two distinct (mobile and immobile) regions. Advective–dispersive solute flux occurs within the mobile water region, with solutes entering and leaving the immobile water region by diffusion. The ADE for solute transport within the mobile water region can then be written as:

$$\frac{\partial C_m}{\partial t} = D\frac{\partial^2 C_m}{\partial x^2} - v_m\frac{\partial C_m}{\partial x} - \left(\frac{\theta_{im}}{\theta_m}\right)\frac{\partial C_{im}}{\partial t} \qquad [12]$$

where the subscripts 'm' and 'im' refer to the mobile and immobile regions, respectively. The kinetic term

$\theta_{\mathrm{im}}\mathrm{d}C_{\mathrm{im}}/\mathrm{d}t$ is usually expressed in terms of a first-order rate constant and the concentration difference $C_{\mathrm{m}} - C_{\mathrm{im}}$. Eqn [12] can be solved analytically with or without making any assumptions about the spatial arrangement of solids and voids. In the first case, the mobile and immobile liquid regions are explicitly defined in terms of pore-space geometry. The majority of these models are for packed beds of uniform-sized, spherical soil aggregates; here $\theta_{\mathrm{m}}$ is the macropore or inter-'aggregate' water content, and $\theta_{\mathrm{im}}$ is the intra-'aggregate' water content. While it is possible to extend this approach to nonuniform and nonspherical aggregates, its predictive capability is limited. This is because the relationship between packed beds of sieved aggregates and the pore characteristics of undisturbed soil is far from clear. In the second case, the mobile–immobile water fraction is simply a fitting parameter.

Eqn [12] is commonly employed to parameterize asymmetric solute breakthrough curves with 'heavy tails.' Deviations from the Gaussian distribution associated with the ADE are often observed in breakthrough curves where more of the observed distribution is in the tails. Improvements in fit associated with the MIM as compared to the ADE can probably often be attributed to the presence of additional model parameters. Since it is rarely possible to measure the mobile and immobile water fractions independently, this parameter is usually estimated inversely. A two-site conceptualization of the soil, where a fraction of the adsorption sites are considered to have noninstantaneous adsorption, leads to an equivalent mathematical formulation.

## Fractional Advection Dispersion and Continuous Time Random Walk Models

Recently, a fractional advection–dispersion equation (or FADE) and the continuous time random walk (CTRW) model have been proposed. These models appear to address the issues of scale-dependent dispersivity and heavy-tailed breakthrough curves well.

The FADE makes use of fractional calculus and is a generalization of the ADE that incorporates the family of Lévy distributions, of which the Gaussian distribution is a member. At the root of the FADE is the incorrect assumption in the derivation of the ADE above that the mass balance based on fluxes and fluid content applies in the infinitesimally small continuum limit. In fact it cannot, as the soil clearly divides into discrete pores and solids, and the averaged concepts of flux and fluid content lose meaning.

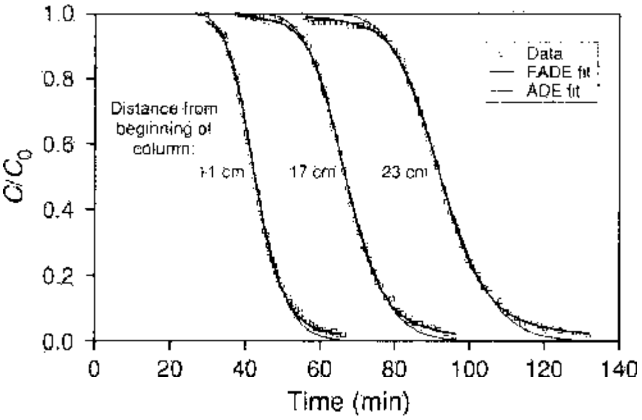


**Figure 10** Comparison of fractional advection–dispersion equation (FADE) and ADE fits to data from an unsaturated soil column. (Adapted from Pachepsky Y. Benson DA, and Timlin D (2001) Transport of water and solutes in soils as in fractal porous media. In: Sparks D and Selim M (eds) *Physical and Chemical Processes of Water and Solute Transport/Retention in Soils*, pp. 51–77. Special Publication 56. Madison, WI: Soil Science Society of America.)

**Table 2** Dispersion coefficients fitted to advection–dispersion equation (ADE) and fractional ADE (FADE) for breakthrough curves in **Figure 10**

| | *Dispersion coefficient* | |
|---|---|---|
| *Length (cm)* | *ADE ($cm^2\ min^{-1}$)* | *FADE ($cm^{1.6}\ min^{-1}$)* |
| 11 | 0.035 | 0.030 |
| 17 | 0.038 | 0.029 |
| 23 | 0.042 | 0.028 |

Adapted from Pachepsky Y. Benson DA, and Timlin D (2001) Transport of water and solutes in soils as in fractal porous media. In: Sparks D and Selim M (eds) *Physical and Chemical Processes of Water and Solute Transport/Retention in Soils*, pp. 51–77. Special Publication 56.
Madison, WI: Soil Science Society of America, with permission.

Instead of using a space-dependent dispersion coefficient to accommodate the apparent increase in $D$ with scale, the FADE uses a constant dispersion coefficient and a non-Fickian fractional derivative for the diffusive–dispersive flux. The one-dimensional, symmetrical FADE can be written:

$$\frac{\partial C}{\partial t} = -v\frac{\partial C}{\partial x} + \frac{1}{2}D\left[\frac{\partial^{\gamma} C}{\partial x^{\gamma}} + \frac{\partial^{\gamma} C}{\partial(-x)^{\gamma}}\right] \qquad [13]$$

where $\gamma$ is the fractional order of the derivative and $D$ has dimensions $L^{\gamma}/T$.

Figure 10 shows resident solute concentration data (measured with time-domain reflectometry probes) at different distances from the end of a single column. In this case, solute is being leached from the column. The data are fitted to both the ADE and the FADE. Clearly, the FADE is superior in its ability to fit the heavy tails of the breakthrough curves. Table 2 lists the dispersion coefficients for each of the models at each distance. Perhaps more important than the fitting of the tails is the fact that a single dispersion coefficient (approx. $0.03\,\text{cm}^{1.6}\,\text{min}^{-1}$) represents all of the curves in the case of the FADE, while the dispersion coefficients required to fit the ADE grow from approximately $0.035\,\text{cm}^2\,\text{min}^{-1}$ at 11 cm from the source to more than $0.04\,\text{cm}^2\,\text{min}^{-1}$ at 23 cm as the solute advances through the column. The FADE allows the parameter describing the changes with scale ($\gamma \approx 1.6$ in this case) to be separated from the transport parameter $D$, which ideally needs to be measured at one scale only.

Like the FADE, the CTRW model is founded on a conceptual picture in which solute particles move along various paths and encounter spatially varying velocities. The particle spatial transitions (direction and distance given by displacement vector s) in time $t$ are represented by a joint probability density function $\Psi(\mathrm{s}, t)$. Estimation of this function is central to application of the CTRW model. The functional form $\Psi(\mathrm{s}, \sim t) \sim t^{-1-\beta}$ ($\beta > 0$) has been found to be of particular interest. $\beta$ characterizes the nature and magnitude of the dispersive processes. $\beta \geq 2$ is reported to be equivalent to the ADE. $\beta$ between 1 and 2 reflects moderate non-Fickian behaviour, while $0 < \beta < 1$ indicates strong 'anomalous' behaviour.

The ability of the CTRW model to fit observations is similar to that of the FADE for the results in Figure 10. Three parameters ($\beta$, $C^*$, and $C_1$) are involved. For breakthrough curves in time, available fitting routines return $\beta$, $T^*$, and $r^*$, which, for $1 < \beta < 2$, are related to $C^*$ and $C_1$ as follows:

$$\begin{array}{c} T^* = LC_1 \text{ and } r^* = C_1^*/\beta L^{-1+1/\beta} \\ \qquad\qquad\qquad\quad C_1 \end{array}$$

where $L$ is the distance from the source. Inverting these equations gives $C^*$ and $C_1$, which can then be used to compute breakthrough curves at different locations. Thus $C^*$ and $C_1$ should be constants for a 'stationary' porous medium (Table 3). Here there is systematic variation, but the precise nature of the medium and the significance of the magnitude of this variation are unknown.

Despite the promise of these models, their application in soil science remains almost completely unexplored currently.

**Table 3** Parameters for continuous time random walk (CTRW) fitted to data in **Figure 10**. $C^*$ and $C_1$ should be constant for solute transport in a 'stationary' porous medium

| *Length* | *CTRW parameters* | | | | |
|---|---|---|---|---|---|
| *(cm)* | $T^*$ | $\beta$ | $r^*$ | $C^*$ | $C_1$ |
| 11 | 44.1 | 1.68 | 0.11 | 1.92 | 4.01 |
| 17 | 68.5 | 1.72 | 0.092 | 2.09 | 4.03 |
| 23 | 94.2 | 1.73 | 0.084 | 2.23 | 4.10 |

## List of Technical Nomenclature

| | |
|---|---|
| $\alpha$ | Longitudinal dispersivity (L) |
| $\beta$ | Characterizes the nature of magnitude of CTRW dispersive processes (–) |
| $\gamma$ | Order of fractional derivative |
| $\Delta P$ | Pressure drop ($\text{MLT}^{-2}$) |
| $\eta$ | Dynamic viscosity ($\text{MLT}^{-1}$) |
| $\theta$ | Volumetric fluid content ($\text{L}^3\text{L}^{-3}$) |
| $\theta_{\text{im}}$ | Volumetric water content of the immobile region ($\text{L}^3\text{L}^{-3}$) |
| $\theta_{\text{m}}$ | Volumetric water content of the mobile region ($\text{L}^3\text{L}^{-3}$) |
| $\rho_{\text{b}}$ | Soil bulk density ($\text{ML}^{-3}$) |
| $\phi$ | Porosity ($\text{L}^3\text{L}^{-3}$) |
| $\Psi(\mathrm{s}, t)$ | Joint probability density function for CTRW (–) |
| $A$ | Cross-sectional area of column and elemental volume ($\text{L}^2$) |
| $Br$ | Brenner number |
| C | Solute concentration ($\text{ML}^{-3}$) |
| $C_0$ | Source solute concentration ($\text{ML}^{-3}$) |
| $C_{\text{im}}$ | Solute concentration in the immobile water region ($\text{ML}^{-3}$) |
| $C_{\text{m}}$ | Solute concentration in the mobile water region ($\text{ML}^{-3}$) |
| $C_{\text{s}}$ | Adsorbed concentration on soil ($\text{MM}^{-1}$) |
| $C^*$, $C_1$, $r^*$, $T^*$ | CTRW parameters |
| $D$ | Apparent diffusion–dispersion coefficient ($\text{L}^2\text{T}^{-1}$) in ADE and ($\text{L}^{\gamma}\text{T}^{-1}$) in FADE |
| $D^*$ | Dimensionless diffusion–dispersion coefficient (–) |
| $D_{\text{b}}$ | Bulk diffusion coefficient ($\text{L}^2\text{T}^{-1}$) |
| $D_{\text{e}}$ | Effective diffusion coefficient ($\text{L}^2\text{T}^{-1}$) |
| $D_{\text{d}}$ | Lumped diffusion–dispersion coefficient ($\text{L}^2\text{T}^{-1}$) |
| $f_{\text{oc}}$ | Fraction of organic carbon in a soil |

| | |
|---|---|
| $J$ | Solute flux ($ML^{-2}T^{-1}$) |
| $J_c$ | Convective flux of solute ($ML^{-2}T^{-1}$) |
| $J_d$ | Diffusive-dispersive flux of solute ($ML^{-2}T^{-1}$) |
| $J_s$ | Diffusive flux of solute ($ML^{-2}T^{1}$) |
| $k$ | First-order rate constant ($T^{-1}$) |
| $K_d$ | Soil and/or water partitioning or soil-water distribution coefficient ($L^3M^{-1}$) |
| $K_{oc}$ | Organic carbon partitioning coefficient ($L^3M^{-1}$) |
| $l$ | Characteristic length (L) |
| $L$ | Length of column, elemental volume, and capillary tube (L) |
| $Pe$ | Peclet number (–) |
| $q$ | Fluid flux ($L^3L^{-2}T^{-1}$) |
| $r$ | Capillary tube radius (L) |
| $R$ | Retardation factor (–) |
| $\mathbf{s}$ | Spatial transition vector in CTRW (L) |
| $T$ | Pore volumes eluted (–) |
| $t$ | Time (T) |
| $v$ | Average fluid velocity ($LT^{-1}$) |
| $v_m$ | Average liquid velocity in the mobile region ($LT^{-1}$) |
| $v_{max}$ | Maximum velocity ($LT^{-1}$) |
| $v_s$ | Solute velocity ($LT^{-1}$) |
| $x$ | Spatial coordinate aligned with average fluid velocity (L) |
| $y$ | Radial distance (L) |

*See also:* **Cation Exchange**; **Chemical Equilibria**; **Colloid-Facilitated Sorption and Transport**; **Leaching Processes**; **Pesticides**; **Pollution**: Groundwater; **Remediation of Polluted Soils**; **Sorption-Desorption, Kinetics**; **Vadose Zone**: Hydrologic Processes; **Waste Disposal on Land**: Liquid; Municipal; **Water Potential**

## Further Reading

Adler PM (1992) *Porous Media: Geometry and Transports.* Boston, MA: Butterworth-Heinemann.

Bear J (1972) *Dynamics of Flow in Porous Media.* New York: American Elsevier.

Bear J and Bachmat Y (1990) *Introduction to Modelling of Transport Phenomena in Porous Media.* Dodrecht, the Netherlands: Kluwer Academic.

Dullien FAL (1992) *Porous Media: Fluid Transport and Pore Structure,* 2nd edn. New York: Academic Press.

Elrick DE and Clothier BE (1990) *Solute Transport and leaching.* Agronomy Monograph 30, pp. 93–126. Madison, WI: American Society of Agronomy.

Fried JJ and Combarnous MA (1971) Dispersion in porous media. *Advances in Hydroscience* 7: 169–282.

Hillel D (1998) *Environmental Soil Physics.* New York: Academic Press.

Jury WA and Flühler H (1992) Transport of chemicals through soil: mechanisms, models, and field applications. *Advances in Agronomy* 47: 141–201.

Kutílek M and Nielsen DR (1994) *Soil Hydrology.* Reiskirchen, Germany: Catena-Verlag.

Leij FJ and Dane JH (1989) *A Review of Physical and Chemical Processes Pertaining to Solute Transport.* Agronomy and Soils Department Series No. 132, Alabama Agricultural Experiment Station. Auburn, AL: Auburn University Press.

Parker JC and van Genuchten MT (1984) *Determining Transport Parameters from Laboratory and Field Tracer Experiments.* Bulletin 84-3. Blacksburg, VA: Virginia Agricultural Experiment Station.

Sahimi M (1995) *Flow and Transport in Porous Media and Fractured Rock.* Weinheim, Germany: VCH-Verlag.

Selim HM and Sparks DL (eds) (2001) *Physical and Chemical Processes of Water and Solute Transport/Retention in Soils,* Special Publication 56. Madison, WI: Soil Science Society of America.

Toride N, Leij FJ, and van Genuchten MT (1995) *The CXTFIT Code for Estimating Transport Parameters From Laboratory or Field Tracer Experiments.* Research Report No. 137, US Salinity Laboratory. Riverside, CA: USDA-ARS.

van Genuchten MT and Wierenga PJ (1986) Solute dispersion coefficients and retardation factors. In: Klute A (ed.) *Methods of Soil Analysis,* 2nd edn, part 1, pp. 1025–1054. Madison, WI: American Society of Agronomy.

# SORPTION

Contents

## Metals

**D L Sparks**, University of Delaware, Newark, DE, USA

### Introduction

The sorption (retention) of metals, including alkali (e.g., K), alkaline earth (e.g., Ca), and transition (e.g., Cd and Ni), on soil mineral and organic constituents is one of the most important processes in controlling the fate, transport, and bioavailability of metals in soil and water environments. This review discusses terminology for sorption processes, the role of surface functional groups on inorganic and organic soil components in sorption of metals, the types of surface complexes and products that metals can form in soils, characteristics of metal sorption in soils, and the importance of surface precipitation of metals such as Co, Ni, and Zn in sequestering metals in soils.

### Terminology and Role of Sorption Processes in Soils

Adsorption, surface precipitation, and polymerization are all examples of sorption, a general term that is used when the retention mechanism at a surface is unknown. Adsorption can be defined as the accumulation of a substance or material at an interface between the solid surface and the bathing solution. Adsorption can include the removal of solute (a substance dissolved in a solvent) molecules from the solution, solvent (continuous phase of a solution, in which the solute is dissolved) from the solid surface, and attachment of the solute molecule to the surface. Adsorption does not include surface precipitation (the formation of a three-dimensional phase product on a surface), or polymerization (the formation of small multinuclear inorganic species such as dimers or trimers) processes. There are various metal sorption mechanisms that can occur at soil mineral surfaces involving both physical and chemical processes (Figure 1). It is useful to define a number of terms pertaining to retention (adsorption/sorption) of metal ions. The adsorbate is the material that accumulates at an interface, the solid surface on which the adsorbate accumulates is referred to as the adsorbent, and the ion in solution that has the potential of being adsorbed is the adsorptive. If the general term 'sorption' is used, the material that accumulates at the surface, the solid surface, and the ion in solution that can be sorbed are referred to as sorbate, sorbent, and sorptive, respectively.

Adsorption is one of the most important chemical processes in soils. It determines the quantity of plant nutrients, metals, pesticides, and other organic chemicals that are retained on soil surfaces and therefore is one of the primary processes that affects transport of nutrients and contaminants in soils. Adsorption also affects the electrostatic properties, e.g., coagulation and settling, of suspended particles and colloids. Both physical and chemical forces are involved in adsorption of solutes from solution. Physical forces include van der Waals forces (e.g., partitioning) and electrostatic outer-sphere complexes (e.g., ion exchange). Chemical forces result from short-range interactions that include inner-sphere complexation (which is defined below) involving a ligand exchange mechanism, covalent bonding, and hydrogen bonding.

### Surface Functional Groups

Surface functional groups in soils play a significant role in adsorption processes. A surface functional group is a chemically reactive molecular unit attached at the boundary of a solid with the reactive groups of the unit exposed to the soil solution. Surface functional groups can be organic (e.g., carboxyl, carbonyl, phenolic) or inorganic molecular units. The major inorganic surface functional groups in soils are the siloxane surfaces associated with the plane of oxygen atoms bound to the silica tetrahedral layer of a phyllosilicate (clay mineral) and hydroxyl groups that are associated with the edges of inorganic minerals such as kaolinite, amorphous materials, and metal oxides, oxyhydroxides, and hydroxides.

### Surface Complexation

When the interaction of a surface functional group with an ion or molecule present in the soil solution

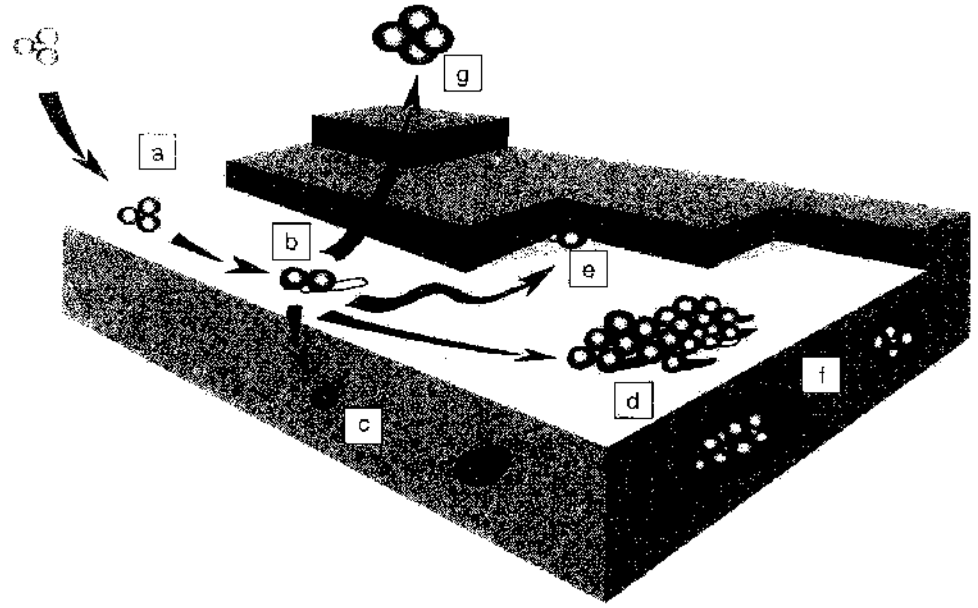


**Figure 1** Various mechanisms of sorption of an ion at the mineral–water interface: (1) adsorption of an ion via formation of an outer-sphere complex (a); (2) loss of hydration water and formation of an inner-sphere complex (b); (3) lattice diffusion and isomorphic substitution within the mineral lattice (c); (4) and (5) rapid lateral diffusion and formation either of a surface polymer, (d), or adsorption on a ledge (which maximizes the number of bonds to the adatom) (e). Upon particle growth, surface polymers end up embedded in the lattice structure (f); finally, the adsorbed ion can diffuse back in solution, either as a result of dynamic equilibrium or as a product of surface redox reactions (g). Reproduced from Charlet L and Manceau A (1993) Structure, formation, and reactivity of hydrous oxide particles: insights from X-ray absorption spectroscopy. In: *Environmental Particles*, pp. 117–164. Boca Raton, FL: Lewis Publishers.

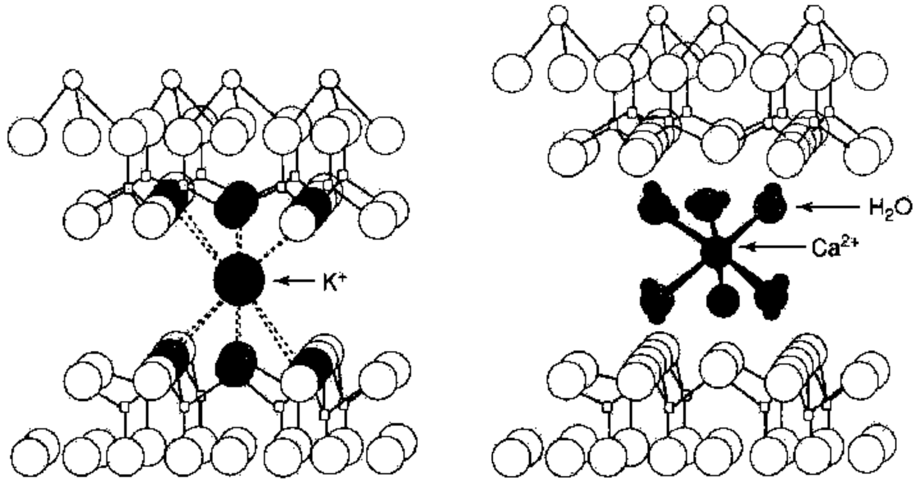
creates a stable molecular entity, it is called a surface complex. The overall reaction is referred to as surface complexation. There are two types of surface complexes that can form: outer-sphere and inner-sphere. Figure 2 shows surface complexes between metal cations and siloxane ditrigonal cavities on 2:1 clay minerals. Such complexes can also occur on the edges of clay minerals. If a water molecule is present between the surface functional group and the bound ion or molecule, the surface complex is termed outer-sphere. If there is not a water molecule present between the ion or molecule and the surface functional group to which it is bound, this is an inner-sphere complex. Inner-sphere complexes can be monodentate (metal is bonded to only one oxygen) and bidentate (metal is bonded to two oxygens) and mononuclear and binuclear.

Outer-sphere complexes involve electrostatic coulombic interactions and are thus weak compared to inner-sphere complexes in which the binding is covalent or ionic. Outer-sphere complexation is usually a rapid process that is reversible, and adsorption only occurs on surfaces that are of opposite charge to the adsorbate.

Inner-sphere complexation is usually slower than outer-sphere complexation; it is often irreversible. Inner-sphere complexation can increase, reduce, neutralize, or reverse the charge on the sorptive regardless of the original charge. Adsorption of ions via inner-sphere complexation can occur on a surface regardless of the original charge. It is important to remember that outer- and inner-sphere complexation can, and often does, occur simultaneously.

Ionic strength effects on metal sorption are often used as indirect evidence for whether an outer-sphere or inner-sphere complex forms. For example, if metal sorption is dependent on ionic strength ($I$), outer-sphere complexes are assumed to form, while if metal sorption is independent of $I$, inner-sphere complexes are assumed to be the predominant surface complex. However, such conclusions are based strictly on macroscopic investigations. To determine definitively the type of surface complex or product that metals form on soil surfaces, one must employ molecular-scale investigations.

Over the past decade or more, *in situ* (under natural conditions where water is present) molecular-scale analytical techniques such as X-ray absorption fine-structure spectroscopy (XAFS) and Fourier transform infrared spectroscopy (FTIR) have been employed to determine the type of metal surface complexes and products that form on soil surfaces. Based on these studies, one can conclude that alkali metals and alkaline earth metals primarily form outer-sphere complexes, while trace and heavy metals primarily form inner-sphere complexes. At higher



Inner-sphere surface complex: K$^+$ on vermiculite

Outer-sphere surface complex: $Ca(H_2O)_6^{2+}$ on montmorillonite

**Figure 2** Examples of inner- and outer-sphere complexes formed between metal cations and siloxane ditrigonal cavities on 2:1 clay minerals. Reproduced with permission from Sposito G (1984) *The Surface Chemistry of Soils*. New York: Oxford University Press.

metal loadings and pHs, sorption of metals such as Co, Cr, Ni, and Zn on phyllosilicates and metal hydr(oxides) can result in the formation of surface precipitates. Thus, environmental factors such as pH, surface loading, ionic strength, type of sorbent, and time all affect the type of sorption complex or product. An example of this is shown for Pb sorption on montmorillonite over an $I$ of 0.006–0.1 and a pH range of 4.48–6.77 (**Table 1**). Employing XAFS analysis, at a pH of 4.48 and an $I$ of 0.006, outer-sphere complexation on basal planes in the interlayer regions of the montmorillonite predominated. At a pH of 6.77 and $I$ of 0.1, inner-sphere complexation on edge sites of montmorillonite was most prominent, and at pH of 6.76, $I$ of 0.006 and pH of 6.31, $I$ of 0.1, both inner- and outer-sphere complexation occurred. These data are consistent with other findings that inner-sphere complexation is favored at higher pH and ionic strength. Clearly, there is a continuum of adsorption complexes that can exist in soils.

## Sorption of Metal Cations

Sorption of metal cations is pH-dependent and is characterized by a narrow pH range where sorption increases to nearly 100%, traditionally known as an adsorption edge (Figure 3). The pH position of the adsorption edge for a particular metal cation is related to its hydrolysis or acid-base characteristics. In addition to pH, sorption of metals is dependent on sorptive concentration, surface coverage, and the type of the sorbent.

One can measure the relative affinity of a cation for a sorbent or the selectivity. The properties of the cation, sorbent, and the solvent all affect the selectivity. With monovalent alkali metal cations, electrostatic interactions predominate and the general order of selectivity is $Li^+ < Na^+ < K^+ < Rb^+ < Cs^+$. This order is related to the size of the hydrated radius. The ion in the above group with the smallest hydrated radius, $Cs^+$, can approach the surface the closest and be held the most tightly. However, on some hydrous oxides, the reverse order is often observed.

This has been particularly noted for some hydrous metal oxides. The reason for this selectivity is not well understood, but may be related to the effect of the solid on water that is present on the oxide surface or to variation in the solution matrix.

With divalent metal ions there is little consistency in the selectivity order. The type of surface and the pH both appear to have major effects on the selectivity sequence. Differences in the $H^+/M^{n+}$ (where $M^{n+}$ is a metal ion of valence n) could cause reversals in selectivity since the ion with the higher $H^+/M^{n+}$ stoichiometry would be favored at higher pH. Divalent transition and heavy metal cations, both of which are often sorbed as inner-sphere complexes, are more strongly sorbed than alkaline earth cations.

## Metal Surface Precipitation

As the amount of metal cation sorbed on a surface (surface coverage or loading which is affected by the pH at which sorption occurs) increases, sorption can proceed from mononuclear adsorption to surface precipitation. There are several thermodynamic reasons for surface precipitate formation: 1) the solid surface may lower the energy of nucleation by providing sterically similar sites; 2) the activity of the surface precipitate is <1; and 3) the solubility

**Table 1** Effect of ionic strength (*I*) and pH on type of Pb adsorption complexes on montmorillonite

| *I (mol l$^{-1}$)* | *pH* | *Removal from solution (%)* | *Adsorbed Pb(II) (mmol kg$^{-1}$)* | *Primary adsorption complex*[a] |
|---|---|---|---|---|
| 0.1 | 6.77 | 86.7 | 171 | Inner-sphere |
| 0.1 | 6.31 | 71.2 | 140 | Mixed |
| 0.006 | 6.76 | 99.0 | 201 | Mixed |
| 0.006 | 6.40 | 98.5 | 200 | Outer-sphere |
| 0.006 | 5.83 | 98.0 | 199 | Outer-sphere |
| 0.006 | 4.48 | 96.8 | 197 | Outer-sphere |

[a] Based on results from X-ray absorption fine-structure spectroscopy data analysis.

Reproduced with permission from Strawn and Sparks (1999).

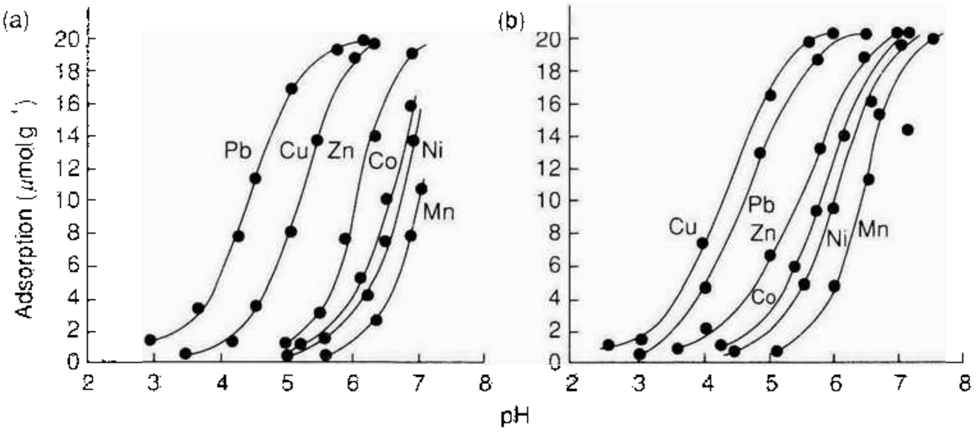


**Figure 3** Sorption of a range of metals on (a) hematite and (b) goethite when they were added at a rate of 20 μmol g$^{-1}$ of adsorbate. The values for the p$K_1$ for dissociation of the metals to give the monovalent $MOH^+$ ions are Pb, 7.71; Cu, 8; Zn, 8.96; Co, 9.65; Ni, 9.86; and Mn, 10.59. Reproduced with permission from McKenzie RM (1980) The adsorption of lead and other heavy metals on oxides of manganese and iron, *Australian Journal of Soil Research* 18: 61–73.

of the surface precipitate is lowered because the dielectric constant of the solution near the surface is less than that of the bulk solution. There are several types of surface precipitates. They can arise via polymeric metal complexes (dimers, trimers, etc.) that form on mineral surfaces and via the sorption of aqueous polymers. Homogeneous precipitates can form on a surface when the solution becomes saturated and the surface acts as a nucleation site. When adsorption attains monolayer coverage sorption continues on the newly created sites causing a precipitate on the surface. When the precipitate consists of chemical species derived from both the aqueous solution and dissolution of the mineral, it is referred to as a coprecipitate. The composition of the coprecipitate varies between that of the original solid and a pure precipitate of the sorbing metal. The ionic radius of the sorbing metal and sorbent ions must be similar for coprecipitates to form. Thus Co(II), Mn(II), Ni(II), and Zn(II) form coprecipitates on sorbents containing Al(III) and Si(IV) but not Pb(II) which is considerably larger (1.20 Å). Coprecipitate formation is most limited by the rate of mineral dissolution, rather than the lack of thermodynamic favorability. If the formation of a precipitate occurs under solution conditions that would, in the absence of a sorbent, be undersaturated with respect to any known solid phase, this is referred to as surface-induced precipitation.

Thus there is often a continuum between surface complexation (adsorption) and surface precipitation. This continuum depends on several factors: 1) ratio of the number of surface sites vs. the number of metal ions in solution; 2) the strength of the metal-oxide bond; and 3) the degree to which the bulk solution is undersaturated with respect to the metal hydroxide precipitate. At low surface coverages surface complexation (e.g., outer- and inner-sphere adsorption) tends to dominate. As surface average increases, nucleation occurs and results in the formation of distinct entities or aggregates on the surface. As surface loadings increase further, surface precipitation becomes the dominant mechanism (**Figure 4**).

Using an *in-situ* molecular-scale technique, X-ray absorption fine structure spectroscopy (XAFS), it has been shown by a number of scientists that multinuclear metal hydroxide complexes and surface precipitates of Co(II), Cr(III), Cu(II), Ni(II), Pb(II), and Zn(II) can form on metal oxides, phyllosilicates, soil clays, and soils. These metal hydroxide phases occur

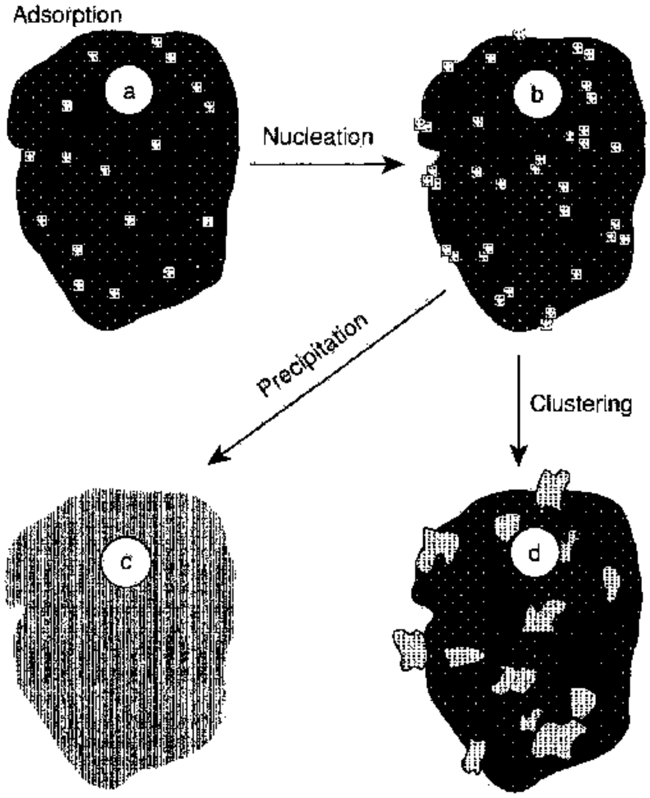


**Figure 4** An illustration of metal ion sorption reactions on (hydr)oxide. (a) At low surface coverage, isolated site binding (adsorption) is the dominant sorption mechanism; (b) with increased metal loading, M-hydroxide nucleation begins; (c) further increases in metal loadings results in surface precipitation or (d) surface clusters. Reproduced with permission from Fendorf SE (1992) *Oxidation and Sorption Mechanisms of Hydrolysable Metal Ions on Oxide Surfaces.* PhD dissertation. Newark, DE: University of Delaware.

at metal loadings below a theoretical monolayer coverage and in a pH range well below the pH where the formation of metal hydroxide precipitates would be expected according to the thermodynamic solubility product.

Sorption of Co, Ni, and Zn, on an array of phyllosilicates and Al-oxide, can result in formation of mixed metal-Al hydroxide surface precipitates which appear to be coprecipitates. The precipitate phase shares structural features common to the hydrotalcite group of minerals and the layered double hydroxides (LDH) observed in catalyst synthesis. The LDH structure is built of stacked sheets of edge-sharing metal octahedra containing divalent and trivalent metal ions separated by anions between the interlayer spaces. The general structural formula can be expressed as $[Me^{2+}_{1-x}Me^{3+}_{x}(OH)_2]^{x+} \cdot (x/n)A^{-n} \cdot mH_2O$, where for example, $Me^{2+}$ could be Mg(II), Ni(II), Co(II), Zn(II), Mn(II), and Fe(II) and $Me^{3+}$ is Al(III), Fe(III), and Cr(III). The LDH structure exhibits a net positive charge x per formula unit which is balanced by an equal negative charge from interlayer anions $A^{n-}$ such as $Cl^-$, $Br^-$, $I^-$, $NO_3^-$, $OH^-$, $ClO_4^-$, and $CO_3^{2-}$; water molecules occupy the remaining interlayer space. The minerals takovite, $Ni_6Al_2(OH)_{16}CO_3 \cdot H_2O$ and hydrotalcite, $Mg_6Al_2(OH)_{16}CO_3 \cdot H_2O$ are among the most common natural mixed-cation hydroxide compounds containing Al.

The mechanism for the formation of metal hydroxide surface precipitates is not clearly understood. It is clear that the type of metal ion determines whether metal hydroxide surface precipitates form, and the type of surface precipitate formed, i.e., metal hydroxide or mixed metal hydroxide, is dependent on the sorbent type. The precipitation could be explained by the combination of several processes. First, the electric field of the mineral surface attracts metal ions, e.g., Ni, through adsorption, leading to a local supersaturation at the mineral–water interface. Second, the solid phase may act as a nucleation center for polyhydroxy species and catalyze the precipitation process. Third, the physical properties of water molecules adsorbed at the mineral surface are different from free water, causing a lower solubility of metal hydroxides at the mineral–water interface. With time Al, which is released by weathering of the mineral surface, slowly diffuses into the octahedral layer of the mineral and partially replaces the metal (e.g., Ni) in the octahedral sites. A Ni-Al (LDH) is formed, which is thermodynamically favored over α-Ni hydroxide.

The formation of metal hydroxide surface precipitates appears to be an important way to sequester metals. As the surface precipitates age, metal release is greatly reduced. Thus, the metals are less prone to leaching and being taken up by plants. This is due to silication of the interlayer of the LDH phases creating precursor phyllosilicate surface precipitates.

*See also:* **Colloid-Facilitated Sorption and Transport; Heavy Metals; Metal Oxides; Metals and Metalloids, Transformation by Microorganisms**

## Further Reading

Barrow NJ (1985) Reactions of anions and cations with variable-charged soils. *Advances in Agronomy* 38: 183–230.

Charlet L and Manceau A (1993) Structure, formation, and reactivity of hydrous oxide particles: Insights from X-ray absorption spectroscopy. In: *Environmental Particles*, pp. 117–164. Boca Raton, FL: Lewis Publishers.

Fendorf SE (1992) *Oxidation and Sorption Mechanisms of Hydrolysable Metal Ions on Oxide Surfaces*. Ph.D. Dissertation. Newark, DE: University of Delaware.

McBride MB (1991) Processes of heavy and transition metal sorption by soil minerals. In: *Interactions at the Soil Colloid–Soil Solution Interface*, vol. 190, pp. 149–176. Dordrecht: Kluwer Academic Publishers.

McKenzie RM (1980) The adsorption of lead and other heavy metals on oxides of manganese and iron. *Australian Journal of Soil Research* 18: 61–73.

Scheidegger AM, Lamble GM, and Sparks DL (1997) Spectroscopic evidence for the formation of mixed-cation, hydroxide phases upon metal sorption on clays and aluminum oxides. *Journal of Colloid Interface Science* 186: 118–128.

Scheidegger AM and Sparks DL (1996) A critical assessment of sorption-desorption mechanisms at the soil mineral/water interface. *Soil Science* 161: 813–831.

Schindler PW (1981) Surface complexes at oxide-water interfaces. In: *Adsorption of Inorganics at Solid-Liquid Interfaces*, pp. 1–49. Ann Arbor, MI: Ann Arbor Science.

Sparks DL (2002) *Environmental Soil Chemistry*, 2nd Ed. San Diego, CA: Academic Press.

Sposito G (1984) *The Surface Chemistry of Soils*. New York: Oxford University Press.

Sposito G (1989) *The Chemistry of Soils*. New York: Oxford University Press.

Strawn DG and Sparks DL (1999) The use of XAFS to distinguish between inner- and outer-sphere lead adsorption complexes on montmorillonite. *Journal of Colloid Interface Science* 216: 257–269.

Strawn DG and Sparks DL (1999) Sorption kinetics of trace elements in soils and soil materials. In: Selim HM and Iskandar I (eds) *Fate and Transport of Heavy Metals in the Vadose Zone*, pp. 1–28. Chelsea, MI: Lewis Publishers.

Stumm W (1992) *Chemistry of the Solid–Water Interface*. New York: John Wiley & Sons.

Towle SN, Bargar JR, Brown GE Jr, and Parks GA (1997) Surface precipitation of Co(II) (aq) on $Al_2O_3$. *Journal of Colloid Interface Science* 187: 62–82.

# Organic Chemicals

**B Xing**, University of Massachusetts, Amherst, MA, USA

**J J Pignatello**, Connecticut Agricultural Experimental Station, New Haven, CT, USA

## Introduction

Organic chemicals enter soils through the use of pesticides, application of sewage sludge, or the disposal or accidental release of petrochemicals or industrial wastes. Some have negative effects on the environment and public health. A key process in soils regulating the transport, bioavailability, and reactivity of organic compounds is sorption. 'Sorption' and the opposite term, 'desorption,' refer to the process of molecular exchange between the gas or solution phase and an immobile phase (the sorbent). The sorbed component is sometimes referred to as the sorbate (American) or sorptive (British).

Sorption may be categorized by the type of intermolecular attractive forces between sorbate and sorbent. Physisorption and ion exchange sorption involve relatively weak intermolecular forces, while chemisorption involves strong bonding of a covalent nature. Physisorption is, by far, the most common type encountered in the context of soil and environmental sciences.

The reactive mineral components in soils are silicate clays and oxyhydroxide minerals of Si, Fe, Al, Mg, Mn, etc. Soil organic matter (SOM) consists mainly of polymeric humic substances derived from decomposition of biological materials such as lignin and cutin. Other substances within the scope of SOM are carbonaceous materials such as soot and charcoal. SOM is usually the predominant sorbent unless the SOM content is extremely low – some estimates suggest less than 0.01% of organic carbon by weight – or unless the water content is low. Therefore, the focus here is on sorption by SOM. Emphasis is placed on sorption of nonionic compounds in water-saturated systems, because the majority of organic compounds, including pesticides, are uncharged within the range of pH usually encountered in soils (5.5–7.5), and because water is abundant, especially in the subsurface.

## Physisorption of Nonionic Compounds

### General Considerations

**Forces involved** Physisorption involves weak intermolecular forces such as London–van der Waals forces and hydrogen bonding. These forces act simultaneously in combinations permitted by the structures of the interacting species. London–van der Waals forces include dipole–dipole, dipole–induced dipole (attraction of a permanent dipole with the dipole it induces in its neighbor), and induced dipole–induced dipole (mutual attraction of momentary dipoles produced by the synchronization of electronic motion in interacting neighbors). Additionally, the interaction of a charge with a neutral molecule may involve charge–dipole and charge–induced-dipole forces. The strengths of London–van der Waals interactions depend on the molecular area of contact, separation distance (to the inverse sixth power), and applicable molecular properties of the interacting species such as dipole moment, ionization potential, and polarizability.

The most common type of hydrogen bond is between an acidic proton and a lone pair of electrons (i.e., $XH\cdots:Y$) which are located on O, N, or S atoms. This bond is highly oriented (XHY angle, 180° ± 15°), ranges in enthalpy from 12 to $35\,\mathrm{kJ\,mol^{-1}}$, and involves a combination of dipole–dipole attraction and molecular orbital overlap. Other much weaker hydrogen bonds are known, such as the one formed between an acidic proton and the $\pi$ system of an aromatic ring, but whether they are important in soil systems in competition with hydrogen bonds from water molecules is unclear.

**Categories** Physisorption is often broadly classified as absorption or adsorption. Absorption (often called partitioning) occurs when molecules penetrate the solid–fluid interface and are allowed to roam the internal molecular or atomic matrix of the sorbent. SOM is the only natural component of soils that is penetrable in this way, but polluted soils may contain additional absorptive phases such as solvents, oils, tars, etc.

Adsorption is a term used in the phenomenological sense to imply association at a solid–fluid interface. Mechanistically, however, adsorption may encompass several qualitatively different processes, including binding of molecules independently to discrete surface sites, partitioning of molecules into an ordered hydration phase near the surface, condensation of the sorbate into a liquid-like or solid-like state in small pores, and adsorption on the surfaces of water films associated with the particles.

**Thermodynamic considerations** Net sorption from water (eqn [1e]) can be written as a series of steps in a thermodynamic cycle:

$$\{x\} \rightleftharpoons \{\ldots\} + x_g \qquad [1a]$$

$$\{\ldots\} \rightleftharpoons (\ ) \qquad [1b]$$

$$\mathbf{M} \rightleftharpoons \mathbf{M}' \qquad [1c]$$

$$x_g + \mathbf{M}' \rightleftharpoons x\mathbf{M}' \qquad [1d]$$

$$\{x\} + \mathbf{M} \rightleftharpoons (\ ) + x\mathbf{M}' \qquad [1e]$$

Equation [1a] represents removal of solute $x$ from water to the gas phase, where {} symbolizes the solvation shell around $x$. Equation [1b] represents collapse of the solvation shell into the cavity vacated by the solute and reconfiguration of solvation-shell water molecules to the bulk water state symbolized by ( ). Equation [1c] is reorganization of the sorbent matrix M to some altered state M′ that may be necessary to accommodate the sorbate; and eqn [1d] represents formation of bonds between sorbate and the altered sorbent.

Matrix reorganization may include displacement of water or other sorbed molecules into solution, or structural changes in the sorbent such as swelling and pore flexing in response to the adsorbate. Little is known about matrix changes and their effects on sorption thermodynamics.

When sorption is from the vapor state, eqns [1a] and [1b] do not apply, and net sorption is the sum of eqns [1c] and [1d]:

$$x_g + \mathbf{M} = x\mathbf{M}' \qquad [1f]$$

The curve defining the equilibrium relationship between sorbed-phase concentration $S$ and fluid-phase concentration $C$ at constant temperature is the sorption isotherm. Various isotherm models have been used depending on the mechanism(s) involved.

It is convenient to define the sorption distribution ratio, $K_d$, as:

$$K_d = \frac{S}{C} \qquad [2]$$

Fluid concentration is expressed in units of pressure or in moles per unit volume. The units of $S$ depend on convention and the implied mechanism being invoked. Ideally, adsorption processes should be based on surface area (or pore volume if condensation is considered) and absorption on solid volume. However, sorbent dry weight is commonly used as a surrogate for these properties, since they are often not accurately known. $K_d$ is concentration-dependent unless the isotherm is linear.

$K_d$ is related to the Gibbs free energy of sorption, $\Delta G^0_{\mathrm{sorp}}$ by:

$$K_d = \exp\left(\frac{\Delta G^0_{\mathrm{sorp}}}{RT}\right) \qquad [3]$$

where $R$ is the gas constant and $T$ is temperature. $\Delta G^0_{\mathrm{sorp}}$ – which is concentration-dependent if $K_d$ is concentration-dependent – encompasses all enthalpy and entropy changes that occur during sorption. This means that the position of equilibrium depends on (1) the difference in the strengths between all sorbate–sorbent and all solute–water interactions, and (2) the entropy changes associated with eqns [1a–1d]. There are virtually no examples where the contributions of these factors have been fully delineated.

Parts of molecules that carry no charge, are not strongly polarized, and lack functional groups that can form strong hydrogen bonds with water are said to be hydrophobic. The hydrophobic effect operates on such molecules when dissolved in aqueous solution and is due to entropy and enthalpy costs associated with the ordered alignment of water molecules near the hydrophobic regions of the solute (apparently, with –OH groups pointed toward the organic), and the consequent disruption of water–water hydrogen bonds. The hydrophobic effect reduces a compound's water solubility and thus contributes to the overall driving force for its sorption from water. For a series of compounds of related structure (e.g., aromatic hydrocarbons), $K_d$ is found to correlate with the molecular hydrophobic surface area or with the $n$-octanol–water partition coefficient ($K_{\mathrm{ow}}$), a commonly used parameter reflecting a compound's net hydrophobic character. This is true whether the sorbent domain is the polar surface of a mineral, or the comparatively apolar surface or interior of organic matter.

## Physisorption of Nonionic Compounds to Mineral Solids

Minerals present mainly two kinds of environments for sorption: (1) the hydroxylated surface found on hydrous oxide minerals and the edges of aluminosilicate clays, and (2) the interlayers of layer silicate clays, which contain siloxane surfaces of variable charge density.

Hydroxylated surfaces have –OH groups protruding into solution from the topmost layer of lattice metal ions. Sorption of apolar compounds from dilute solution to nonmicroporous hydrous oxide surfaces is proportional to mineral surface area and is typically not very sensitive to surface charge, solution pH, and solution ionic strength (although high ionic strength may reduce water solubility and increase the tendency for sorption). It is also proportional to the hydrophobic surface area and to the $K_{\mathrm{ow}}$ of the sorbate, at least for closely related compounds. While data are sparse, sorption to hydroxylated surfaces tends to be endothermic and sorption isotherms tend to be linear to slightly nonlinear.

In the case of apolar compounds, there is good reason to believe that sorption to hydroxylated surfaces occurs by partitioning (i.e., absorption) of the sorbate into the semiordered multilayer of water under the influence of the surface, known as the vicinal water layer. This is depicted as 'B' in **Figure 1a.** Apolar molecules are thought to be unable to penetrate the closest one or two layers of water molecules

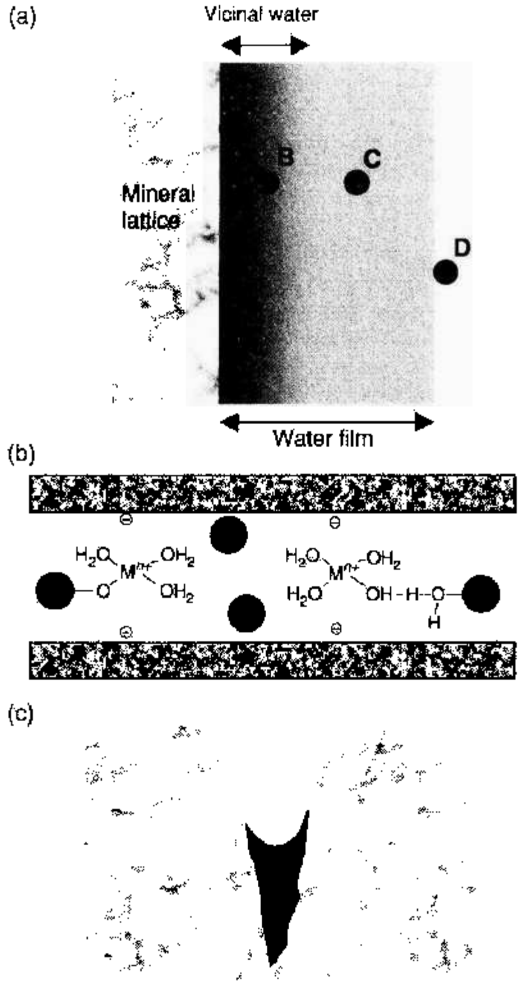


**Figure 1** Depiction of different types of sorption in mineral particles: (a) interaction with discrete sites on the hydroxylated surface, A; partitioning in the vicinal water layer on the hydroxylated surface, B; dissolution in water films, C; and adsorption on the surface of water films, D; (b) sorption in a clay interlayer: left to right, direct coordination to interlayer metal ion; association with the uncharged siloxane surface; and hydrogen-bonding to a metal-coordinated water molecule; (c) condensation in a small pore.

and hence do not actually contact the surface. This is inferred from the observation that sorption of apolar organic vapors is strongly depressed by water vapor, and from the fact that per unit surface area of bare hydrous oxide surface heats of adsorption of organic molecules are much smaller than that of water. As the compound becomes more capable of hydrogen bonding to surface hydroxyl groups, it is better able to compete with strongly bound water molecules for surface sites ('A' in **Figure 1a**).

The siloxane surface existing on the faces of many layer silicate clays consists of oxygen atoms bridging underlying $Si^{IV}$ atoms. The oxygens may bear permanent negative charge, depending on whether they lie near a point of isomorphic substitution in the underlying lattice. The charged sites are closely associated with metal or organic cations, and the surface in the vicinity of the charge is strongly hydrophilic (**Figure 1b**). Polar compounds can hydrogen-bond to metal ion-coordinated water molecules, or can coordinate directly with metal ions if they have an appropriate functional group. The neutral regions between charges are hydrophobic or only weakly hydrophilic; hence, hydrophobic molecules or parts of molecules may compete favorably with water for the bare surface in those regions. Sorption in clay interlayers is complicated by accessibility. The charge density and the exchangeable cation with its hydration sphere control the interlayer spacing and the size of the adsorption 'sites' available for molecules.

Under relatively dry conditions, multilayer adsorption, in which the sorbate condenses as a liquid-like phase on mineral surfaces, may also occur. Due to capillary forces, condensation progresses from the smallest to the largest of pores (or narrowest to widest part of a pore) as the partial pressure increases (**Figure 1c**).

Water has a strong negative effect on vapor adsorption to soils. Sorption of organic vapors decreases by an order of magnitude or more as the soil moisture content increases to approximately the wilting point, which corresponds to an average of 3–5 molecular layers of water (**Figure 2**). This inhibition is due to competition by water molecules for pore volume space and polar surface sites on inorganic surfaces. At moisture contents above the wilting point, sorption approaches what it would be in water-saturated soil.

When dealing with sorption of vapors to unsaturated soils, it is often necessary to take into account the association of organic sorbates with the water located in surface films and pores. This water is technically considered part of the immobile phase. Sorption may occur by a combination of aqueous-phase dissolution ('C' in **Figure 1a**) and vapor-phase adsorption at the gas–water interface ('D' in **Figure 1a**). The latter is especially important for compounds such as alkanes that are highly insoluble in water.

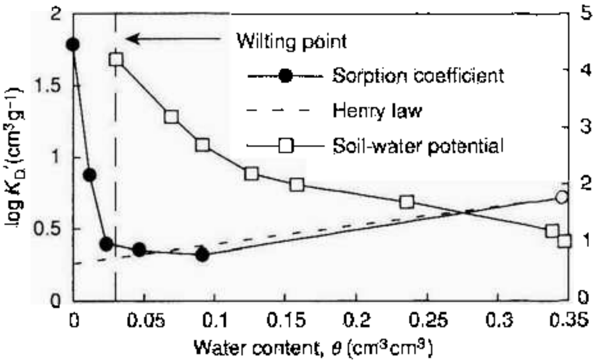


**Figure 2** Toluene sorption coefficient ($K_D'$) and soil-water potential ($\psi$) as functions of volumetric water content in Lundgaard soil (1.1% organic carbon). The Henry law line is the calculated sorption assuming dissolution into water and aqueous–solid partitioning only (i.e., zero vapor–solid sorption). (Reproduced with permission from Poulsen TG, Yamaguchi T, Moldrup P, de Jonge LW, and Rolston DE (2000) Predicting volatile organic vapor sorption from soil specific surface area and texture. *Journal of Environmental Quality* 29: 1642–1649.)

### Physisorption of Nonionic Compounds to Natural Organic Matter

**General considerations** SOM is usually the dominant sorbent of neutral compounds in soils, because it presents a hydrophobic domain for hydrophobic molecules. Sorption isotherms in SOM and SOM-rich soils usually obey the semiempirical Freundlich equation:

$$S = K_F C^N \qquad [4]$$

where $K_F$ and $N$ are constants. The value of $N$ is almost always less than or equal to 1, reflecting a constant ($N = 1$)-to-diminishing ($N < 1$) tendency for sorption with increasing loading. The Freundlich relationship is only approximate and $N$ may vary depending on the range of concentrations over which it is measured. It should be noted that all sorption isotherms tend toward linearity below a certain concentration, a result the Freundlich equation fails to predict.

Many researchers have assumed $N$ is 1, or have taken $N$ to be 1 for purposes of simplifying the sorption term in transport or bioavailability models. In that case, the isotherm is given simply by:

$$S = K_p C \qquad [5]$$

where $K_p$ is referred to as the partition coefficient.

On the assumption that sorption occurs exclusively to the organic matter component of the soil, the distribution, Freundlich, or partition coefficient is often normalized to the fraction of organic carbon ($f_{oc}$) in the sample; i.e.:

$$K_{oc} = K_d(\text{or } K_F \text{ or } K_p) \cdot f_{oc}^{-1} \qquad [6]$$

Linear free energy relationships between $K_{oc}$ and parameters that represent hydrophobic properties such as $K_{ow}$ or saturated water solubility ($C_{sat}$) have been extensively tested, e.g.:

$$\log K_{oc} = a + b\log K_{ow} \qquad [7]$$

$$\log K_{oc} = c - d\log C_{sat} \qquad [8]$$

where *a* and *b* or *c* and *d* are fitting parameters. Such relationships work best for a set of closely related compounds in a given soil; the reason being that the enthalpy and entropy of eqns [1a–1d] depend on sorbate molecular structure, and because soil organic matter is not uniform from place to place. Other structure–property relationships have been studied based on, for example, parachor, connectivity index, or UNIFAC group contribution method. Solution properties such as pH and ionic strength have modest – and often contradictory – effects on sorption of nonionic compounds to SOM.

**Natural organic materials in soils** A discussion of sorption to soil natural organic matter must take into account the nature and abundance of all the different types of refractory components that make it up. Ordinarily the most abundant component is humic substances. Humic substances are highly altered, refractory macromolecules derived from biomass. They may also contain recognizable fragments of carbohydrates and proteins covalently attached to, or sequestered in, the humic solid. Humic substances are operationally defined as consisting of fulvic acid (soluble at all pH), humic acid (soluble in alkaline solutions), and humin (insoluble at all pH and typically strongly associated with inorganic particles). The soil water may also contain 'dissolved' humic substances – ranging from small, truly dissolved molecules to colloidal particles – that can sorb organic compounds, although somewhat less effectively than particulate humic substances. Kerogens are biogenic, macromolecular substances that have undergone low-temperature chemical and physical diagenesis over geologic time. Kerogens are abundant in soft coals and shale and may be found in soils as a result of reworking processes. Hard coals are biogenic macromolecular substances that have been subjected to high-temperature and -pressure alteration (catagenesis) over geologic time and are found in some soils. Black carbon is the carbonaceous by-product of biomass burning (charcoal) or fossil fuel combustion (soot). Black carbon may constitute a minute-to-small fraction of total organic C in soil.

**Physisorption to humic substances and kerogen** Humic substances and kerogens are considered to be random-network macromolecular solids; that is, they are composed of macromolecules of heterogeneous structure that are randomly aggregated to form a three-dimensional phase. This phase is hydrated to a degree depending on the type of humic substance and relative humidity, and may contain metal ions coordinated to charged groups, such as carboxylate. From fulvic acid to humic acid to humin to kerogen, there is a trend of increasing molecular weight, apolarity, C content, fused ring size, density, cross-linking, and glass-to-rubber transition temperature.

For such materials, an absorption rather than an adsorption model is better supported by the evidence and is widely accepted. External surface areas are too low – approx. $10^0\,m^2\,g^{-1}$ for humic acid particles based on $N_2$ as the probe – to account for their high uptake of organic chemicals. Furthermore, their surfaces are not hard like a mineral crystal, but diffuse and flexible due to solvation and charge repulsion of polar interfacial groups. Organic molecules reaching the interface pass through and enter interior regions, where they intermingle with the strands of the macromolecules. A similar paradigm for small-molecule sorption by synthetic organic polymers is well established. Sorption to humic substances appears to be slightly exothermic, although few data are available.

According to established polymer theory, amorphous macromolecular solids may exist in 'glassy' or 'rubbery' states, depending on temperature. These states differ in physical properties and sorbent behavior. The rubbery state is more expanded and its macromolecular segments have greater mobility. Sorption in rubbery solids is similar to liquid dissolution in that specific sorption sites have fleeting existence and sorbed molecules experience an average chemical environment. This results in a linear isotherm, like eqn [5], and noncompetitive sorption behavior when multiple solutes are present.

The glassy state is denser and more rigid. Also, because the glassy state is incompletely relaxed, it contains internal voids not filled by macromolecules. These voids (or 'holes') are of molecular dimensions and may accumulate sorbate. Hence, sorption in the glassy state exhibits dual-mode behavior. One mode is 'dissolution' in the bulk solid that is linear and noncompetitive, analogous to sorption in the rubbery state. The other mode is 'hole-filling' in the voids that is best modeled by a Langmuir-type equation. The dual-mode isotherm is given by:

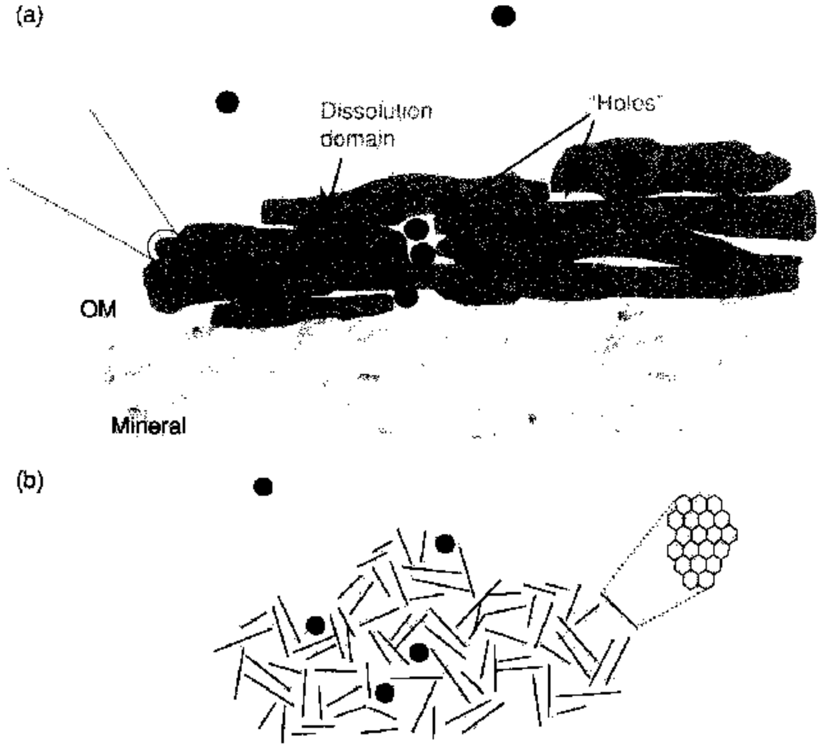


**Figure 3** Sorption in soil organic matter (OM) fractions: (a) Macromolecular solids such as humic substances and kerogens: sorbate may undergo solid-phase dissolution in domains composed of macromolecules in a relatively fluid state and hole-filling in voids within the organic phase and between the organic phase and the mineral surface; (b) micrographitic substances such as black carbon: sorbate may adsorb in shallow micropores and some may penetrate into deep micropores between polyaromatic platelets.

$$S = K_D C + \frac{S_H^{max} K_H C}{1 + K_H C} \qquad [9]$$

where $K_D$ is the dissolution domain partition coefficient, and $S_H^{max}$ and $K_H$ are the hole domain capacity and affinity parameters, respectively. Equation [9] predicts nonlinear and competitive sorption, both of which have been realized for many soils, humic acid reconstituted in particulate form, and kerogens.

In the context of SOM, holes may be regarded as semipermanent voids within the folds of individual macromolecules, between macromolecules, or between macromolecules and the mineral surface to which they are anchored. This is shown in Figure 3a. When eqn [9] is applied to heterogeneous natural materials: (a) the $K_D$ term represents the sum of partition coefficients for all rubbery substances present and the dissolution domains of all glassy substances present; and (b) the Langmuir term represents a weighted average of holes of different sorption potential that may exist. The affinity for a given solute is greater in holes than in the dissolution domain.

As sorbate loading increases, the isotherm can become more complex. The glassy state may become softened ('plasticized') by the sorbate, causing it to become rubbery and the holes to disappear. At high loadings the isotherm in a rubbery solid begins to curve upward with respect to the solution concentration axis owing to the effects of swelling.

Natural organic matter in many samples appears to have glassy character. The glass transition temperature of some natural substances has been measured, including wet and dry humic acids (43–72°C), coals (>300°C), and wet lignin (60–90°C). Lignin is an amorphous, random polymer from the woody parts of plants that is regarded to be a major precursor of terrestrial humic substances. Whole-soil SOM may exhibit multiple glass transitions due to its heterogeneous nature. Synthetic polymers, humic acids, and coals have all been shown to have high internal surface area associated with the holes.

Nonlinearity and competitive effects tend to increase in the expected order of increasing glassy character of the solid:

$$\text{dissolved humic acid} < \text{particulate humic acid} < \text{whole-soil SOM} < \text{humin} < \text{kerogen}$$

Artificially cross-linking humic acid by flocculating it with polyvalent metal ions (e.g., $Al^{3+}$) increases nonlinearity and competition.

Figure 4 shows an example of sorption of a hydrophobic organic compound phenanthrene in SOM. It is evident that phenanthrene sorption is nonlinear with $N = 0.772$. It can also be seen that the dual-mode

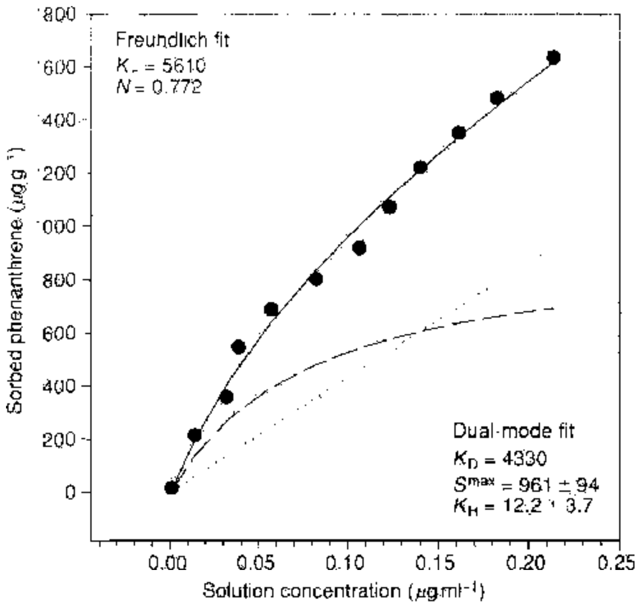


**Figure 4** Phenanthrene sorption in soil with a high organic content. The fitting parameters of both Freundlich and dual-mode equations are shown. The solid line is the dual-mode fit, assuming only one type of Langmuir (hole) sites, and solid circles are actual data points; dotted and dashed lines are the contribution from partitioning and Langmuir (hole)-type sorption, respectively. The maximum surface sorption capacity calculated from surface area and molecular size is 286 $\mu g\, g^{-1}$. See text for details.

model fits the data well. Langmuir (hole) type of sorption dominates the overall uptake till the solution concentration reaches 0.14 $\mu g\, ml^{-1}$, then partitioning contributes more. The overall dissolution coefficient is 4330 $ml\, g^{-1}$ and Langmuir (hole) sorption capacity 961 $\mu g\, g^{-1}$. If the surface of this SOM sample (0.88 $m^2\, g^{-1}$) were rigid as an inorganic mineral, the maximal (external) surface coverage would be 286 $\mu g\, g^{-1}$ calculating from the molecular volume of phenanthrene as 182 $cm^3\, mol^{-1}$. Clearly, phenanthrene molecules have diffused into and sorbed within the SOM matrix, in contrast to surface adsorption.

SOM is heterogeneous and polydisperse in nature, and its characteristics can vary substantially as a function of age, source, diagenesis, and mode of transformation. Variations in SOM such as polarity and hydrophobicity strongly affect $K_{oc}$, up to several orders of magnitude. This is one reason why many different $K_{oc}$–$K_{ow}$ free energy relationships are reported in the literature.

Aromatic and aliphatic moieties are the main structural components of SOM. Many investigations report positive correlations between the amount of HOC sorption and percentage of aromatic C of SOM. In addition, condensed aromatic domains are suggested to be the locations in SOM samples and isolated humic materials where nonlinear and competitive sorption occurs. Generally speaking, diagenetically altered organic matter samples are enriched in aromatic groups and show a higher degree of isotherm nonlinearity and competitive sorption, and a larger sorption capacity than geologically young samples. Organic matter in shale compared with recently deposited peat samples is an example in case. Therefore, HOC sorption is thought to be dominated by aromatic components of SOM rather than aliphatic components.

Recent research indicates that aliphatic components of SOM, particularly the paraffinic groups, can significantly sorb HOC. One example is that some soil humin samples often have a higher HOC sorption capacity than the humic acids extracted from the same soils, but these humic acids have higher aromatic content than the soil humins. It has been suggested that HOC is partitioned into the flexible, paraffinic carbon moieties to explain the high HOC sorption capacity. However, sorption of HOC by many soil humin samples is reported to be highly nonlinear and competitive, inconsistent with partitioning as the only sorption mechanism.

To account for both high sorption capacity and nonlinearity in soil humins, a new substructural conformation is proposed for the paraffinic carbons (**Figure 5**). Crystalline and amorphous aliphatic domains can coexist in SOM, which is supported by solid-state nuclear magnetic resonance (NMR) studies. A proposed conformation of semicrystalline polymer interphase is shown in **Figure 5a**. Sorption by crystalline regions is negligible due to their crystalline nature (inability for HOC molecules to diffuse). Upon interaction of the crystalline–amorphous complex with soil mineral surfaces, the first several molecular layers of the amorphous region may rearrange to take a more condensed form (**Figure 5b**). This condensed region can cause enhanced nonlinear sorption and both condensed and expanded amorphous aliphatic regions contribute to high HOC sorption. This proposed configuration is further supported by the experimental data reported in the early literature, where isotherm nonlinearity and competitive sorption of HOC by humic acids adsorbed on minerals are greater than unsorbed humic acids.

**Physisorption to black carbon** Black carbons are carbonaceous materials composed of single and stacked polyaromatic sheets in a highly disordered arrangement. The sheets may be functionalized along the edges with hydroxyl, carboxyl, keto, or other groups. The pore structure of black carbon depends on the source material and conditions of synthesis. From gas adsorption measurements it has high surface

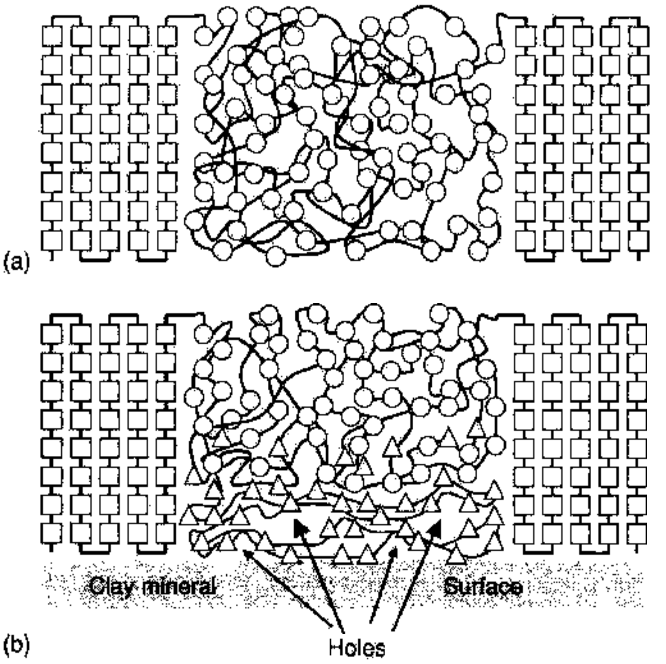


**Figure 5** Schematic configuration showing the distribution of crystalline (squares) and amorphous (circles) aliphatic organic carbons (a). Interaction of the aliphatic carbons with a clay mineral surface rearranges part of the amorphous domain into a condensed configuration (triangles) having nanometer-sized holes (b). These holes are the locus of Langmuir adsorption sites, responsible for nonlinear and competitive sorption. (Reproduced with permission from Gunasekara AS and Xing B (2003) Sorption and desorption of naphthalene by soil organic matter: importance of aromatic and aliphatic components. *Journal of Environmental Quality* 32: 240–246.)

area (typically, more than $50\,m^2\,g^{-1}$), high porosity ($0.1–0.2\,cm^3\,g^{-1}$), and small pore size – i.e., below approximately 20 Å, with much porosity in the range 4–10 Å. Recent evidence shows that small apolar molecules can penetrate the interior of black carbon and cause swelling. Sorption isotherms of apolar compounds on prepared black carbons tend not to fit Freundlich or Langmuir models very well. For example, the isotherm of benzene in a wood charcoal is reverse S-shaped, with a linear region from $10^{-7}$ to $10^{-4}$ times its water solubility. While heats of sorption for black carbons have not been reported, sorption of apolar compounds from water to related carbonaceous sorbents such as graphite and activated carbon is exothermic.

A model that takes into account the existence of both macromolecular substances (humics plus kerogen) and black carbons in soils is clearly desirable. Models have been proposed of the form:

$$S = r_{HS} \cdot K_p C + r_{BC} \cdot f_{BC}(C) \qquad [10]$$

where $r_{HS}$ and $r_{BC}$ are the fractions of total organic carbon as humic substance and black carbon, respectively; $S$, $K_p$, and $C$ are as defined above; and $f_{BC}(C)$ is a nonlinear function describing sorption to black carbon. In one example, $f_{BC}(C)$ is a Freundlich term like that in eqn [4]. In another, $f_{BC}(C)$ is a term that represents the potential to condense into liquid-like or solid-like form in pores that are assumed fixed in size and shape and similar to those that exist in synthesized activated carbons. Such models assume linear sorption in humic substances, which is not always true. Multicomponent models will be improved by advances in our understanding of the true nature of sorption to black carbon.

## Kinetics of Physisorption

The rates of uptake and release from soil particles often control the transport and bioavailability of chemicals in soil systems. In well-mixed soil-water suspensions it is found experimentally that a large fraction of the sorbate is taken up or released within a few hours, but days, weeks, or even months may be required to reach complete equilibrium, depending on compound and soil. Moreover, exhaustive desorption at ordinary temperatures to air or water may reveal a minor fraction that seems permanently bound, yet is recoverable by solvent extraction or other vigorous means.

In general, soils contain structures that can hinder the diffusion of molecules to and from all available sites. These structures include pore networks in the interstices of particle aggregates as well as organic matter, which can be considered a highly viscous absorptive phase. Models based on pore diffusion and intraorganic matter diffusion being rate-limiting to sorption have been presented in the literature. Because most organic compounds have such a high affinity for the organic matter, it is tempting to regard intraorganic matter diffusion as rate-limiting unless the fraction of organic matter is very low. Generally, intraparticle sorption or desorption rate coefficients decrease with molecular size (or $K_{ow}$ as a surrogate for size) and also appear to decrease with the introduction of hydrogen-bonding functional groups. This finding alone, however, does not distinguish pore diffusion and intraorganic matter diffusion. Fractions that become highly resistant to desorption have been generated in purely inorganic as well as purely organic reference materials. (Sorption or desorption in clay interlayers appears to be rapid, however, and does not generate resistant fractions.) Diffusion rates in soils are frequently independent of nominal particle radius, implying that the controlling diffusion length-scale is much smaller, perhaps as small as 10–100 nm.

Diffusion is highly dependent on the nature and geometry of the diffusing medium and the boundary conditions. Classic diffusion laws are difficult to apply to soils owing to their heterogeneous nature in both composition and particle geometries. From studies on homogeneous inorganic reference materials, it is known that diffusion in fixed-pore networks containing mesopores (widths of 2–50 nm) or macropores (greater than 50 nm) is slowed by sorption to pore walls and by tortuosity. Tortuosity includes deviation from straight-line paths and pore noninterconnectedness. Diffusion in fixed micropore (less than 2 nm) networks, where pore openings are not much larger than the diffusant diameter, is additionally slowed by steric effects.

Intraorganic matter diffusion behavior depends on whether the solid is in the rubbery or glassy state. Diffusion in rubbery solids involves a more-or-less cooperative interchange between diffusant and polymer strands. The ideal rubbery sorbent is characterized by linear sorption isotherms, concentration-independent diffusion coefficients, and curves describing fractional mass uptake and release with time that are coincident. Diffusion in glassy sorbents generally shows the opposite. Molecules traversing the glassy solid experience a more condensed matrix and they encounter holes along the way where the diffusant can be momentarily detained. For these reasons, diffusion in glassy solids may be several orders of magnitude slower than in rubbery solids. The apparent diffusion coefficient in glassy solids is concentration-dependent because the availability of holes is concentration-dependent, and because the solid can undergo transition to a rubbery state at high concentrations. Diffusion coefficients of toluene, *n*-hexane, and acetone in pressed humic acid disks are in the range $10^{-8}$–$10^{-9}\,\mathrm{cm^2\,s^{-1}}$, about the same as in rubbery polymers and 3–4 orders of magnitude smaller than those in water. There is evidence to suggest that the diffusivity of a given compound follows the order: humic acid > humin > kerogen/coal. Diffusion within black carbons has received little attention so far.

To avoid difficulties inherent in the application of diffusion models to soils, many people have employed nonmechanistic rate models based on simpler mathematics. Recent ones include the two-site, multicompartment, and stochastic models. Although these models have met with some success, they usually suffer from an inability to translate to new systems. The two-site model, which is often used in soil-column elution studies, partitions the sorbate into instantaneous and kinetic compartments. The kinetic compartment is described by first-order terms for the forward and reverse steps. Overall the rate law has three parameters: fraction in the kinetic compartment, bulk solid equilibrium sorption coefficient, and rate constant for desorption from the kinetic compartment. The multicompartment model is popular in batch experiments in which molecules are instantly removed as they desorb by a stream of gas or by a polymeric sorbent included in the suspension. This model assumes that desorption occurs concurrently from 'fast,' 'slow,' and sometimes 'very slow' compartments, each described by a first-order term having two parameters: the rate constant and the fraction of chemical in that compartment. Stochastic models assume that the rate can be described with an array of equilibrium sorption coefficients and/or first-order rate constants that are distributed continuously according to a probability density function. Such models have three or two parameters, respectively.

## Physisorption Hysteresis

Hysteresis in the sorption isotherm exists when the curve constructed by allowing the solid to take up the compound (the sorption branch) is different from the curve constructed in the reverse manner, by diluting the fluid and reequilibrating (the desorption branch). Typically, the desorption branch is shifted upward of the sorption branch, and may appear to extrapolate to a nonzero sorbed concentration at zero solution-phase concentration. Hysteresis has been observed in many studies. An example is given in Figure 6. Hysteresis poses a challenge to the elucidation of sorption mechanism and to using sorption models predictively.

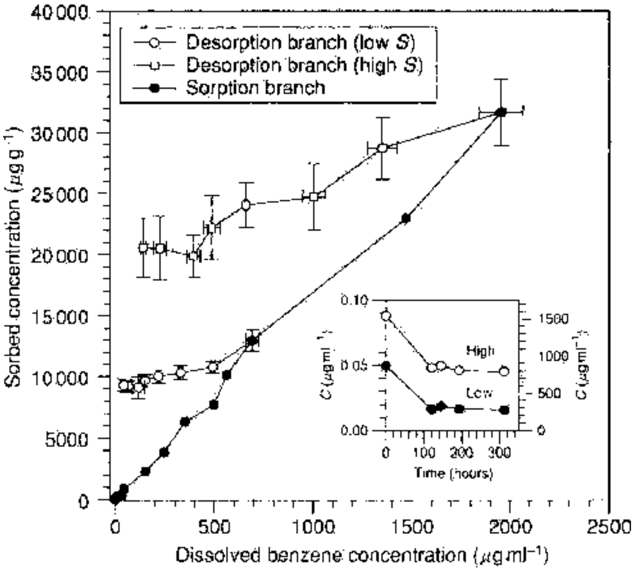


**Figure 6** Hysteresis of benzene in a reference soil (Pahokee) from the International Humic Substances Society. Sorption was carried out in one step for 14 days. Desorption was carried out in consecutive steps (triplicate samples), each step for 14 days. Inset shows uptake rate curves (change in solution concentration) at two concentrations, indicating apparent equilibrium after approximately 4 days.

Hysteresis may be an artifact of the experimental technique. When sorbed concentrations are determined indirectly based on the amount taken up from solution, as is often the practice, artificial hysteresis may result from chemical or biological reactions or from dilution of competing substances such as cosolutes or colloidal sorbents in the desorption step. Failure to allow enough time for diffusion to all sites can give artificial hysteresis regardless of whether sorbed concentration is determined directly or indirectly. Nevertheless, in many cases hysteresis is observed even when those artifacts have been eliminated.

True hysteresis (also referred to as irreversible sorption) is caused by spontaneous mechanisms associated with irreversible pore-filling and -emptying. Since it would violate the laws of classical thermodynamics for both isotherm branches simultaneously to represent the reversible state, true hysteresis always indicates a nonequilibrium condition. Proposed explanations for true hysteresis fall into three categories: formation of metastable states of the sorbate, network percolation effects, and irreversible pore deformation. The first two are associated with capillary condensation of vapors in mesoporous solids. During uptake in a pore there can be a delay in the transition from the metastable multilayer film to the thermodynamic condensed plug state; evaporation from the meniscus of the condensate is less favorable than from the film surface. Hysteresis due to network percolation effects may arise in complex systems of pores, and it results when desorption is delayed because condensate in some pores does not have a continuous vapor channel to the surface. Capillary condensation hysteresis applies primarily to fixed-pore mesoporous materials.

Irreversible pore deformation has been identified in sorption of organic vapors or solutes by organic sorbents, including dry smectite clays, organosmectite clays, glassy polymers, and (recently) natural organic matter in soils and prepared charcoals. By this mechanism, the incoming molecule causes dilation of the pore due to the flexibility of the organic matrix, and the pore does not instantly relax when the molecule leaves. When this happens it may be surmised that some of the free energy of sorption goes into expanding the pores and the sorbate leaves from pores existing in a different thermodynamic state than the ones they entered. When desorption abruptly occurs, it is possible that some molecules get trapped in the collapsing organic matrix.

## Ion-Exchange Sorption

Permanently charged and ionizable organic molecules in their ionic form can undergo 'ion-exchange' reactions with charged sites in soil particles from which existing ions are displaced. Cation exchange is far more common than anion exchange due to the prevalence of negative charges in soil particles. Examples of permanently charged compounds that can undergo cation exchange include tetraalkylammonium surfactants and cationic herbicides such as paraquat and diquat. Alkylamines, aromatic amines, and *N*-heterocyclic aromatic compounds are examples of ionizable compounds that can sorb via ion exchange, depending on the degree of their protonation, which, in turn, depends on soil pH.

Typically, cation exchange is nonlinear and subject to competition by other cations present. According to diffuse double-layer theory, cation exchange on mineral surfaces occurs by an accumulation of organic ions in the vicinal water film near the surface due to electrostatic force between the ion and the oppositely charged surface. However, sorption of organocations often exceeds the cation exchange capacity (CEC) of the solid in proportion to the apolar content of the molecule, such as alkyl chain length or aromatic ring content. This may be rationalized by a contribution to overall sorption from the hydrophobic effect operating on the apolar parts of the ion. In order to maintain electrical neutrality, organocations in excess of the CEC must carry along a counterion (e.g., $Cl^-$) to the interfacial region. Thus, sorption of organocations is best represented as the sum of true ion exchange and sorption of the ion pair:

$$R^+ + M^+ \cdots \equiv \rightleftharpoons M^+ + R^+ \cdots \equiv \qquad [11]$$

$$R^+ + X^- + \equiv \rightleftharpoons R^+X^- \cdots \equiv \qquad [12]$$

where $R^+$ represents the organocation, $X^-$ a counter ion, $M^+$ a competing cation, and $\equiv$ the surface.

Ion-exchange reactions in SOM also occur. Organocations interact electrostatically with dissociated carboxyl or phenoxyl groups on humic molecules; hence, ion exchange in SOM is strongly pH-dependent. Ion exchange in SOM may be facilitated by physisorptive interactions of the noncharged parts of the molecule with humic groups other than the negative charge itself. Just like ion exchange on minerals, the hydrophobic effect may be expected to play some role in the sorption of organocations. Sorption of organoanions to SOM is typically weak, since positively charged sites on SOM are rare and the solid is negatively charged overall at ordinary pH. However, some organoanions – e.g., alkylbenzene sulfonate surfactants and the anion of pentachlorophenol – appear to be capable of sorbing to SOM, possibly as their ion pairs with $Na^+$ or other cations.

## Chemisorption

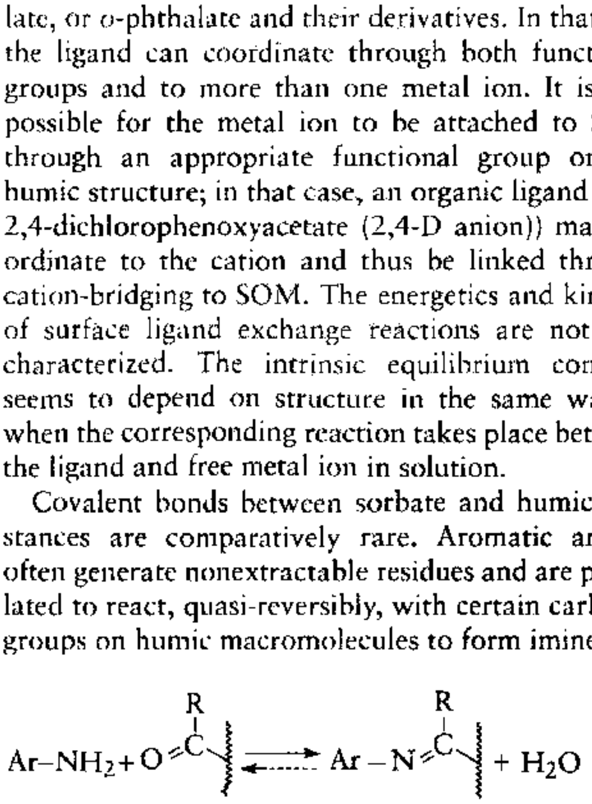
Chemisorption entails appreciable overlap of molecular orbitals between sorbate and sorbent, such as ligand–metal coordination bonds and covalent bonds. Molecules capable of chemisorption may also be capable of physisorption.

Ligand–metal bonds may be formed between carboxylate, phenolate, thiolate, sulfonate, or phosphate groups on the solute and metal ions – especially $Fe^{3+}$ and $Al^{3+}$ – on the surface of minerals, usually in exchange for hydroxyl or water ligands:

$$\equiv \mathrm{M{-}OH(-OH_2)} + \mathrm{L} \rightleftharpoons \equiv \mathrm{M{-}L} + \mathrm{OH^-(H_2O)} \qquad [13]$$

where ≡ represents the surface, M a metal ion, and L a ligand. Equation [13] is more favorable when L is bidentate; for example, salicylate, oxalate, catecholate, or o-phthalate and their derivatives. In that case the ligand can coordinate through both functional groups and to more than one metal ion. It is also possible for the metal ion to be attached to SOM through an appropriate functional group on the humic structure; in that case, an organic ligand (e.g., 2,4-dichlorophenoxyacetate (2,4-D anion)) may coordinate to the cation and thus be linked through cation-bridging to SOM. The energetics and kinetics of surface ligand exchange reactions are not well characterized. The intrinsic equilibrium constant seems to depend on structure in the same way as when the corresponding reaction takes place between the ligand and free metal ion in solution.

Covalent bonds between sorbate and humic substances are comparatively rare. Aromatic amines often generate nonextractable residues and are postulated to react, quasi-reversibly, with certain carbonyl groups on humic macromolecules to form imines:



$$\mathrm{Ar{-}NH_2} + \mathrm{O{=}C(R)} \rightleftharpoons \mathrm{Ar{-}N{=}C(R)} + \mathrm{H_2O} \qquad [14]$$

The reactive carbonyl may be present in ketone, aldehyde, or quinone functional groups. The rate and extent of imine formation are currently unpredictable.

Other cases are known in which the organic solute is converted to reactive forms that then couple covalently with humic substances. Concern about potential remobilization of these so-called bound residues has spawned an active area of research. Coupling reactions are technically outside the definition of sorption, since the molecules undergoing coupling are transformed species, not the original ones, and since they are seldom reversible in the sense that the original chemical is liberated. Examples are found among pesticides, explosive compounds, phenols, aromatic amines, and polycyclic aromatic hydrocarbons (PAHs). Typically, the initial reaction is an enzyme- or mineral surface-catalyzed oxidation reaction, although reduction reactions have been reported. Rarely has the coupling reaction itself been assigned with certainty to a biotic or an abiotic origin, much less a specific mechanism. In most cases coupling is inferred by the inability to recover by solvent extraction a portion of an isotope when the compound is added in isotopically labeled form. Oxidative coupling of some compounds – for example the fungicide cyprodinil and the herbicide 2,4-dichlorophenoxyacetic acid – has been verified by nuclear magnetic resonance spectroscopy. In another example, trinitrotoluene was microbially reduced to amino toluenes that subsequently coupled to organic matter, most likely abiotically via eqn [14]. It has been proposed that aromatic amines are oxidized by manganese oxide minerals in soils to give free radicals that subsequently couple with humic substances.

*See also:* **Hydrocarbons**; **Organic Matter**: Interactions with Metals; **Pesticides**

## Further Reading

Alexander M (1995) How toxic are toxic chemicals in soil? *Environmental Science and Technology* 29: 2713–2717.

Cheng HH (1990) *Pesticides in the Soil Environments: Processes, Impacts, and Modeling*. Madison, WI: SSSA.

Chiou CT (2002) *Partition and Adsorption of Organic Contaminants in Environmental Systems*. Englewood Cliffs, NJ: John Wiley.

Dec J and Bollag J-M (1997) Determination of covalent and noncovalent binding interactions between xenobiotic chemicals and soil. *Soil Science* 162: 858–874.

Gunasekara AS and Xing B (2003) Sorption and desorption of naphthalene by soil organic matter: importance of aromatic and aliphatic components. *Journal of Environmental Quality* 32: 240–246.

Israelachvili JN (1992) *Intermolecular and Surface Forces*, 2nd edn. London, UK: Academic Press.

Naidja A, Huang PM, and Bollag J-M (2000) Enzyme–clay interaction and their impact on transformations of natural and anthropogenic organic compounds in soil. *Journal of Environmental Quality* 29: 677–691.

Pignatello JJ (2000) The measurement and interpretation of sorption and desorption rates for organic compounds in soil media. *Advances in Agronomy* 69: 1–73.

Poulsen TG, Yamaguchi T, Moldrup P, de Jonge LW, and Rolston DE (2000) Predicting volatile organic vapor sorption from soil specific surface area and texture. *Journal of Environmental Quality* 29: 1642–1649.

Rouquerol F, Rouquerol J, and Sing K (1999) *Adsorption by Powders and Porous Solids*. San Diego, CA: Academic Press.

Schwarzenbach RP, Gschwend PM, and Imboden DM (2002) *Environmental Organic Chemistry*, 2nd edn. Englewood Cliffs, NJ: John Wiley.

Senesi N (1992) Binding mechanisms of pesticides to soil humic substances. *Science of the Total Environment* 123: 63–76.

Xing B and Pignatello JJ (1997) Dual-mode sorption of low-polarity compounds in glassy polyvinychloride and soil organic matter. *Environmental Science and Technology* 31: 792–799.

# Oxyanions

**C P Schulthess**, University of Connecticut, Storrs, CT, USA
**H Wijnja**, Connecticut Agricultural Experiment Station, New Haven, CT, USA
**W Yang**, University of Connecticut, Storrs, CT, USA

## Introduction

An oxyanion is a net negatively charged compound consisting of a cation at its center surrounded by oxygens. There are many different oxyanions present in soils. The most common ones of interest to soils include phosphates ($PO_4^{3-}$), arsenite ($AsO_3^{3-}$), arsenate ($AsO_4^{3-}$), selenite ($SeO_3^{2-}$), selenate ($SeO_4^{2-}$), chromate ($CrO_4^{2-}$), molybdate ($MoO_4^{2-}$), carbonate ($CO_3^{2-}$), sulfate ($SO_4^{2-}$), nitrate ($NO_3^-$) and perchlorate ($ClO_4^-$). Oxyanions are prevalent in nature, and most of these are essential nutrients for plants. Most all of them, however, are also toxic to plants and the environment when their concentrations rise above a certain critical level. For example, excessive amounts of some oxyanions, namely nitrates and phosphates, are directly responsible for the eutrification of our ponds and lakes.

There are, of course, numerous other oxyanions that are discussed in the environmental literature, particularly in connection to industrial and nuclear waste disposal, or spills and remediation. Industrial oxyanion wastes include many of those already mentioned above. Radioactive waste and environmental studies may include oxyanions such as pertechnetate ($TcO_4^-$). Since no stable isotopes of Tc exist (which are prepared in very small amounts in cyclotrons), and since it is rather dangerous to work with this radionuclide, a surrogate oxyanion is typically studied instead, namely perrhenate ($ReO_4^-$).

The fate of oxyanions in soils in the environment is strongly dependent on their sorption to soil mineral components. That is, an oxyanion that is adsorbed to an immobile solid phase will not be transported to a new location, because movement of oxyanions in the environment is through the liquid phase. Retention by solids results in a longer availability of oxyanions in a defined area. For example, leaching from the soil's upper horizons down toward groundwater does not occur as long as it is sorbed to the immobile solid phase. Lateral liquid movement to nearby streams is also minimized, as is its subsequent transport to larger bodies of water.

All this, of course, assumes that the solid phase that the oxyanion is sorbed to is itself immobile. Keep in mind, however, that various solids (such as colloidal-sized particles) are very small and can be translocated by the moving water. This translocation of mobile solids along with the ions sorbed on them can be either vertically through soil horizons, or horizontally with the groundwater flow, or as soil erosion moving horizontally over the soil surface.

An understanding of sorption of oxyanions to solid soil components is very important for an evaluation of the quality of the environment. If the oxyanions are retained weakly, then the moving water will easily remove them from the soil. If the retention of oxyanions by the immobile solid-phase is moderate in strength, then the moving water will remove the oxyanions from the area slowly. Conversely, if the oxyanions are retained strongly by immobile solids, then the moving water will remove them from the area very, very slowly. If the solid phases are small enough to be mobilized by the moving water, then attention to the turbidity of the liquid phase or the presence of surface soil erosion must be included in any evaluation of the environmental quality present. Once again, this is because the mobile solids typically will have sorbed contaminants on them.

## Techniques for Detection and Quantification of Oxyanions in Solution

Quantitative studies of the adsorption of oxyanions on soils and minerals require determination of the oxyanion concentration remaining in the aqueous phase. That is, the amount adsorbed equals the amount of anion added to the solid-liquid suspension minus the amount remaining in the liquid phase at equilibria. Many different instruments can be used for the liquid-phase analysis, such as ion-selective electrodes, colorimetric methods, ion chromatography, atomic absorption, and inductively coupled plasma-atomic emission spectrometry. With some of these methods (namely, the first three mentioned), one must be particularly cautious to avoid problems attributed to speciation and complexation interferences. These complications can usually be avoided by measuring the liquid phase concentration with

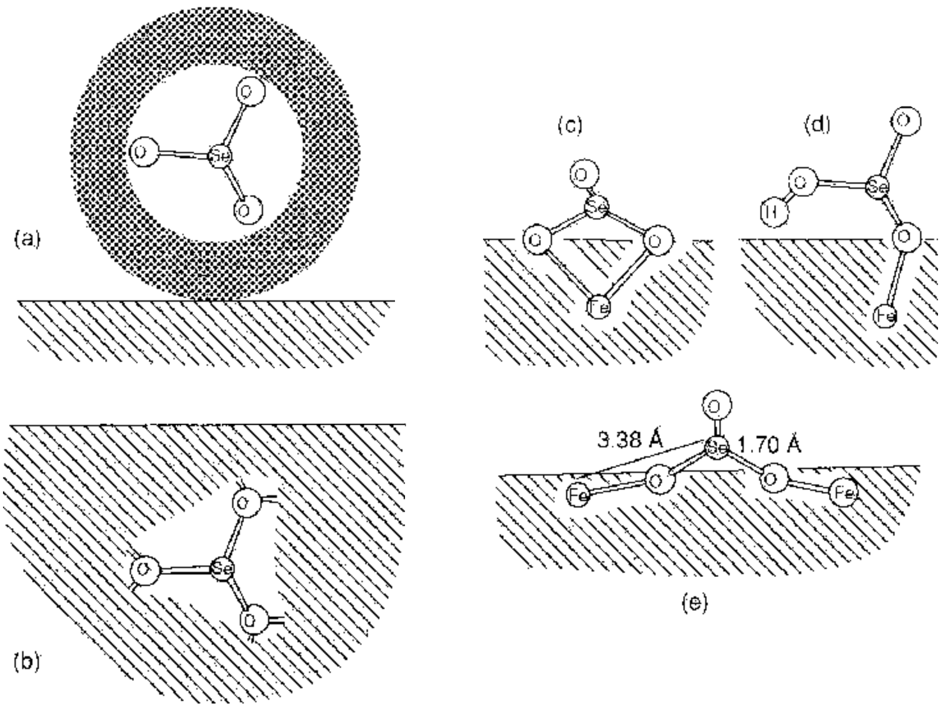


**Figure 1** Schematic of several of the possible geometric arrangements of an oxyanion at the solid–water interface. Selenite is the adsorbate oxyanion and an iron oxide the adsorbent mineral. The area above each surface (indicated by solid horizontal lines) is water, while the solid mineral phases are below the solid horizontal lines. (Reprinted from Brown GE Jr, Parks GA, and Chisholm-Brause CJ (1989) *In-situ* X-ray absorption spectroscopy studies of ions at oxide–water interfaces. *Chimia* 43: 248–256.)

a high background ionic strength and at a specified pH condition. One does not generally quantify the amount of oxyanions adsorbed by direct measurement of the solid-phase concentration. This is because many of the solid-phase measuring methods require expensive instruments and/or the quantitative precision is not satisfactory.

## Techniques for Investigation of Oxyanions Adsorbed at Mineral–Water Interfaces

Various surface-sensitive techniques have been used to obtain structural information of surfaces and surface species that are needed to identify adsorption mechanisms and reactions. An important aspect of the application of spectroscopic techniques to the investigation of adsorbed species is the experimental conditions under which the spectra of the sample are collected. Invasive or *ex situ* methods commonly require extensive sample dessication and vacuum, and therefore have the potential for modifying or destroying sorption complexes normally present at the solid–liquid interface in an aqueous environment. Conversely, noninvasive or *in situ* methods require little or no alteration of a sample from its natural state. Clearly, in order to resolve the structure of surface species present at a mineral–water interface of natural particles, *in situ* spectroscopy techniques are needed that can be applied in the presence of liquid water. Three examples of *in situ* techniques that have been used for the investigation of oxyanions on mineral–water interfaces are attenuated total reflectance–Fourier-transformed infrared (ATR-FTIR), Raman, and X-ray absorption spectroscopy (XAS). A very brief overview of these is given below.

The predominant application of spectroscopic analysis of adsorbed oxyanions at the solid–liquid interface is to understand the symmetry and binding mechanism of the anion on the surface. Quantitative measurements ("How many ions are adsorbed?") is not the primary focus, because, as was noted above, this is much more easily determined based on the mass balance of the elements added and recovered from the liquid phase. Spectroscopy is the only method, however, that can determine how the ions are held on the surface. **Figure 1** shows five commonly encountered possibilities. Using selenite as the example oxyanion, it can be held as an outer-sphere (**Figure 1a**), where at least one molecule of water exists between the anion and the surface. It may be embedded in the internal regions of the mineral (**Figure 1b**) due to solid diffusion of the anion into the mineral or due to some coprecipitation of the anion with a cation onto the surface of the mineral. **Figure 1c–e** shows the anion held as an inner sphere, where the anion directly bonds with the mineral surface. An inner-sphere anion can be monodentate (**Figure 1d**) or bidentate

(**Figure 1c, e**). ‘Monodentate’ means that there is only one bond formed between the anion and the surface, whereas ‘bidentate’ means that there are two bonds formed between the anion and the surface. The bidentate inner-sphere bonds can be mononuclear (**Figure 1c**) or binuclear (**Figure 1e**), which refer to the minimum number of mineral cations that are needed to link the anion’s oxygens to the mineral.

## Vibrational Spectroscopies

Vibrational spectroscopy techniques such as FTIR and Raman spectroscopy probe the vibrational properties of molecules. Characteristic vibrational properties of oxyanions originate from X–O(H) stretching and bending vibrations. Each oxyanion has its characteristic pattern of vibrations that relates to its symmetry state. The symmetry, and consequently the vibration pattern, changes when an oxyanion complexes with other ions such as protons or metal cations. The changes in the number and position of the virbational bands provide information about the speciation and coordination state of the oxyanion. This behavior is the underlying principle of the determination of the types of surface species using vibrational spectroscopy.

Spectra of outer-sphere complexes are similar to the spectra of the free anion in solution. Differences in position and number of adsorption bands between the spectra of solution and adsorbed species indicate coordination with the surface, that is, inner-sphere complexation. Determination of the coordination state of an inner-sphere complex (mono- or bidentate and mono- or binuclear) is usually based on the comparison of spectra of adsorbed species with the literature data of spectra of coordinated compounds with known structures. Extensive infrared and Raman spectral data of inorganic and coordination compounds are available.

Attenuated total reflectance–Fourier-transformed infrared (ATR-FTIR) spectroscopy is an *in situ*, infrared sampling technique that can be used to provide information about surface species at the mineral–water interface. Note that, since water strongly adsorbs infrared, it was the introduction of ATR (and other similar detection modifications, such as cylindrical internal reflectance, CIR) that has given us the needed breakthrough to very effectively use FTIR in aqueous (i.e., natural) mediums. This technique has been applied successfully to investigations of various anions ($PO_4^{3-}$, $B(OH)_4^-$, $SO_4^{2-}$, $CO_3^{2-}$, $NO_3^-$) at metal oxide–solution interfaces.

The use of ATR-FTIR does not fully avoid water absorption of the infrared signal. The spectral window in metal oxide aqueous suspensions is limited by the strong adsorption by water and metal oxide lattice vibrations in the region below 1000 cm$^{-1}$ and the strong adsorption by water in the region of 3000–4000 cm$^{-1}$. For example, the vibrations of the oxyanions of Se and As are expected below 1000 cm$^{-1}$ and, consequently, are difficult to observe in these systems with the ATR-FTIR technique.

Diffuse reflectance FTIR (DRIFT) is an infrared sampling technique that requires air-drying of the sample. Usually this does not significantly affect the surface speciation and coordination of the adsorbed oxyanion because a very thin layer of water should remain on the mineral surface. Since there is less interference with the strong absorbing water, this sampling technique offers the possibility to investigate OH-stretching in the spectral range of 3400–3800 cm$^{-1}$. The OH-stretching bands can provide information about the effect of oxyanion adsorption on the surface OH groups of the mineral surface.

Raman spectroscopy can provide a window in the spectral region below 1000 cm$^{-1}$ in the investigation of aqueous minerals and, accordingly, can offer complementary data to IR spectroscopy. One of the attractive aspects of Raman spectroscopy for its application to minerals in aqueous suspension is that Raman is transparent to water and, therefore, offers the possibility of *in situ* investigation of surface species. The recent advances in Raman instrumentation have brought new opportunities for applications to the investigation of surface species at mineral–water interfaces. Raman spectroscopy has been used to study the interfacial species on metal oxides of $SeO_4^{2-}$, $SO_4^{2-}$, $AsO_4^{3-}$ and $AsO_3^{3-}$.

## X-Ray Absorption Spectroscopy

X-ray absorption spectroscopy (XAS) is another techniques that can be applied to *in situ* studies of adsorbates at mineral–water interfaces. XAS is an element-specific, short-range probe, providing information about the local structural and compositional environment of an adsorbing atom. X-ray absorption near-edge structure (XANES) spectra provide information about the bonding and oxidation state of the adsorbing atom. Extended X-ray absorption fine structure (EXAFS) can provide information about the composition and structure in the vicinity of the element. The absence of back-scattering beyond the first shell indicates an outer-sphere complex. In the case of inner-sphere complexes, the number, distance, and type of back-scatterers provide information for identifying the type of inner-sphere complex (mono-versus bidentate). XAS has been used to identify the adsorption mechanisms of the oxyanions of Se, As, and Cr. The technique requires high-energy X-rays, which are only available at synchrotron facilities.

## Sorption Strength of Oxyanions

The strength of the bond of oxyanions with mineral surfaces strongly affects the fate of the oxyanion in the environment. Weakly held anions leach easily and are lost from the upper soil horizons down toward the groundwater or are readily carried in the surface runoff to nearby streams. There are a few ways to determine the relative strength of these bonds. One common method used involves adsorption isotherms and another involves adsorption edges.

An adsorption isotherm analysis is based on measuring the amount of anion adsorbed by a solid phase as a function of the equilibrium concentration present in the liquid phase, as shown in **Figure 2** for sulfate adsorption onto an aluminum oxide. The curvature of the isotherm, which is particularly visible at low aqueous concentrations, is a function of the adsorption strength. Qualitatively, the sharper the rate of increase and the sharper the amount of curvature observed in the adsorption isotherm, the stronger the adsorption strength of the compound with the surface. It is the curvature of the adsorption isotherm, not the adsorption maximum, that gives an indication of the adsorption strength.

Efforts to quantify the adsorption strength have been the focus of much debate in the literature for nearly a century. For example, if we assume a simple adsorption mechanism:

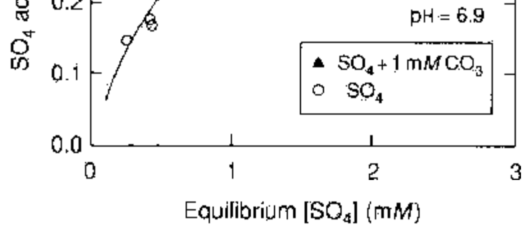


**Figure 2** Adsorption isotherms of sulfate on an aluminum oxide (bayerite). Notice that the presence of carbonates increased the adsorption strength of sulfate, which is contrary to the competitive adsorption behavior typically observed when two adsorbing anions are in solution. (Reprinted from Wijnja H and Schulthess CP (2000) Interaction of carbonate and organic anions with sulfate and selenate adsorption on an aluminum oxide. *Soil Science Society of America Journal* 64: 898–908.)

$$\text{Solid Surface} + \text{Aqueous Anion} \rightleftharpoons \text{Adsorbed Anion} \qquad [1]$$

then the amount adsorbed can be predicted mathematically using data-fitting techniques on the Langmuir isotherm equation:

$$\text{Amount Adsorbed} = MKc/(1 + Kc) \qquad [2]$$

where $M$ is maximum adsorption, which can be obtained from the isotherm at high aqueous concentrations; $K$ is the Langmuir constant, which is equal to the ratio of $P/R$, with $P$ the adsorbed anion concentration, $R$ the (concentration of unoccupied surface sites) × (aqueous anion concentration); and $c$ is the aqueous anion concentration at reaction equilibria. The adsorption strength is quantified here by the Langmuir constant, $K$. Much debate arises, however, about just how the adsorption mechanism should be modeled; that is, modeling the adsorption reaction by eqn [1] and other similar equations are not yet fully validated. Note that a model that does not fit the data is clearly disproved. However, a model that does fit the data is not automatically adequately validated.

One can also infer the relative strength of an adsorbing anion based on the pH value of the adsorption edge, which is shown in **Figure 3**. For a given mixture, the amount of aqueous anion adsorbed by the solid phase is strongly dependent on the equilibrium pH of the system. Working down the pH scale, there is no adsorption of anions at high pH values. At intermediate pH values, there is an abrupt increase in adsorption. The pH value measured at the midpoint of this abrupt adsorption increase is the pH value of the adsorption edge. At much lower pH values, there is a decrease in the amount of anions adsorbed. The entire bell-shaped curve is known as an adsorption envelope.

If we assume that the adsorption mechanism involves the coadsorption of hydroxyl anions ($OH^-$), for example:

$$\text{Surface–OH} + \text{Anion}^{n-}(aq) \rightleftharpoons \text{Surface–Anion}^{1-n} + OH^-(aq) \qquad [3]$$

then we can use the pH value of the adsorption edge to compare the relative adsorption strength of the anion. At equimolar concentrations, the anion that has a higher adsorption edge value is the anion that binds more strongly with the surface. Notice in eqn [3] that a strong surface–anion bond will require a higher concentration of $OH^-$ anions to reverse the reaction. The midpoint of equilibria for eqn [3] occurs at the adsorption edge. Accordingly, if the adsorption of an anion is strong, then the quantity of OH$^-$ anions needed to reverse the reaction will be high, and the pH value of the adsorption edge will be high also.

Efforts to quantify the adsorption strength based on the pH value of the adsorption edge are not yet fully resolved. The problem arises in that we are not definitively sure that eqn [3] or other similar equations fully describe the adsorption process. Consequently, the mathematical interpretation of the pH-dependent data is still discussed extensively in soil science research articles. We do know, however, that if the pH is kept constant, then eqn [3] will simplify to eqn [1], and equations similar to eqn [3] will simplify to equations similar to eqn [1]. We also know that anion adsorption does involve the release of hydroxyl anions. Note that it is difficult to distinguish hydroxyl release (which chemically removes protons from solution) from proton coadsorption (which physically removes protons from solution). Accordingly, essentially identical concepts can be presented using equations that illustrate proton coadsorption to the surface rather than hydroxyl release from the surface.

Using the tools outlined above, a general review of the published data suggests the following trend: the larger the negative charge present on the oxyanion, the stronger the affinity of the oxyanion toward the surface. Within each group, there is much uncertainty on which is more strongly adsorbed, and this is probably due to numerous reasons (such as, presence of surface contaminants). In general, the affinity of trivalent oxyanions follows the sequence: $AsO_3^{3-} > PO_4^{3-} \approx AsO_4^{3-}$. Divalent oxyanions adsorb more weakly, and in general follow the affinity sequence $SeO_3^{2-} > CrO_4^{2-} > CO_3^{2-} > SeO_4^{2-} \approx SO_4^{2-} > MoO_4^{2-}$. Monovalent oxyanions such as $ReO_4^-$, $NO_3^-$ and $ClO_4^-$ are very weakly adsorbed. Another, more useful observation is the close correlation of adsorption affinities of oxyanions with the stability constants ($pK_a$ values) of their conjugate acids; noteworthy is the high adsorption affinity of $B(OH)_4^-$ by minerals. For these reasons, nitrates or perchlorates are sometimes used as background electrolytes in adsorption studies and are often assumed not to adsorb at all or to cause no significant impact on the adsorption process of the other compounds present.

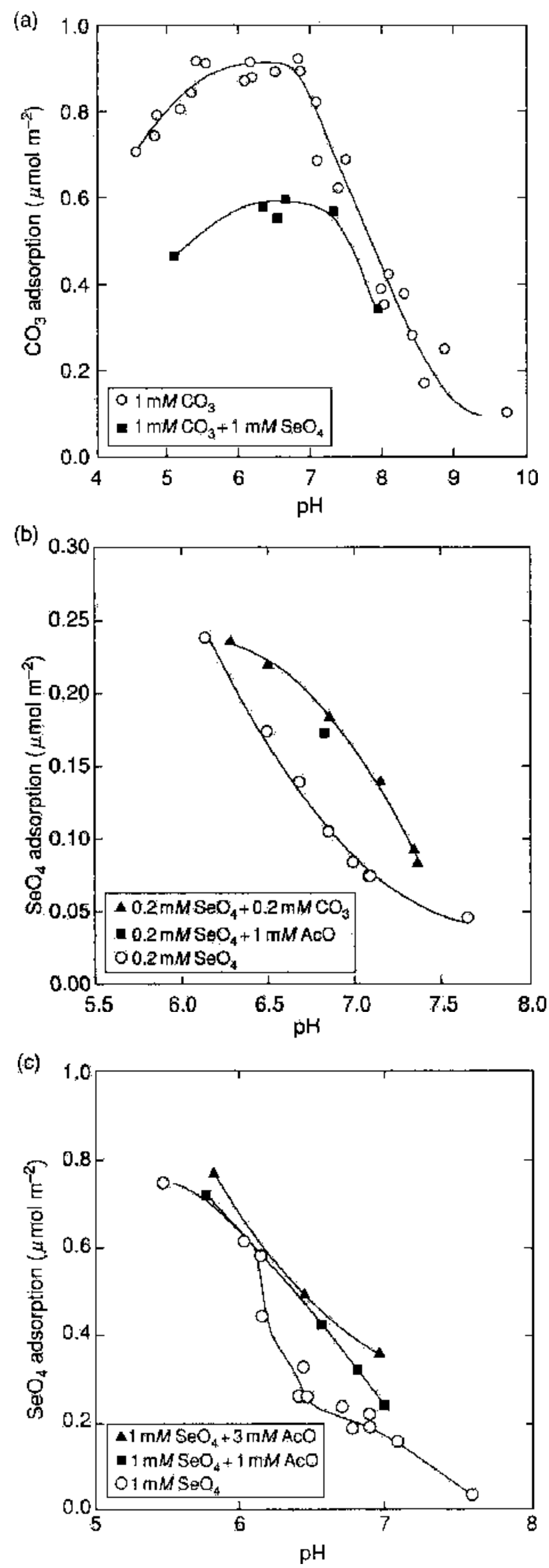


**Figure 3** Adsorption envelopes (adsorption as a function of pH) for various oxyanions on goethite: (a) adsorption of carbonate. Notice the competitive behavior of selenate on carbonate adsorption; (b, c) adsorption of selenate in the presence of carbonate and acetate. Notice the enhancement by these oxyanions of selenate adsorption. (Reprinted from Wijnja H and Schulthess CP (2002) Effect of carbonate on the adsorption of selenate and sulfate on goethite. *Soil Science Society of America Journal* 66:1190–1197.)

## Impact of Ions on Adsorption of Oxyanions

The presence of dissolved electrolytes in the mixture affects the adsorption of oxyanions. Adsorption studies, therefore, are pursued under constant ionic-strength conditions. These background electrolytes consist of dissolved cations and anions such as $Na^+Cl^-$ or $Na^+NO_3^-$. It is critical to the adsorption study that the reactivity of these ions with the surface be minimal, preferably completely unreactive or at least constant in their reactivity. Sometimes, an aqueous phase complexation reaction occurs between the oxyanion being studied and the components of the background electrolyte. This will complicate the analysis and should be avoided as much as possible, preferably by using a different background electrolyte.

The presence of dissolved electrolytes affects the adsorption process in numerous ways. First, it raises the ionic strength of the solution. This tends to lower the electrostatic attraction of the adsorbate due to the lowering of surface potential or the masking of the surface charge. This impact on the adsorption process is not a serious concern as long as the ionic strength is kept constant throughout the study.

Second, the background anion can compete with the oxyanion for adsorption sites, which is very likely at high ionic-strength concentrations. Although the adsorption of the background anion (e.g., $Cl^-$) is typically extremely weak, it is rarely zero and its competitive impact increases significantly as its concentration in solution increases. This impact on the adsorption process is not serious as long as the concentration of the background ions are kept constant. It will, however, cause some complications to attempts to understand or quantify the adsorption strength of the oxyanion. When comparing adsorption isotherms or adsorption edges of adsorption envelopes, it is helpful if the oxyanions are compared at the same ionic strength and initial conditions. This impact is shown in **Figure** 3a at low pH, where the concentration of weakly competitive aqueous $Cl^-$ anions has increased to the point of rendering it significantly competitive with the carbonate oxyanion. The source of the $Cl^-$ anion at low pH is from the addition of acids (namely, HCl) needed to achieve the low-pH conditions.

Competition between anions is also shown in **Figure** 3a, where the presence of selenate significantly reduces the adsorption of carbonate. Competition between oxyanions is commonly observed. Competitive behavior should be expected due to the structural similarities of the oxyanions and the similarity of their adsorption sites. High $Cl^-$ concentrations will also lower the amount adsorbed over a broad pH range due to competitive adsorption.

Third, the reactivity of the background anion (e.g., $Cl^-$) with the surface may be pH-dependent. Note that the first two problems mentioned above assume, for simplicity, that the reactivity of the background anion is pH-independent. However, this is perhaps one of the most serious problems that can arise in the study of oxyanion-adsorption processes. In these situations, the reactivity of the background anion must be monitored if the adsorption reactivity of the oxyanion with the solid surface is to be properly understood.

Fourth, some anions may be present in the mixture that actually enhance the adsorption reaction of the aqueous oxyanion with the solid surface. Note that this is a counterintuitive process because traditional views do not have an anion causing an increase in adsorption of another similarly charged ion. **Figure 2** shows a sharp increase in adsorption of sulfate on an aluminum oxide in the presence of carbonate. **Figure 3b** and c shows that this enhancing effect is also pH-dependent. The enhancement of oxyanion adsorption, such as sulfate and selenate adsorption on aluminum oxide and iron oxide, has been observed in the presence of carbonates, acetate, and formate anions. At high concentrations, these anions seem to become competitive. Hence, an enhancing effect does not exclude the existence of a concurrent competitive effect. The enhancing effect is most pronounced at low concentration levels, where any possible competitive effect would also be at its lowest level (**Figure 4**). It is

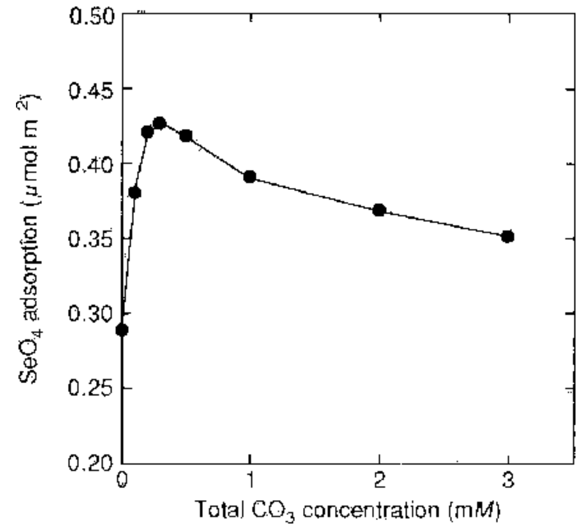


**Figure 4** Adsorption of 1 m*M* selenate at pH 6.8 on an aluminum oxide (bayerite) as a function of the amount of carbonate added. Notice that the presence of carbonates increases the adsorption of selenate at low carbonate concentrations. (Reprinted from Wijnja H and Schulthess CP (2000) Interaction of carbonate and organic anions with sulfate and selenate adsorption on an aluminum oxide. *Soil Science Society of America Journal* 64: 898–908.)

not clear how prevalent the enhancing effect is in nature or in laboratory studies. We do know, however, that carbonates are nearly always present in aquatic systems. Their concentration in laboratory solutions varies widely depending on the pure water sources used or on the techniques used to minimize or control the carbonate concentrations. In other words, aqueous carbonate impurities can easily enter the sample preparations unbeknownst to the experimenters and cause a wide range in the variance of the data collected. This creates a serious problem for those that wish to understand and quantify the adsorption of oxyanions by solid surfaces in aqueous mediums.

## Sorption Mechanisms of Oxyanions

The adsorption reaction of oxyanions at a mineral surface is described as a surface complexation reaction between the oxyanion and a surface functional group forming a surface complex. With respect to the types of surface complexes or adsorption mechanisms, there are commonly two major types distinguished: inner-sphere and outer-sphere complexes. An inner-sphere complex is formed via a ligand-exchange reaction with a surface-OH or surface-$OH_2^+$ group and is directly coordinated with the underlying surface metal ion. In a monodentate inner-sphere complex, one bond is formed with a surface group. In a bidentate complex, two bonds are formed, either with one surface metal ion (mononuclear) or with two surface metal ions (binuclear). In contrast, outer-sphere complexes form mainly by electrostatic interactions and at least one water molecule is interposed between the oxyanion and the mineral surface. Diffuse-ion swarm ions do not form a complex with a surface functional group, but instead are only bound through electrostatic interactions that neutralize the surface charge. Spectroscopic tools are essential for the development of detailed descriptions of the adsorbed oxyanion. A mechanistic description of the adsorption process requires spectroscopic information of the adsorbed species, plus information on the stoichiometry of the reaction and reaction rates, as well as a measure of the adsorption strength.

Oxyanions with a relatively high adsorption affinity such as $PO_4^{3-}$, $AsO_4^{3-}$, $SeO_3^{2-}$, and $B(OH)_4^-$ form predominantly inner-sphere complexes on metal oxide surfaces. Both monodentate and bidentate complexes have been found on metal oxides for several of these oxyanions. For example, it has been shown that the relative distribution of monodentate and bidentate inner-sphere complexes of $AsO_4^{3-}$ and $CrO_4^{2-}$ varies with the surface coverage.

The $CO_3^{2-}$ anion forms monodentate inner-sphere complexes at metal oxide–water interfaces. Oxyanions with low affinity, such as $NO_3^-$ and $ClO_4^-$, absorb mainly as outer-sphere and diffuse-swarm ions.

Several anions that have high or moderate sorption affinity, such as $AsO_3^{3-}$, $SeO_4^{2-}$, and $SO_4^{2-}$, show intermediate behavior, where both inner- and outer-sphere surface complexes coexist at metal oxide–water interfaces. For these moderately adsorbing anions, the relative distribution of inner- versus outer-sphere adsorption depends on pH, ionic strength, and type of mineral.

When an anion is held as an inner-sphere compound, then it has probably been preceded by an initial outer-sphere attraction to the mineral surface. The mechanism of forming an inner-sphere bond must involve the release of a molecule of water from the solid–liquid interface into the bulk solution. A net volume change occurs to the system whenever this process occurs, and this can be measured with a dilatometer. Based on measured volume changes, there is approximately one molecule of $H_2O$ released into the bulk solution for every aqueous molecule of $SO_4^{2-}$, $SeO_4^{2-}$, or $H_2PO_4^-$ adsorbed onto an amorphous iron hydroxide at pH 4.5, which should be expected for inner-sphere adsorption of these oxyanions.

Anion adsorption is accompanied by proton uptake. The net proton uptake is the result of a ligand exchange reaction (OH release), surface protonation, oxyanion protonation on the surface, or a combination of these. Quantifying the number of protons coadsorbed (or OH released) by the surface per molecule of oxyanion adsorbed is an important component of our understanding of the adsorption process. This stoichiometry of the reaction directly affects how the surface reaction will be described, such as the 1:1 stoichiometry described in eqn [3]. This, in turn, also affects how the surface complexation model will be developed. The best way to obtain data on the adsorption or coadsorption of protons is by an acid–base titration method known as the backtitration technique.

Anion adsorption affects the electrostatic properties of mineral surfaces. The net surface charge of a mineral surface is the result of proton-transfer reactions and surface-complexation reactions. Electrophoretic mobility measurements can reveal the net surface charge changes due to oxyanion adsorption. In general, the adsorption of an oxyanion tends to lower the positive charge present on a surface. The positively charged surface does not necessarily become negatively charged following the adsorption of oxyanions. Instead, it tends to approach a net neutral surface charge. Taking these data in conjunction with the proton coadsorption data mentioned above, the adsorption of oxyanions and proton coadsorption tends to form net neutral products on the mineral surfaces.

The adsorption of oxyanions generally reaches a maximum at around the p$K_a$ value for the conjugate Bronsted acid, as shown in Figure 3a for carbonate adsorption (p$K_a$ = 6.3). One can generally infer that the higher the value of the oxyanion's acidity constant (p$K_a$), the stronger will be its adsorption by a hydrophilic surface. This further reflects the importance of the coadsorption of protons on the adsorption mechanism of oxyanions. The mechanism of bringing the coadsorbed protons to the mineral surface does not necessarily follow a separate, independent path from that taken by the adsorbing oxyanion. The proton can be brought toward the mineral surface by the protonated oxyanions. The fate of the coadsorbed proton may be: (1) to remain complexed with the adsorbed oxyanion (hence, the oxyanion is held as a Bronsted acid on the mineral surface); (2) to remain complexed with the surface (hence, the oxyanion's negative charge is neutralized by a positively charged surface-$OH_2^+$ surface, which may or may not be directly below or adjacent to the adsorbing oxyanion); or (3) it may react with a surface-OH group and form an $H_2O$ product that diffuses back to the bulk solution (this may also be described schematically as a hydroxyl exchange mechanism, where the hydroxyl is neutralized by free protons in the aqueous phase).

## Site Preference and Site Reactivity

Surface hydroxyl groups constitute the complexation sites on metal oxide surfaces suspended in aqueous solutions. The surface groups behave as weak Bronsted acids or bases, capable of ionizing to yield a pH-dependent charged surface or serving as a ligand in surface complexation reactions. Analysis of the crystal structure of metal oxides indicates the existence of various different types of surface hydroxyl groups with different reactivities. Their reactivity depends, among other factors, on their coordination environment (singly, doubly, and triply metal-coordinated). The different crystal planes of a metal oxide mineral have a different composition with respect to the types of surface groups and densities of these groups.

With respect to complexation of oxyanions, the singly coordinated surface hydroxyls are believed to be most reactive. An FTIR study on the effect of As oxyanions on the various different types of surface groups has indicated that the singly coordinated surface groups are involved in the ligand exchange reaction to form inner-sphere complexes.

There are several things that the adsorbed oxyanion can do after it has adsorbed to the mineral surface. It can, of course, desorb from the surface and diffuse back to the aqueous bulk solution. Given the high energy of hydration of many of these oxyanions, this desorption step is very likely and is the basis for establishing the equilibrium condition where only a specific density of anions remain adsorbed to the surface for a given concentration of anions in solution.

The adsorbed oxyanion can also diffuse into the mineral through a process known as solid diffusion (product is shown in Figure 1b). Solid diffusion at ambient temperatures is a very slow process. However, the oxyanion may also became entrapped inside the mineral through the process of oxyanion-enhanced dissolution of the mineral followed by coprecipitation of the oxyanion–metal complex.

A final option for the adsorbed oxyanion is to migrate laterally, traveling horizontally on the mineral surface along a two-dimensional matrix. This is dependent on the strength of the bond of the oxyanion with the surface. The lateral, two-dimensional mobility of the oxyanion allows it to migrate toward a more stable adsorption site (more stable, that is, relative to its point of initial collision and reaction with the surface). Lateral mobility of adsorbed ions is known to occur on solid–gas interfaces, but data are missing that accurately quantify this mobility on solid–liquid interfaces.

*See also:* **Cation Exchange**

## Further Reading

Fendorf S, Eick MJ, Grossl P, and Sparks DL (1997) Arsenate and chromate retention mechanisms on goethite. 1. Surface structure. *Environmental Science and Technology* 31: 315–320.

Nakamoto K (1997) *Infrared and Raman Spectra of Inorganic and Coordination Compounds*, 5th edn. New York: John Wiley & Sons, Inc.

Schulthess CP and Sparks DL (1986) Backtitration technique for proton isotherm modeling of oxide surfaces. *Soil Science Society of America Journal* 50: 1406–1411.

Schulthess CP and Sparks DL (1991) Equilibrium based modeling of chemical sorption on soils and soil constituents. *Advances in Soil Science* 16: 121–163.

Schulthess CP and Hu Z (2001) Impact of carbonate on the adsorption of selenate and sulfate on goethite. *Soil Science Society of America Journal* 66: 1190–1197.

Sun XH and Doner HE (1996) An investigation of arsenate and arsenite bonding structures on goethite by FTIR. *Soil Science* 161: 865–872.

Yamaguchi NU, Okazaki M, and Hashitani T (1999) Volume changes due to $SO_4^{2-}$, $SeO_4^{2-}$, and $H_2PO_4$ adsorption on amorphous iron (III) hydroxide in an aqueous suspension. *Journal of Colloid and Interface Science* 209: 386–391.

# SORPTION–DESORPTION, KINETICS

**D L Sparks**, University of Delaware, Newark, DE, USA

## Introduction

In the past two decades, as concerns and interests about soil and water quality have increased, soil and environmental chemists, environmental and chemical engineers, and geochemists have increasingly realized that sorption/desorption reactions in soil and other subsurface environments are time-dependent. Thus, to predict accurately the fate, mobility, speciation, and bioavailability of environmentally important plant nutrients, trace elements, radionuclides, and organic chemicals in soils, one must understand the kinetics and mechanisms of the reactions. This article will provide an overview of the topic of kinetics of sorption/desorption on soils with applications to environmentally important reactions.

## Rate-Limiting Steps and Time Scales for Sorption Processes

A number of transport and chemical reaction processes can affect the rate of soil chemical reactions. The slowest of these will limit the rate of a particular reaction. The actual chemical reaction (CR) at the surface, for example, adsorption, is usually very rapid and not rate-limiting. Transport processes (**Figure 1**) include: (1) transport in the solution phase, which is rapid, and in the laboratory, can be eliminated by rapid mixing; (2) transport across a liquid film at the particle/liquid interface (film diffusion (FD)); (3) transport in liquid-filled macropores (>2 nm), all of which are nonactivated diffusion processes and occur in mobile regions; (4) diffusion of a sorbate along pore wall surfaces (surface diffusion); (5) diffusion of sorbate occluded in micropores (<2 nm) (pore diffusion); and (6) diffusion processes in the bulk of the solid, all of which are activated diffusion processes. Pore and surface diffusion can be referred to as interparticle diffusion while diffusion in the solid is intraparticle diffusion.

The energy of activation value ($E_a$), which can be obtained by employing the Arrhenius equation ($\ln k = \ln A_f - E_a/RT$ where $k$ is the rate constant, $A_f$ is the frequency factor, $R$ is the universal gas constant, and $T$ is absolute temperature), is a useful kinetic parameter to assess rate-limiting steps. Some $E_a$ values for various reactions or processes are shown in Table 1. Low $E_a$ values usually indicate diffusion-controlled transport and physical adsorption processes, whereas higher $E_a$ values would indicate chemical reaction or surface-controlled processes.

Sorption reactions can range over a wide time scale – from seconds to years. The type of soil component can drastically affect the reaction rate. For example, sorption reactions are often more rapid on clay minerals such as kaolinite and smectites than on vermiculitic and micaceous minerals. This is in large part due to the availability of sites for sorption. For example, kaolinite has readily available planar external sites and smectites have primarily internal sites that are also quite available for retention of sorbates. Thus, sorption reactions on these soil constituents are often quite rapid, even occurring on time scales of seconds and milliseconds.

On the other hand, vermiculite and micas have multiple sites for retention of metals and organics, including planar, edge, and interlayer sites, with some of the latter sites being partially to totally collapsed. Consequently, sorption and desorption reactions on these sites can be slow, tortuous, and mass transfer-controlled. Often, an apparent equilibrium may not be reached even after several days or weeks. Thus, with vermiculite and mica, sorption can involve two to three different reaction rates: high rates on external sites, intermediate rates on edge sites, and low rates on interlayer sites.

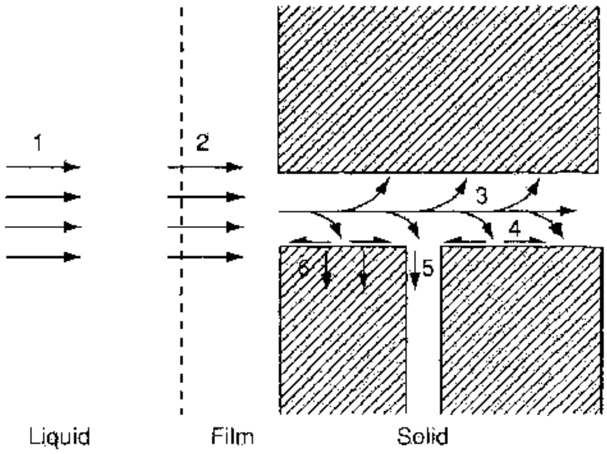


**Figure 1** Transport processes in solid–liquid soil reactions – nonactivated processes: (1) transport in soil solution; (2) transport across a liquid film at the solid–liquid inerface; (3) transport in a liquid-filled macropore; activated processes: (4) diffusion of a sorbate at the surface of the solid; (5) diffusion of a sorbate occluded in a micropore; (6) diffusion in the bulk of the solid. Reproduced with permission from Aharoni C and Sparks DL (1991) *Soil Science Society of America Special Publication* 27.

**Table 1** Energy of activation ($E_a$) values for various reactions or processes

| *Reaction of process* | *Typical range of* $E_a$ *values (kJ mol$^{-1}$)* |
|---|---|
| Physical adsorption | 8 to 25 |
| Aqueous diffusion | $<$21 |
| Pore diffusion | 20 to 40 |
| Cellular and life-related reactions | 21 to 84 |
| Mineral dissolution or precipitation | 34 to 151 |
| Mineral dissolution via surface reaction control | 42 to 84 |
| Polymer diffusion | 60 to 70 |
| Ion exchange | $>$84 |
| Isotopic exchange in solution | 75 to 201 |
| Solid-state diffusion in minerals at low temperatures | 84 to 503 |

Reproduced with permission from Lasaga AC and Kirkpatrick RJ (eds) (1981) *Kinetics of Geochemical Processes*. Washington, DC: Mineralogical Society of America; and Langmuir D (1997) *Aqueous Environmental Geochemistry*. Englewood Cliffs, NJ: Prentice Hall.

Metal sorption reactions on oxides, hydroxides, and humic substances depend on the type of surface and metal being studied, but the CR appears to be rapid. For example, CR rates of metals and oxyanions on goethite occur on millisecond time scales. Half-times for divalent Pb, Cu, and Zn sorption on peat range from 5 to 15 s. Heavy metal sorption on oxide and clay minerals increases with longer residence times (contact time between metal and sorbent). The mechanisms for these lower reaction rates are not well understood, but have been ascribed to diffusion phenomena, sites of lower reactivity, and surface nucleation/precipitation.

Sorption/desorption of metals, oxyanions, radionuclides, and organic chemicals on soils can be very slow, and may demonstrate a residence time effect, which has been attributed to diffusion into micropores of inorganic minerals and into humic substances, retention on sites of varying reactivity, and surface nucleation/precipitation.

## Sorption/Desorption Kinetics of Metals

The kinetics of heavy-metal sorption on soil components and soils is typically characterized by a biphasic process in which sorption is initially rapid followed by slow reactions (**Figure 2**). The rapid step, which occurs over milliseconds to hours, can be ascribed to CR and film diffusion processes. During this rapid reaction process, a large portion of the sorption may occur. For example, in **Figure 2a** one sees that ~90% of the total Ni sorbed on kaolinite and pyrophyllite occurred within the first 24 h. For Pb sorption on a Matapeake soil, 78% of the total Pb sorption

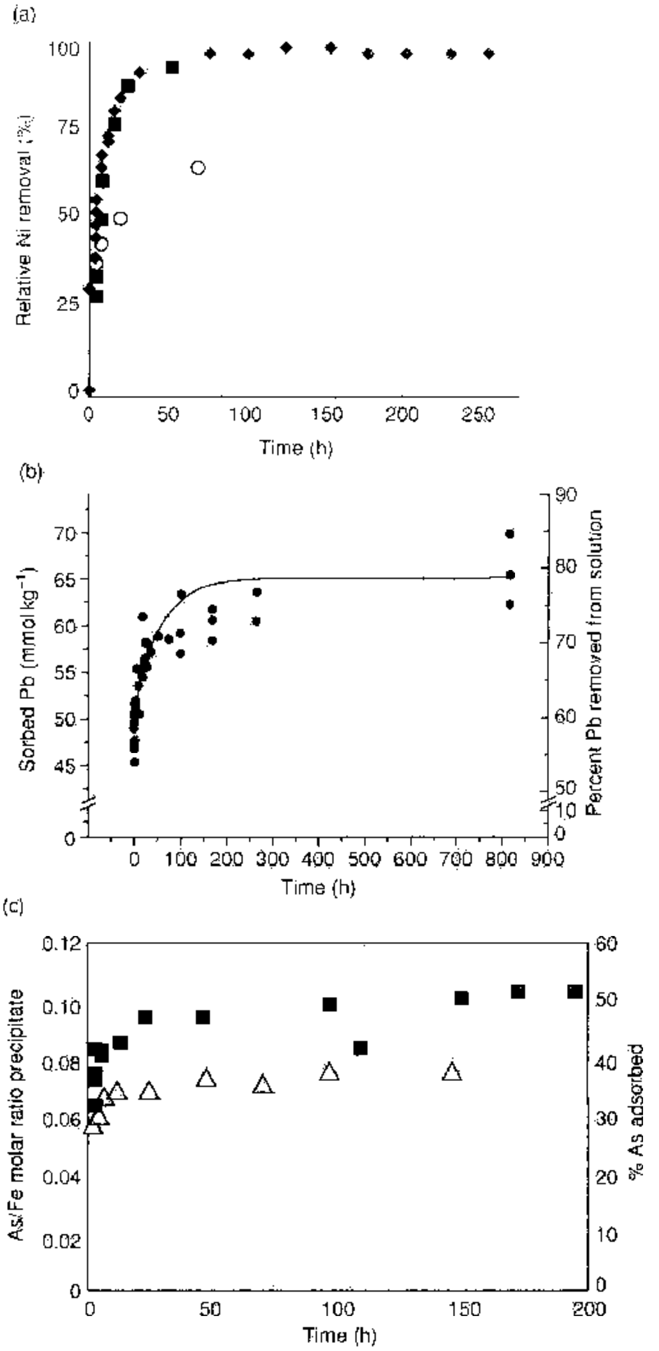


**Figure 2** Kinetics of metal and oxyanion sorption on soil minerals and soil. (a) Kinetics of Ni sorption (%) on pyrophyllite (◆); kaolinite (■); gibbsite ( ); and montmorillonite (○) from a 3 mmol l$^{-1}$ Ni solution, an ionic strength $I = 0.1$ mol l$^{-1}$ NaNO$_3$, and a pH of 7.5. Reproduced with permission from Scheidegger AM, Lamble GM, and Sparks DL (1997) Spectroscopic evidence for the formation of mixed-cation, hydroxide phases upon metal sorption on clays and aluminum hydroxides. *Journal of Colloid and Inter. Science* 186: 118–128. (b) kinetics of sorption on a Matapeake soil from a 12.25 mmol l$^{-1}$ Pb solution, an ionic strength $I = 0.05$ mol l$^{-1}$, and a pH of 5.5. Reproduced with permission from Strawn DG and Sparks D (2000) Effects of soil organic matter on the kinetics and mechanisms of Pb(II) sorption and desorption in the soil. *Soil Science Society of America Journal* 64: 144–156; (c) kinetics of As(V) sorption on ferrihydrite at pH 8.0 and 9.0. As(V) Total $1 \times 10^{-4}$ mol l$^{-1}$; Fe(III) Total $5 \times 10^{-4}$ mol l$^{-1}$; ■ Adsorption at pH 8.0; △ Adsorption at pH 9.0. Reproduced with permission from Juller CC, Davis JA, and Waychunas GA (1993) Surface chemistry of ferrihydride: part 2, kinetics of arsenate adsorption and coprecipitation. *Geochimica Cosmochimica Acta* 57: 2271–2282.

occurred in 8 min. Following the initial fast reaction, slow sorption continued, but only about 1% additional Pb was sorbed after 800 h (**Figure 2b**). **Figure 2c** shows a biphasic reaction for As(V) sorption on ferrihydrite. Within 5 min, a majority of the total sorption had occurred. Slow sorption continued for at least 192 h.

However, in some cases, the magnitude of sorption can greatly increase with longer reaction times. For example, it has been found that metal (e.g., Cd, Ni, and Zn) sorption on porous Fe-oxides such as goethite can significantly increase as time increases. These observations have been ascribed to a three-step mechanism: (1) adsorption of metals on external surfaces; (2) solid-state diffusion of metals from external to internal sites; and (3) metal binding and fixation at positions inside the goethite particle.

The amount of contact time between metals and soil sorbents (residence time) can dramatically affect the degree of desorption, depending on the metal and sorbent. Examples of this are shown in **Figure 3** and **Table 2**. In **Figure 3** the effect of residence time on $Pb^{2+}$ and $Co^{2+}$ desorption from hydrous Fe-oxide (HFO) was studied. With Pb, between pH 3 and 5.5 there was a minor effect of residence time (over 21 weeks) on $Pb^{2+}$ desorption, with only minor hysteresis occurring (hysteresis varied from <2% difference between sorption and desorption to ~10%). At pH 2.5, $Pb^{2+}$ desorption was complete within a 16-h period and was not affected by residence time (**Figure 3a**). In a soil where 2.1% soil organic matter (SOM) was present, residence time had little effect on the amount of Pb desorbed, but marked hysteresis was observed at all residence times (**Table 2**). This could be ascribed to the strong metal–soil complexes that occur and perhaps to diffusion processes.

With $Co^{2+}$, extensive hysteresis was observed over a 16-week residence time (**Figure 3b**) and the hysteresis increased with residence time. After a 16-week residence time, 53% of the $Co^{2+}$ was not desorbed, and even at pH 2.5, hysteresis was observed. The extent of Co reversibility with residence time was attributed to Co incorporation into a recrystallizing solid by isomorphic substitution and not to micropore diffusion. Similar residence time effects on Co desorption have been observed with both Fe- and Mn-oxides and with soil clays.

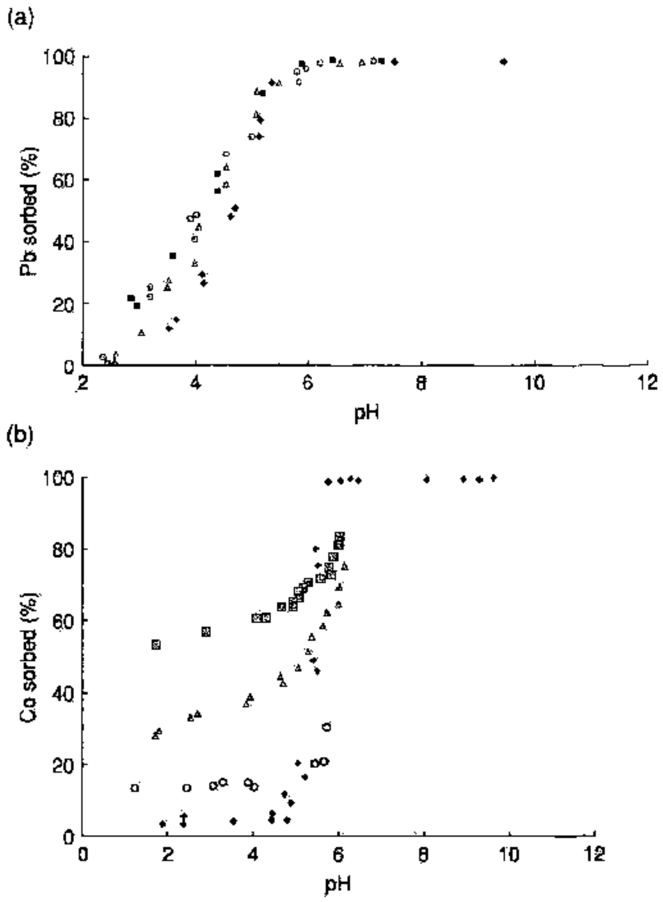


**Figure 3** Fractional sorption–desorption of (a) $Pb^{2+}$ (adsorption pH edge ◆ aging time = 0 weeks; desorption ■ aging time = 2 weeks; ▲ aging time = 12 weeks; ▩ aging time = 21 weeks) and (b) $Co^{2+}$ (adsorption pH edge ◆ aging time = 0 weeks; desorption ○ aging time = 2 weeks; ▲ aging time = 9 weeks; ▩ aging time = 16 weeks) to hydrous Fe(III) as a function of pH and HFO-$Pb^{2+}$ (a) and HFO-$Co^{2}$ (b) aging time. From Ainsworth CC, Pilou JL, Gassman PL, and Van Der Sluys WG (1994) Cobalt, cadmium, and lead sorption to hydrous iron oxide: residence time effect. *Soil Science Society of America Journal* 58: 1615–1623.

**Table 2** Effect of residence time on Pb desorption from a Matapeake soil

| *Residence time (days)* | *Sorbed Pb (mmol kg⁻¹)* | *Desorbed Pb (mmol kg⁻¹)* | *Percentage Pb desorbed (mmol kg⁻¹)* |
|---|---|---|---|
| 1 | 54.9 | 27.9 | 50.8 |
| 10 | 60.0 | 28.7 | 47.0 |
| 32 | 66.1 | 30.5 | 46.1 |

Reproduced with permission from Strawn DG and Sparks DL (2000) Effects of soil organic matter on the kinetics and mechanisms of Pb(II) sorption and desorption in soils. *Soil Science Society of America Journal* 64: 144–156.

## Kinetics of Metal Hydroxide Surface Precipitation/Dissolution

In addition to diffusion processes, the formation of metal hydroxide surface precipitates and subsequent residence time effects on soil sorbents can greatly affect metal release and hysteresis. Metal hydroxide precipitates can form on phyllosilicates, metal oxides, and soils. It has generally been thought that the kinetics of formation of surface precipitates was slow. However, recent studies have shown that metal hydroxide precipitates can form on time scales

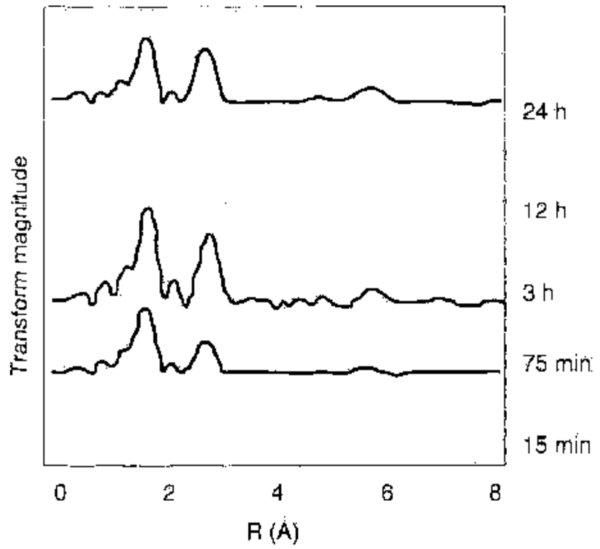


**Figure 4** Radial structure functions (derived from X-ray absorption fine structure (XAFS) analyses) for Ni sorption on pyrophyllite for reaction times up to 24 h, demonstrating the appearance and growth of the second shell (peak at approximately 2.8 Å) contributions due to surface precipitation and growth of a mixed Ni-Al hydroxide phase. Reproduced with permission from Scheidegger AM, Strawn DG, Lamble GM, and Sparks DL (1998) The kinetics of mixed Ni-Al hydroxide formation on clay and aluminum oxide minerals: a time-resolved XAFS study. *Geochimica Cosmochimica Acta* 62: 2233–2245.

of minutes. In Figure 4 one sees that mixed Ni-Al hydroxide precipitates formed on pyrophyllite within 15 min, and they grew in intensity as time increased. Similar results have been observed with other soil components and with soils.

The formation and subsequent 'aging' of the metal hydroxide surface precipitate can have a significant effect on metal release. In **Figure 5** one sees that as residence time ('aging') increases from 1 h to 2 years, Ni release from pyrophyllite, as a percentage of total Ni sorption, decreased from 23 to ~0%, when $HNO_3$ (pH 6.0) was employed as a dissolution agent for 14 days. This enhanced stability is due to the transformation of the metal-Al hydroxide precipitates to a metal-Al phyllosilicate precursor phase as residence time increases. This transformation occurs via a number of steps (**Figure 6**). There is diffusion of Si originating from weathering of the sorbent into the interlayer space of the layered double hydroxide precipitate (LDH), replacing the anions such as $NO_3^-$. Polymerization and condensation of the interlayer Si slowly transform the LDH into a precursor metal-Al phyllosilicate. The metal stabilization that occurs in surface precipitates on Al-free sorbents (e.g., talc) may be due to Ostwald ripening, resulting in increased crystallization. Thus, with time, metal (e.g., Co, Ni, and Zn) sorption on soil minerals often results in a continuum of processes from adsorption to precipitation to solid-phase transformation.

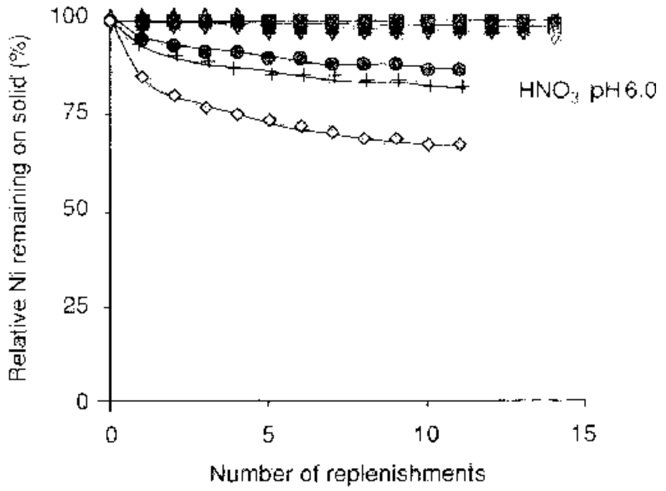


**Figure 5** Dissolution of Ni from surface precipitates formed on pyrophyllite at residence times of 1 h to 2 years. The figure shows the relative amount of $Ni^{2+}$ remaining on the pyrophyllite surface following extraction for a 24-h period (each replenishment represents a 24-h extraction) with $HNO_3$ at pH 6.0. ✚ 2 year, ⊠ 1 year; △ 6 months; × 3 months; ✧ 1 month; ● 24 h; + 12 h; ◇ 1 h. Reproduced with permission from Scheckel KG and Sparks DL (2001) Dissolution kinetics of Ni surface precipitates on clay mineral and oxide surfaces. *Soil Science Society of America Journal* 65: 719–728.

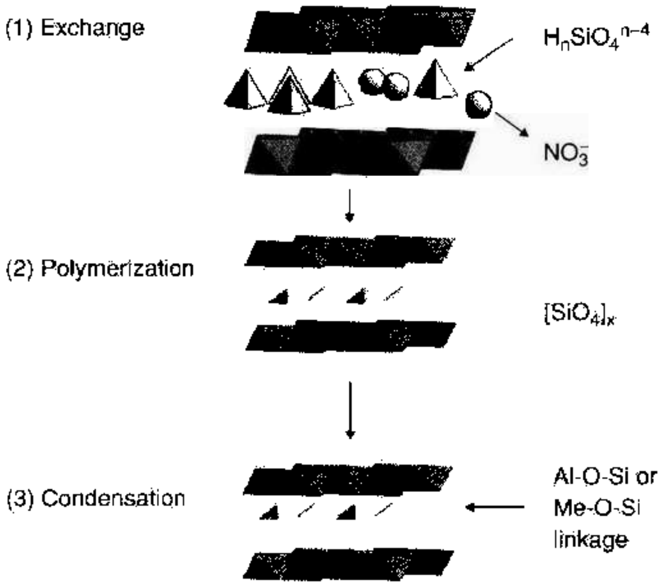


**Figure 6** Hypothetical reaction process illustrating the transformation of an initially precipitated Ni-Al LDH to a phyllosilicate like phase during aging. The initial step involves the exchange of dissolved silica for nitrate within the layered double hydroxide (LDH) interlayer followed by polymerization and condensation of silica on to the octahedral Ni-Al layer. The resultant solid possesses structural features common to 1:1 and 2:1 phyllosilicates. Reproduced with permission from Ford RG, Scheinost AC, and Sparks DL (2001) Frontiers in metal sorption/precipitation mechanisms on soil mineral surfaces. *Advances in Agronomy* 74: 41–62.

The formation of metal surface precipitates could be an important mechanism for sequestering metals in soils such that they are less mobile and bioavailable.

Such products must be considered when modeling the fate and mobility of metals such as $Co^{2+}$, $Mn^{2+}$, $Ni^{2+}$, and $Zn^{2+}$ in soil and water environments.

## Kinetics of Organic Contaminant Sorption/Desorption

Similarly to metals and oxyanions, many investigations have shown that sorption/desorption of organic contaminants is characterized by a rapid, reversible stage followed by a much slower, nonreversible stage or biphasic kinetics. The rapid phase has been ascribed to retention of the organic chemical in a labile form that is easily desorbed. The labile form of the chemical is also available for microbial attack. However, the much slower reaction phase involves the entrapment of the chemical in a nonlabile form that is difficult to desorb and is resistant to biodegradation. This slower sorption/desorption reaction has been ascribed to diffusion of the chemical into micropores of organic matter and inorganic soil components.

Recent theories have explained the slow diffusion of the organic chemical in SOM by considering SOM as a combination of 'rubbery' and 'glassy' polymers. The rubbery-like phases are characterized by an expanded, flexible, and highly solvated structure with pores of subnanometer dimensions (holes). Sorption in the rubbery phase results in linear, noncompetitive, and reversible behavior. The glassy phases have pores that are of subnanometer size and sorption in this phase is characterized by nonlinearity and is competitive.

The above theories, relating slow diffusion into organic matter to diffusion in polymers, are partially validated in recent studies measuring $E_a$ values for organic chemical sorption. $E_a$ values of 60–70 kJ mol$^{-1}$ have been determined for some chlorobenzenes, polychlorinated biphenyls (PCBs) and polycyclic aromatic hydrocarbons (PAHs) which are in the range for diffusion in polymers (Table 1). These values are much higher than those for pore diffusion (20–40 kJ mol$^{-1}$), suggesting that intraorganic matter diffusion may be a more important mechanism for slow organic chemical sorption than interparticle pore diffusion.

An example of the biphasic kinetics that is observed for many organic chemical reactions in soils/sediments is shown in Figure 7. In this study 55% of the labile PCBs was desorbed from sediments in a 24-h period, while little of the remaining 45% nonlabile fraction was desorbed in 170 h (Figure 7a). Over another 1-year period about 50% of the remaining nonlabile fraction desorbed (Figure 7b).

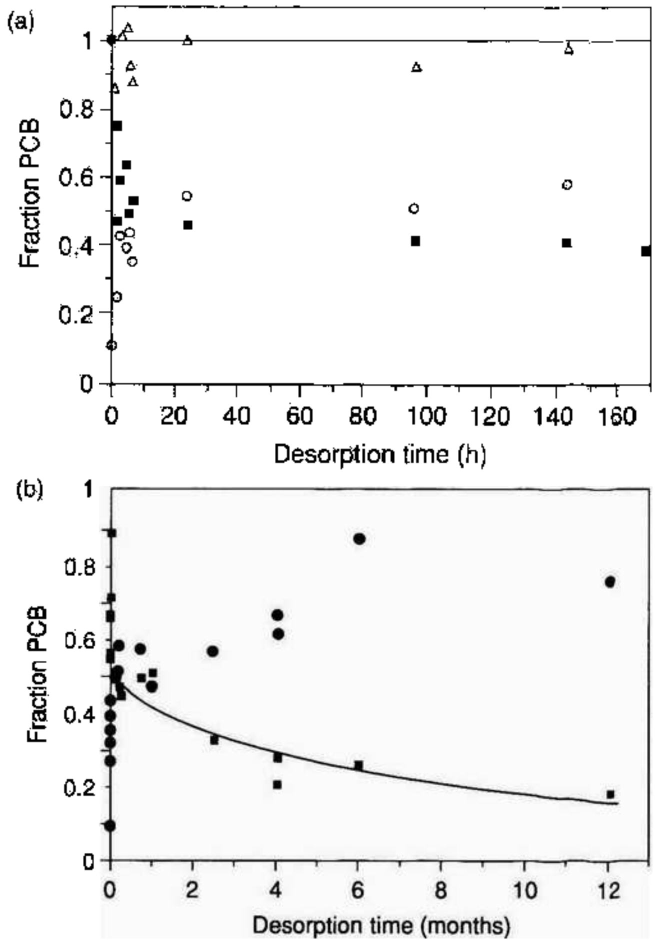


**Figure 7** (a) Short-term polychlorinated biphenyl (PCB) desorption in hours (h) from Hudson river sediment contaminated with 25 mg kg$^{-1}$ PCB. Distribution of the PCB between sediment (■) and XAD-4 resin (○) is shown, as well as the overall mass balance (△). The resin acts as a sink to retain the PCB that is desorbed. (b) Long-term PCB desorption in months from Hudson river sediment contaminated with 25 mg kg$^{-1}$ PCB. Distribution of the PCB between the sediment (■) and XAD-4 resin (●) is shown. The line represents a nonlinear regression of the data by the two-box (site) model. Reproduced with permission from Carroll KM, Harkness MR, Bracco AA, and Balcarcel RB (1994) Application of a permeant/polymer diffusional model to the desorption of polychlorinated biphenyls from Hudson river sediments. *Environmental Science Technology* 28: 253–258.

Similarly to some metals, it has been demonstrated that with increased residence time the nonlabile portion of the organic chemical in the soil/sediment becomes more resistant to release and is less available for microbial degradation. Soils reacted in the laboratory over a short period of time often show more rapid organic chemical release than field soils that have been contaminated for many years. The difference in release has been ascribed to greater diffusion into micropores of clay minerals and humic components that occurred at longer times.

One of the implications of these results is that while many transport and degradation models for organic contaminants in soils and water assume that the sorption process is an equilibrium process, the above studies clearly show that kinetic reactions must be considered when making predictions about the mobility and fate of organic chemicals. Moreover, $K_d$ values that are determined based on a 24-h equilibration period and which are commonly used in fate and risk assessment models can be inaccurate since 24-h $K_d$ values often overestimate the amount of organic chemical in the solution phase.

The finding that many organic chemicals are quite persistent in the soil environment has both good and bad features. The beneficial aspect is that the organic chemicals are less mobile and may not be readily transported in groundwater supplies. The negative aspect is that their persistence and inaccessibility to microbes may make decontamination more difficult, particularly if *in situ* remediation techniques such as biodegradation are employed.

## Conclusions and Future Directions

This article has discussed various basic and applied aspects of kinetics of sorption/desorption processes on soils and soil components including: rate-limiting steps and time scales for sorption processes, sorption/desorption kinetics of metals, kinetics of metal hydroxide surface precipitation/dissolution, and kinetics of organic contaminant sorption/desorption.

While great strides have been made in our understanding of kinetics and mechanisms of sorption processes on soils, a number of areas need further investigation and will be undoubtedly frontiers in the field. These include: the role of microbes on sorption kinetics and mechanisms at the soil/mineral water interface; the coupling of macroscopic studies with molecular-scale tools to study sorption reactions in real time; development of more predictive kinetic models to describe sorption processes on multireactive, heterogeneous soils; long-term sorption and, particularly, desorption rate studies; and increased knowledge of nucleation/precipitation phenomena at the soil/water interface and their impact on nutrient/contaminant mobility and bioavailability in the soil environment.

*See also:* **Cation Exchange**; **Sorption:** Metals; Oxyanions

## Further Reading

Aharoni C and Sparks DL (1991) Kinetics of soil chemical reactions: a theoretical treatment. In: Sparks DL and Suarez DL (eds) *Rates of Soil Chemical Processes.* SSSA special publication no. 27, pp. 1–18. Madison, WI: Soil Science Society of America.

Ainsworth CC, Pilou JL, Gassman PL, and Van Der Sluys WG (1994) Cobalt, cadmium, and lead sorption to hydrous iron oxide: residence time effect. *Soil Science Society of America Journal* 58: 1615–1623.

Alexander M (2000) Aging, bioavailability, and overestimation of risk from environmental pollutants. *Environmental Science Technology* 34: 4259–4265.

Bruemmer GW, Gerth J, and Tiller KG (1988) Reaction kinetics of the adsorption and desorption of nickel, zinc and cadmium by goethite: I. Adsorption and diffusion of metals. *Journal of Soil Science* 39: 37–52.

Pignatello JJ (2000) The measurement and interpretation of sorption and desorption rates for organic compounds in soil media. *Advances in Agronomy* 69: 1–73.

Sparks DL (1989) *Kinetics of Soil Chemical Processes.* San Diego, CA: Academic Press.

Sparks DL (1998) Kinetics of sorption/release reactions on natural particles. In: Huang PM, Senesi N, and Buffle J (eds) *Structure and Surface Reactions of Soil Particles,* pp. 413–448. New York: John Wiley.

Sparks DL (1999) Kinetics and mechanisms of chemical reactions at the soil mineral/water interface. In: Sparks DL (ed.) *Soil Physical Chemistry,* 2nd edn, pp. 135–191. Boca Raton, FL: CRC Press.

Sparks DL and Suarez DL (eds) (1991) *Rates of Soil Chemical Processes.* SSSA special publication no. 27. Madison, WI: Soil Science Society of America.

Sposito G (1994) *Chemical Equilibria and Kinetics in Soils.* New York: John Wiley.

# SPATIAL PATTERNS

**J H Görres and J A Amador**, University of Rhode Island, Kingston, RI, USA

## Introduction

Biological processes in the soil and associated properties are highly variable at different spatial scales. The sources of spatial variability in soils include both natural and anthropogenic processes that may act at microscopic, aggregate, field, landscape, or regional scale. Together these factors create complex mosaics of environments affecting biological processes at different scales. Identifying the sources of spatial heterogeneity and quantifying the heterogeneity of biological processes in soils are challenging. However, once understood, spatial heterogeneity may offer management options that affect ecosystem productivity, biodiversity, and soil and environmental quality.

## Background

Soil biological processes provide crucial feedback between belowground decomposition and aboveground processes such as productivity, herbivory, and diversity. Soil organisms recycle much of primary productivity in terrestrial ecosystems, contributing to sustained ecosystem growth. At the ecosystem level, the interaction between decomposers and producers is well understood. However, high spatial heterogeneity of soil biological processes may obscure ecologic relationships at a finer scale. For this reason, the origins of spatial variability in soil biological processes have themselves become an important topic of research.

Soil biological processes and their spatial variability have gained practical environmental and economic recognition. For example: bioremediation relies on soil microorganisms that can metabolize organic pollutants; precision agriculture attempts to reduce the amount of agrochemicals applied by variable applications that depend on the spatial distribution of nutrients in a field; and riparian wetland buffers promote denitrification, which reduces nitrate loading to streams, lakes, and estuaries. Understanding soil spatial variability is essential to successfully managing soil biological processes to improve environmental quality.

Organisms are at the root of many soil processes, influencing the development of physical and chemical soil properties. The close relationship between soil organisms and their habitat confounds a clear distinction between biological and physicochemical soil properties. For example, soil porosity is a physical soil property, but biogenic mechanisms such as aggregation and bioturbation contribute to its development. Components of the nitrogen cycle affect soil pH. Conversely, soil biological processes are affected by physicochemical properties such as water content and pH. Here soil biological processes are defined as those that are clearly associated with the distribution of organisms, such as microbial biomass and functional diversity, and their metabolic activity, such as decomposition and nutrient mineralization.

A common characteristic of soil biological processes is their high spatial variability. Depending on sampling intensity, coefficients of variation range from ten to several hundred percent. Spatial heterogeneity of soil biological processes is linked to the patchiness of the distribution of population and community level processes, and to pedological factors. It is clear that their spatial variability derives from a superimposition of processes that occur at multiple scales. Spatial variability has been categorized into variations at microsite (approx. $10^{-3}$–$10^{-1}$ m), field ($10^0$–$10^2$ m), landscape ($10^1$–$10^2$ m), and regional scale (greater than $10^3$–$10^5$ m). A case could also be made for adding a continental and global scale in which climatic and vegetational factors of ecoregions determine soil biological processes.

Regardless of scale, spatial variation can be attributed to a combination of random and deterministic variations. Random variations are those in which soil properties or processes are spatially independent, i.e., the property value at one point is not related to values at any other point. Deterministic variations of a property are characterized by spatial dependence. Spatial dependence develops because of processes that connect points in space, such as migration of organisms, diffusion of chemicals, or the integrative effect of plant roots. In environments favoring patchy distributions of organisms, such as soils, spatial dependence includes within-patch and between-patch components. Statistical techniques exist to characterize the properties of both random and deterministic variations.

## Sources of Variability

### Pedological Factors of Soil Biological Variability

Factors influencing soil formation lay out a coarse, multiscale spatial pattern of factors that influence soil formation. Pedologists distinguish five factors, four

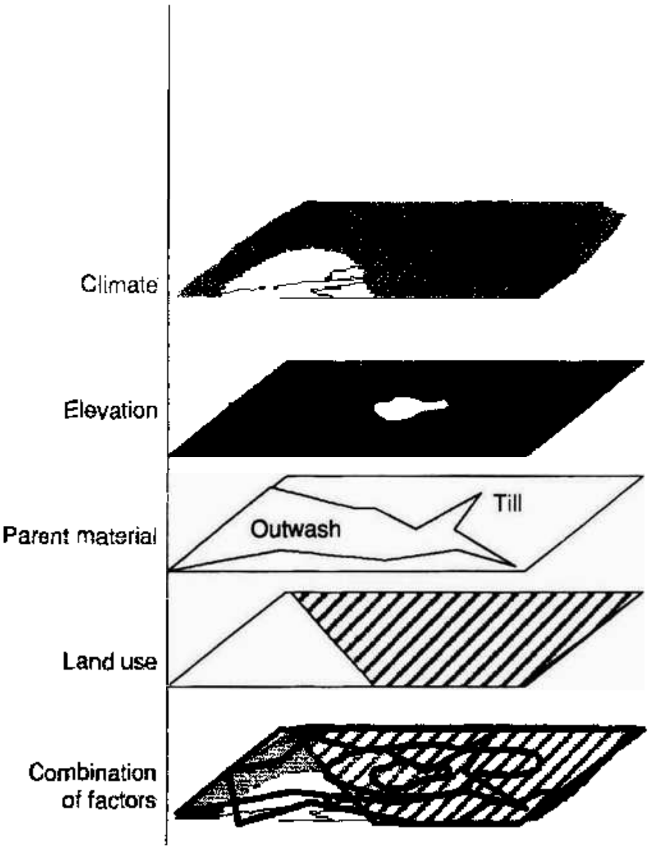


**Figure 1** Superposition of intrinsic and extrinsic properties. Climate, elevation, parent material, and land use combine to form a complex soil-property pattern that affects soil biological processes.

of which are distributed at different spatial scales: parent material, climate, topography, organisms, and time. Parent material and topography are regarded as landscape-scale factors. Climate has an overarching function affecting soil biology at all scales. For example, soil moisture, an important factor of biological activity in soils, is affected by topography, i.e., landscape position, elevation, and aspect, but also by climate. At finer scales, the action of organisms may be affected by resource distribution, microclimates, and microtopography.

The superimposition of parent material, climate, and topography (**Figure 1**) within a region establishes a coarse pattern, which is modified by organisms to form soils, creating multiscale distributions of biological soil properties. For example, nematode distributions are affected by large-scale processes such as texture and climate, and small-scale, organism-related processes such as life-history traits. Denitrification also displays multiscale variability. Microsite characteristics such as localized anaerobiosis at the center of soil aggregates and the availability of detrital resources to denitrifying bacteria are essential for denitrification. Yet anaerobic conditions crucial to denitrification are also controlled by landscape-scale factors such as proximity of groundwater to the soil surface.

In managed ecosystems, additional factors impact the spatial heterogeneity of soil properties. Soil variability is introduced by intrinsic and extrinsic factors. Intrinsic factors are those that act in the absence of soil management and these may coincide with pedological factors of soil formation. Extrinsic factors are associated with land management. In silvicultural, agricultural, and urban ecosystems, management tools such as artificial drainage, tillage, fertilization, planting, and control of pH are superimposed on intrinsic soil processes, adding to the variability of soil biological properties and processes. For example, enzymatic activity has been shown to be high within a crop row and low in the spaces between rows. Nonpoint source pollution, acid deposition, and land-use patterns may add to soil variability at the landscape, regional, and continental scale.

### Spatial Scales Related to Organisms

Biological processes occur at scales ranging from molecular (e.g., enzymes, extracellular polysaccharides, etc.) and organismal, to population and metapopulation. Thus, organisms create connections between points in space in several ways: their bodies occupy space; their motion spans distances between points in space; diffusion and bulk flow transport metabolic waste products (e.g., mineralized nutrients, gases); the presence and activities of larger organisms such as earthworms or plants may create spatially extended spheres of opportunity for other soil dwellers such as nematodes and microorganisms; small organisms may move by phoresis on, or in, larger motile organisms. These processes create spatial correlations that are expressed at different scales.

Whether spatial correlations due to body size are detected depends on the size of the soil sample taken. For example, the soil sample volume, also known as the sample support, is often many times greater than the dimensions of microscopic soil inhabitants, obscuring spatial correlation introduced by the body size of organisms. The longest linear dimensions of soil organisms range from less than a micrometer (microorganisms) to several meters (plants). Even within a group of organisms such as the soil fauna, sizes vary over several orders of magnitude (Table 1). The effect of plants on the spatial variability of mineral nitrogen can thus be detected, using typical soil cores (e.g., 5 cm in diameter and 10 cm long). However, the typical soil core is several orders of magnitude larger than the space occupied by a single nematode. The spatial effects of organisms many times smaller than a typical core, e.g., of a nematode, cannot be resolved

**Table 1** Order of magnitude of body diameter ranges for selected soil animals in relation to pore-size classes

| *Pore-size range* | *Faunal-size class* | *Organism* | *Order of magnitude of organism diameter (μm)* |
|---|---|---|---|
| Mesopores 1–100 μm | Microfauna | Ciliates | $10^0$–$10^1$ |
| | | Amebae | $10^0$–$10^1$ |
| | | Nematodes | $10^1$ |
| | Mesofauna | Collembola | $10^2$–$10^3$ |
| | | Mites | $10^2$–$10^3$ |
| Macropores >100 μm | | Insects | |
| | Macrofauna | Spiders | |
| | | Earthworms | $10^3$–$10^4$ |
| | | Millipedes | $10^3$–$10^4$ |

at that sampling scale. Finer-scale sample support has been used (e.g., 2.5 × 2.5 × 5 mm), particularly in the investigation of genetic structure of microbial communities.

Spatial dependence related to populations and metapopulations may be more evident because these occupy more space. Indeed, the distribution of soil organisms is often clumped, introducing characteristic spatial correlations that are related to life-history traits and migration patterns. For example, entomopathogenic nematodes release their young in insect larvae. Juvenile nematodes spread from these larval contagions, creating spatial patterns that at least initially show strong aggregation around the point of contagion.

Because organisms, even when they are of dissimilar scale, are connected via trophic or symbiotic interactions, the spatial distribution of one organism can affect the distribution of another population. The rhizosphere is defined as soil attached to plant roots. It is the focus of microbial and faunal activity due to rhizodeposits, enriching the food resources of microorganisms. Similarly, the soil surrounding earthworm burrows, known as the drilosphere, has elevated abundances of soil organisms because of detrital resources translocated from the surface. As a result of the elevated biological activity and resource enrichment, carbon and nitrogen mineralization rates are enhanced.

The activities of organisms alter soil properties beyond their immediate neighborhood, thus affecting other organisms. For example, consecutive microbial transformations, e.g., ammonification followed by nitrification, may occur in different locations because molecules move from the transformation sites by diffusion or by bulk movement of water. The diffusion of substrates may link points in space, with different microorganisms acting locally. The relationship of processes at these points may not be immediately apparent, because they occur in different locations and at different times and are affected by microbial populations with different functions.

Migration of organisms introduces another scale of spatial dependence. Migration distances of microorganisms, micro-, and mesofauna are limited by the distribution of soil air and water, as well as resource distribution and transport (phoresis) by other organisms. Aquatic microfauna, soil animals that live in water-filled pores (e.g., nematodes and protozoa), and their aerial counterparts, soil animals that live in the air-filled pores (e.g., mites, Collembola), depend on connected water-filled or air-filled spaces, respectively. Unless macrofauna extend the dispersal distances, microfauna and mesofauna migrate distances up to several decimeters.

Resource distribution directs motion of chemotactic soil animals such as entomopathogenic and bacterivorous soil nematodes; but even in the absence of chemotactic adaptations, resource distribution affects the distribution of consumers. Thus, the range of motion of an organism depends on a complex set of factors such as soil type, water status, distribution of resources, and whether the organism has particular traits for finding its prey or grazing substrates. In addition, many organisms and microbiological transformations are sensitive to abiotic constraints such as pH, texture, moisture, substrate, and $O_2$ availability. For example, C mineralization depends on water content and quality, temperature, organic matter content, bulk density, and texture. Nematode abundance depends on similar factors. The actual scale of potential motion may thus be secondary to other factors.

Some studies aim at characterizing spatial distributions at an even smaller scale. Such studies may be conducted to investigate the distribution of organisms or soil microbial processes at aggregate scale, or genetic diversity at subaggregate scales. To investigate the potential for denitrification in unsaturated soils, microelectrodes have been used to map the distribution of oxygen availability in individual soil peds. Microbial diversity measured using genetic techniques has been conducted at the subaggregate scale (2.5 × 2.5 × 5 mm). Studies conducted at these scales are important for the elucidation of mechanisms which contribute to biodiversity, e.g., the probability of genetic exchange during bacterial encounters, or to a mechanistic understanding of biogeochemical processes.

## Quantification of Spatial Properties

### First-Order Statistics and Spatial Independence

In soil science, both first- and second-order statistics are used to quantify spatial variability. Typical measures of first-order statistics are mean and standard deviation. Because of the patchiness of biological properties, distributions are skewed and the coefficient of variation is often high. For this reason, lognormal transformations or nonparametric techniques are used for inferential analysis of data. Coefficients of variation (CV) of greater than 150% have been reported for abundance of soil fauna. For enzymatic activity, CVs can be as low as 30% for phosphatase and protease, but as high as 130% for urease. CVs of several hundred percent have been reported for microbial transformation rates.

Mean and variance are used beyond inferential statistics. For example, the Taylor power law relates sample mean, $\mu$, and variance, $\sigma$:

$$\sigma^2 = a\mu^b \qquad [1]$$

where b is a measure of aggregation or 'clumpiness' of the distribution, and *a* is a scale factor. The Taylor power law is often invariant for a given population over a wide range of spatial scales. Tests of the invariance of the law might include rhizosphere scale, field scale, and farm scale. Invariance of the law has been interpreted to mean that the distribution of the population depends only on the dispersal traits of the organism, not on spatial constraints. Its scale invariance can be helpful, because, once *a* and *b* are known at one scale, the constants can be applied at another scale. The invariant properties of the Taylor power law are useful in the design of efficient sampling schemes for soil organisms and in the control of plant-parasitic nematodes.

### Second-Order Statistics and Spatial Dependence

First-order estimates are appropriate when soil properties at all sample points are spatially independent. However, the variance of most properties has both random (spatially independent) and deterministic (spatially dependent) components. Total sample variance, $\sigma_T^2$, can be expressed as:

$$\sigma_T^2 = \sigma_R^2 + \sigma_D^2 \qquad [2]$$

where the variances associated with random and deterministic variations are $\sigma_R^2$ and $\sigma_D^2$, respectively. Deterministic variance includes components such as trends, periodicities, and characteristic patch sizes. Organism size and its range of motion as well as physicochemical scales may contribute to the deterministic variations of biological soil properties. Care has to be taken when using the paradigm expressed in Eqn [2]. The perception of deterministic and random variations in soil may depend on sampling or investigative scale, when multiple scales are contributing to spatial variations. Increasing the scale of an investigation may change the ratio of deterministic to random contributions.

The assumption – inherent in first-order statistics – that samples taken from different locations are independent of each other is not valid at all scales. Second-order analyses are more appropriate when spatial dependence in the data is suspected. In soil science, the main tool to parameterize spatial dependence is semivariance analysis. For this analysis, data needs to satisfy the stationarity requirements which prohibits trends. For example, denitrification increases along a hydrosequence from dry, well-drained soils to poorly drained soils. Hence, semivariance analysis of denitrification along the entire hydrosequence would violate the stationarity requirement. Trends are easily detected and can be removed from the data prior to analysis.

Semivariance, $\gamma(h)$, measures how much of the variance of a spatial distribution can be explained by the relationship between points that are a given distance, *h*, apart:

$$\gamma(h) = 1/(2N(h))\Sigma(x_i - x_j)^2 \qquad [3]$$

where $x_i$ and $x_j$ are all point pairs that are a distance *h* apart, and *N* is the number of pairs ($x_i$, $x_j$) separated by distance *h*. A semivariogram is constructed in two steps. First, the covariances of point pairs separated by particular lag distances, *h*, are computed and plotted against the lag distance (**Figure 2**). Then, a regression model is fitted to the covariance data to estimate three parameters: the range lag (*R*), the nugget variance (Co), and the spatial variance (C) (**Figure 2**). Co is given by the *y* intercept of the regression model. The nugget value estimates the compound variance associated with sampling too few points, analytical errors, and random variations. The sum of nugget and spatial structure variance is called the sill variance, C + Co. The sill variance estimates variance at lags beyond the range lag, *R*, where sample pairs become spatially independent. The range lag, *R*, can be interpreted as the characteristic scale of a process.

The ratio C/(C + Co) is called spatial structure and expresses how much of the total variance is attributable to spatial dependence. If the nugget is equal to the sill variance, the data have no spatial structure; it is not spatially dependent. Conversely, if the nugget is zero, spatial structure is high and spatial correlation between points is high.

**Figure 3** shows three examples of variograms relating to microbial biomass N (**Figure 3a**), soil respiration (**Figure 3b**), and predaceous nematode (**Figure 3c**) data collected from an old field in southern New England, USA. The three variograms have distinctly different shapes. The variogram for microbial

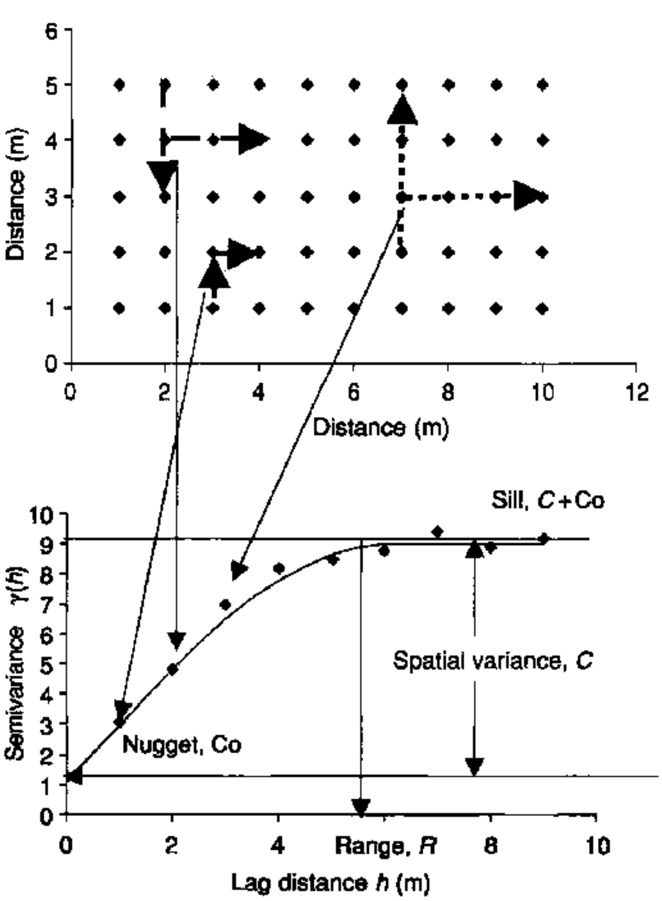


**Figure 2** Examples of point pairs separated by distances of 1 m (thin arrows), 2 m (dashed arrows), 3 m (dotted arrows). Semivariance values are computed using Eqn [3] based on the entirety of point pairs for each lag distance, *h*. Nugget variance is the semivariance, Co, at zero predicted by a model line. Sill variance is the sum of nugget, Co, and spatial variance, *C*.

biomass N displays strong spatial dependence from 0 to 8 m, expressed by the gradual increase in semivariance from zero to the sill variance (Figure 3a). In contrast, the semivariogram of soil respiration has no spatial structure. Semivariance values are scattered about the sample variance, indicating that the nugget and the sill variances are identical and that there is no spatial correlation of respiration for the sample distance and the sample support (Figure 3b). The semivariogram for predaceous nematodes has an additional feature (Figure 3c): after the semivariance reaches the sill, its value drops again. This is known as the 'hole effect.' Semivariance holes are caused by strongly repetitive or patchy patterns. Both range and hole lag distances are characteristic scales of the pattern.

## Sampling to Characterize Spatial Variability

Sampling schemes employed to collect soil samples for first- and second-order statistics are subject to specific sampling rules. Under the assumption that

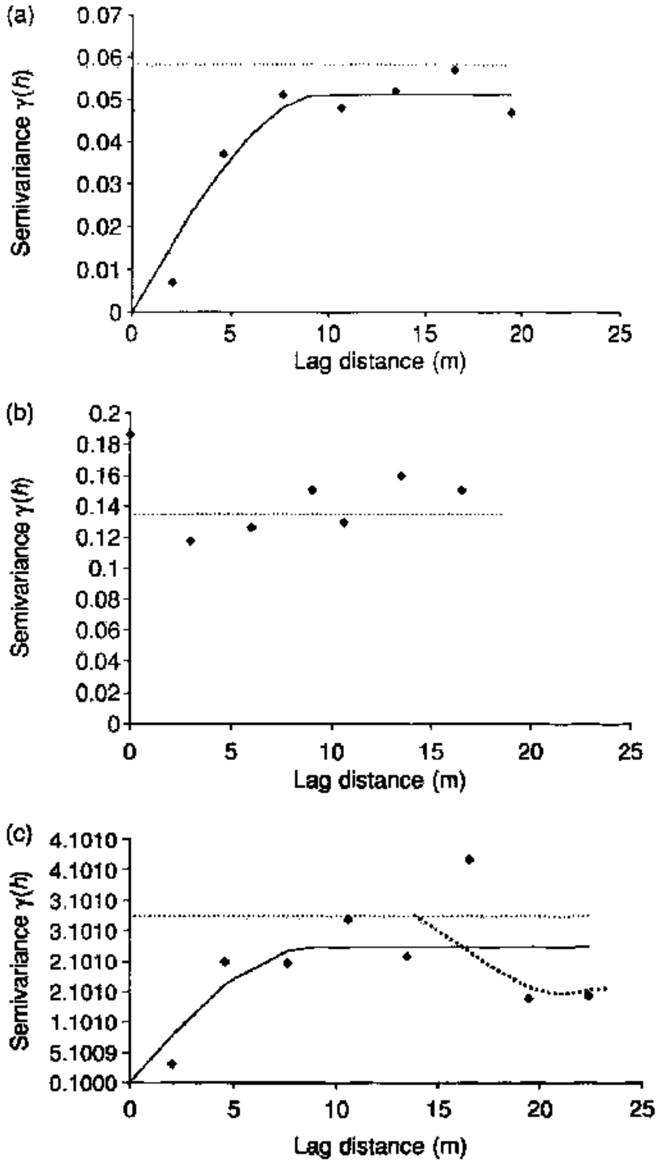


**Figure 3** Typical semivariogram shapes of biological properties. Rhomboid, covariance at lag *h*; continuous line, fitted spherical model; dashed line, sample variance; dotted line, hole effect. (a) Biomass N; (b) soil respiration; (c) predaceous nematodes.

samples are independent of each other, random sampling can be employed, and the design is guided by the central-limit theorem and considerations of accuracy and confidence.

If spatial dependence is assumed, sampling schemes need to fulfill an additional criterion. In theory, sampling ought to comply with the Nyquist theorem to avoid misrepresentation of patterns, which is also called aliasing. The Nyquist theorem states that a pattern must be sampled at a frequency of greater than twice the wavelength of the shortest variation that can be resolved. Sampling design needs to trade off the physical dimension, or sample support, of each sample and the number of samples collected. If the sampling design does not comply with the Nyquist

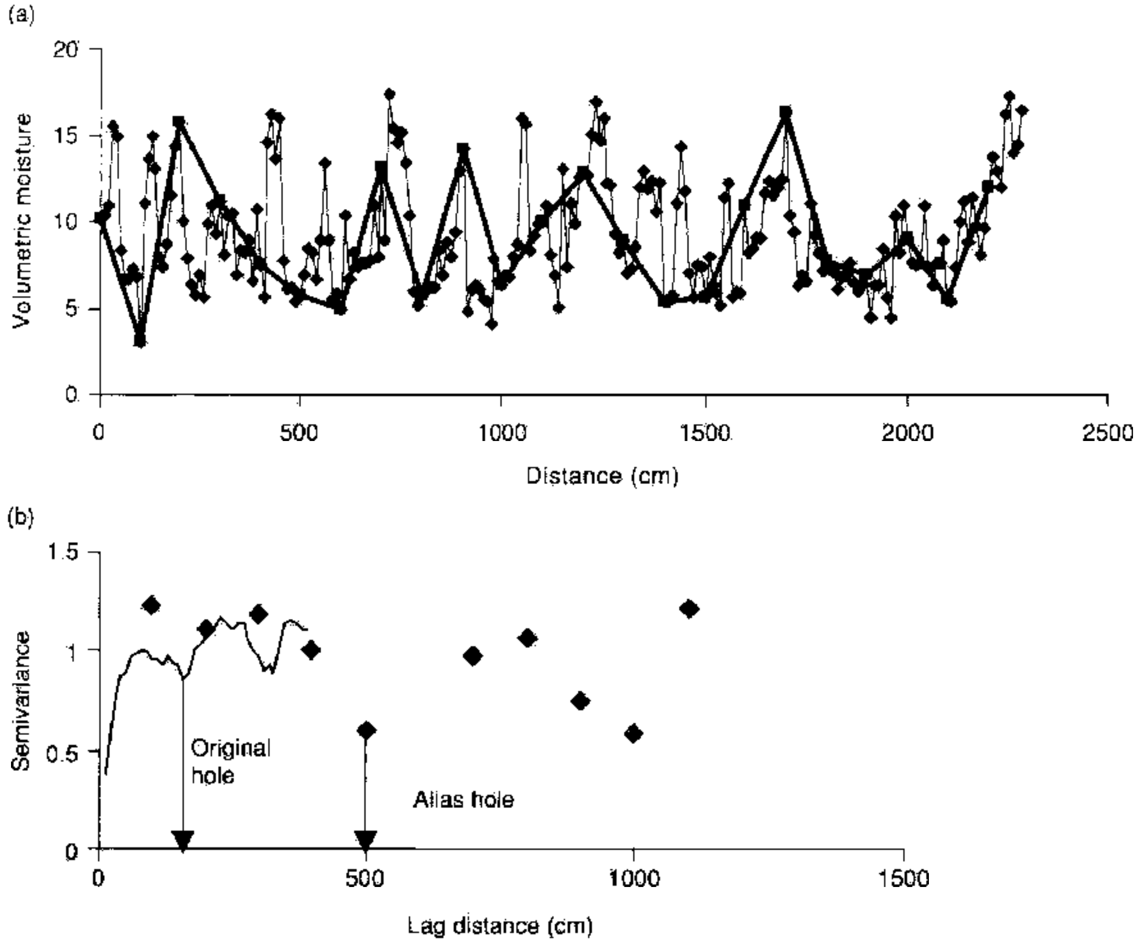


**Figure 4** Sampling at an insufficient rate may result in aliasing, i.e., the misrepresentation of a fine-grain regular pattern by a coarser variation: (a) Moisture data of field transect with a ridge-and-furrow pattern of 1.5 m. Thin line with rhomboids, sampled at intervals of 10 cm (entire data set); thick line, sampled at 100-cm intervals (subsample); (b) semivariance: (continuous line) semivariogram of transect data sampled at 10 cm; rhomboid semivariance values for transect sampled every 100 cm. The 'hole effect' occurs at 140 cm when the entire data set is analyzed; it occurs at 500 cm when a transect is sampled every 100 cm.

theorem the field pattern may be misrepresented by an alias pattern, which usually overestimates the wavelength of the shortest variation. Aliasing may affect mean, variance, and scale estimates.

An example of aliasing is shown in **Figure** 4. Volumetric soil moisture was determined using time-domain reflectometry at 220 points along a field transect, oriented normal to the plow direction (**Figure 4a**, thin line). Sampling points were 10 cm apart. The associated semivariogram displays a hole at 1.4 m, roughly the distance between plough ridges (**Figure 4b**). A subsample was selected from the data sets, with point separations of 1 m (**Figure 4a**, thick line). This sample does not satisfy the Nyquist theorem (1.4 samples per small-scale variation, rather than the more-than-two samples required to satisfy the theorem). As a result, a long-wavelength variation is erroneously detected which finds expression in the semivariogram as a dilation of the hole lag distance. The larger variation seen in the subsample is called an 'alias' of the original data.

Nested sampling schemes are used to give consideration to the multiscale nature of soil variability. Nested sampling distributes sampling at a primary coarse scale with a finer-scale grid being applied locally at the coarse grid, yielding data at two dissimilar scales.

Examples of nested sampling schemes are shown in **Figure** 5. The nested, stratified random sampling scheme shown in **Figure 5a** was used to compare the spatial dependence of nematode distributions in soils from a forest and an old field. Plots (17 × 17 m) were divided into 25 subplots. For each subplot at least one sample was taken from a random location. In five randomly selected subplots, nine regularly spaced samples were taken from a 30 × 30 cm area surrounding the random location (**Figure 5a**). The range over which C mineralization was spatially correlated was dominated by nest-scale variations in the autumn, and by the large-scale variations for winter, spring, and summer sampling. Range values for nematode distributions and moisture indicate spatial dependence

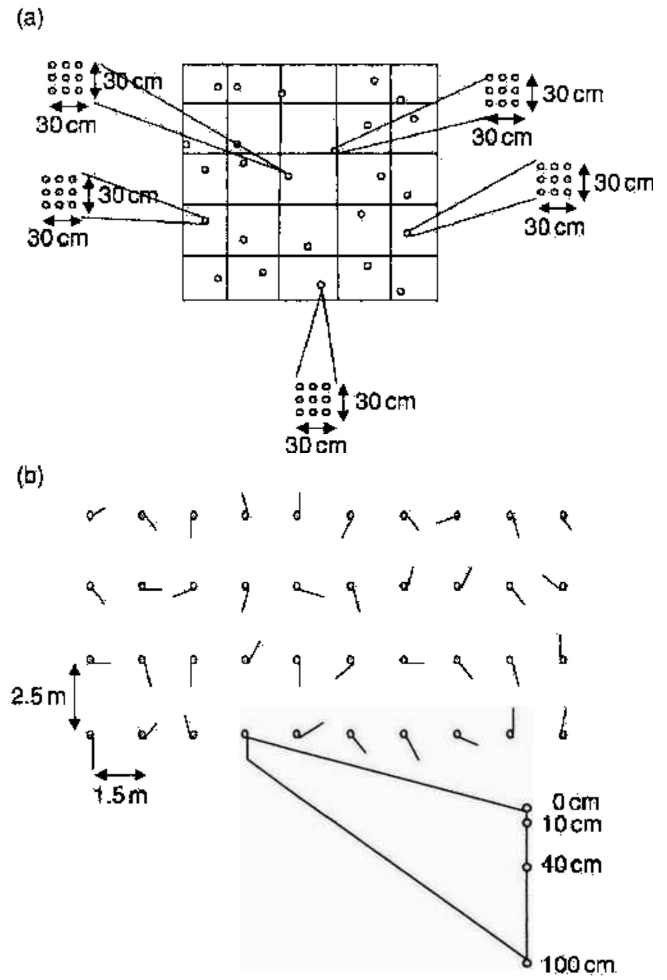


**Figure 5** Examples of nested-sampling designs: (a) Nested sampling used for quantifying spatial dependence of nematode, $CO_2$, and soil-moisture distribution. (Adapted from Görres JH, Dichiaro MJ, Lyons JB, and Amador JA (1998) Spatial and temporal patterns of soil biological activity in a forest and an old field. *Soil Biology and Biochemistry* 30: 219–230 with permission.) (b) Nested sampling design to quantify spatial dependence of $NO_3$ availability. (Adapted from Gross KL, Pregitzer KS, and Burton AJ (1995) Changing scales of nitrogen availability in early and late successional plant communities. *Journal of Ecology* 83: 357–367, with permission.)

**Table 2** Range values determined from semivariogram analysis of C mineralization, soil moisture, and total nematode abundance distributions in an old field and a forest soil using nested sampling

| *Semivariogram range* | | | | | | |
|---|---|---|---|---|---|---|
| | *Total nematodes (m)* | | *C Mineralization (m)* | | *Soil moisture (m)* | |
| *Date* | *Old field* | *Forest* | *Old field* | *Forest* | *Old field* | *Forest* |
| 5 Jun 1994 | 2.9 | 1.0 | 0.9 | 5.6 | 14.0 | 4.9 |
| 4 Nov 1994 | 2.8 | 10.3 | 0.3 | 0.2 | 10.0 | >20.0 |
| 20 Jan 1995 | 3.2 | 1.1 | 1.4 | 3.5 | 2.8 | 2.6 |
| 25 Mar 1995 | 6.0 | 2.6 | 0.9 | 0.9 | 6.7 | 1.3 |

at a larger scale than for $CO_2$ for most sampling dates (Table 2).

The nested sampling scheme in **Figure 5b** was used to detect multiscale variations in soil nutrient availability. A primary grid of a field consisted of 40 points spaced at 2.5-m spacings in the N-S direction and at 1.25 m in the E-W direction. A 1-m-long transect originated at each grid point. Each transect was oriented at random and sampled at distances of 0, 10, 40, and 100 cm along the transect, giving the pattern in **Figure 5b**.

Nested sampling schemes are useful in discerning variability at different scales, an important step in elucidating spatial variability. Usually, only two scale components are investigated, and it is difficult to assess whether these are the only levels of nesting present. Detecting several levels of a spatial hierarchy may not be practicable because of the cost of collecting and analyzing many samples, a prerequisite in many spatial analyses.

## Models of Spatial Variability

### Linear Statistical Models

Statistical descriptions of spatial distributions in themselves do not explain the cause of variability. Attempts to attribute biological effects to physical, chemical, and biological factors with linear models are necessarily incomplete. Typically, variance of biological distributions is explained to approximately 50% even when a multitude of factors are used in the model. The difference between measured variance of a soil biological property and a multivariate model may be caused by analytical error, nonlinear effects of population dynamics, dissimilar scales between the variables, and the absence of important factors from the model.

### Physical Mechanistic Process Models

Decision-making in environmental management is sometimes guided by mechanistic models describing processes with chemical, physical, or biological mechanisms. Mechanistic models are used to estimate the contamination of groundwater and surface water by nutrients and pesticides. They require information about soil physical structure, decomposition, and mineralization rates at the scales of interest.

Most mechanistic models simulate soil biological processes at a point. A point may represent a landscape unit, field, or forest stand. Usually, biological processes are treated as black boxes, while physical concepts in the model have strong mechanistic components. Some models of physical mechanisms use multiple data layers as spatial input. The data required for the analysis are frequently found in and extracted from environmental data bases such as geographic information systems (GISs). Typically, simulations are run for every point independently, but connections between points can be made by including the contribution of groundwater or surface-water flow to simulated contaminant inputs.

### Biological Mechanistic Process Models

Life-history traits and motility affect spatial population characteristics and can often be described by random processes such as the Neyman–Scott point process, which assumes that offspring disperses randomly from contagions, e.g., mother or parasitized organisms, which are themselves randomly distributed. The net effect of a Neyman–Scott process is a clumped distribution. Different levels of aggregation are shown in **Figure 6**. A simple random process (Figure 6a) distributes organisms almost evenly across an area. Any clumping here is a result of chance. In the double random Neyman–Scott process, the appearance of the distribution depends on the initial distribution pattern of the contagions and the distance that offspring could have traveled at the time of sampling. Both **Figure 6b** and c show Neyman–Scott processes, with dispersion distances of offspring of 7.5 m and 15 m, respectively.

An example of a Neyman–Scott process is that of endoparasitic nematodes, i.e., those that live within a root. Many endoparasites lay all their eggs in a single location, from where nematode larvae migrate, resulting in a clumped distribution. Aggregations arising from doubly random processes can have distinct spatial characteristics that increase spatial structure measured by semivariogram analysis even though they were caused by random processes.

## Applications of Spatial Analysis

Environmental scientists and managers have become more aware of soil variability at different scales. Managers may use spatial information at landscape or field scale. Many data sets which are used to manage the environment are stored as spatially referenced information within GISs that have become important databases and analysis tools supporting decision-making and research. For instance, mapping of critical loads for soils has supported strategies to reduce soil acidification in Europe. The critical-loads approach requires several data layers pertaining to spatially variable soil biological processes. For nitrogen deposition, critical loads use spatial information of denitrification, plant uptake, and immobilization. Critical-load maps are essential to estimating where depositions of nitrogen and sulfur exceed the ability of the soil to buffer acidification and thus to taking steps to target specifically reductions in acid emissions.

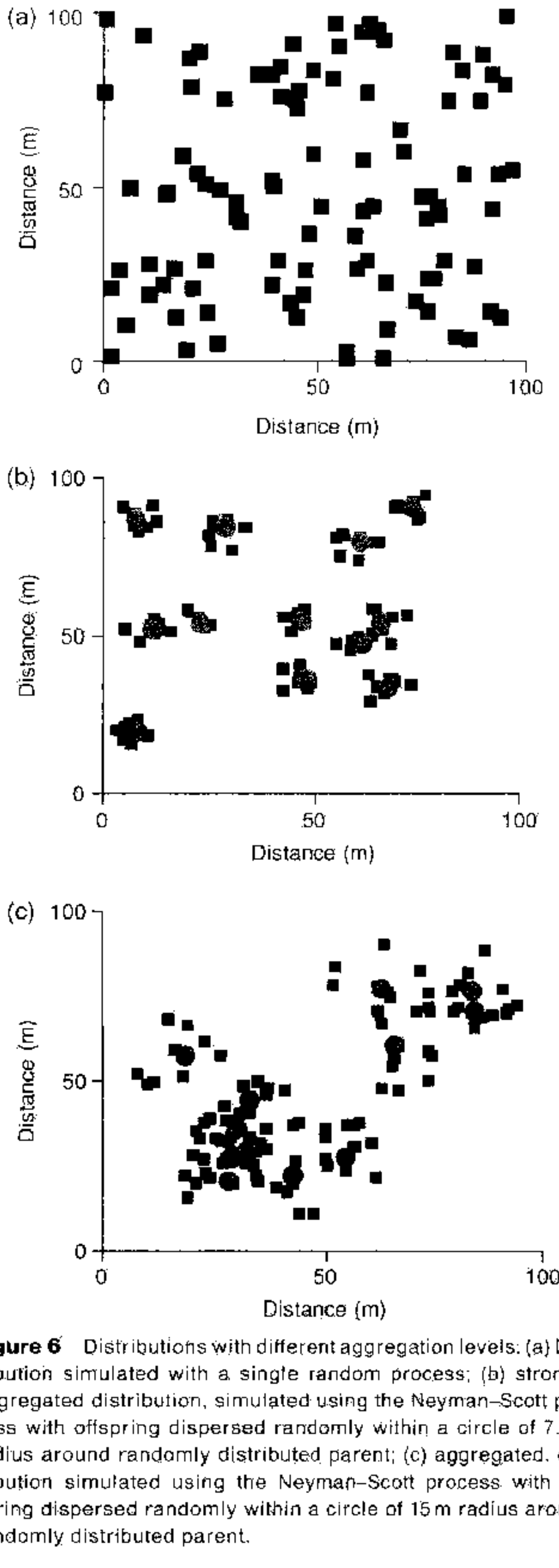


**Figure 6** Distributions with different aggregation levels: (a) Distribution simulated with a single random process; (b) strongly aggregated distribution, simulated using the Neyman–Scott process with offspring dispersed randomly within a circle of 7.5 m radius around randomly distributed parent; (c) aggregated distribution simulated using the Neyman–Scott process with offspring dispersed randomly within a circle of 15 m radius around randomly distributed parent.

Against the backdrop of the large amount of information available in spatially referenced form, first- and second-order statistics parameters of soil variability allow managers to estimate uncertainty associated with mean property estimation. In turn, this enables the manager to assess the risk inherent in any management action informed by measures of central tendency. Farm managers using precision agriculture or integrated pest management use the spatial distribution of soil properties explicitly to dose agrochemical inputs to minimize environmental impacts while increasing farm productivity.

At smaller scales, information on the spatial variability of soil and vegetation properties can guide experimental or survey designs to support mechanistic studies of ecosystem processes. In this way, the power of inferential tests can be increased. In fundamental research addressing questions of soil biodiversity, knowledge of the spatial structure of small habitat units may help estimate the frequency or the chance of genetic exchange between microorganisms. Likewise, knowledge of how habitats of soil-dwelling predators and prey or of grazers and forage overlap can help estimate the flow of energy and matter through soil food chains.

Differences in space may be seen as gradients that drive ecologic processes such as the flow of materials in an ecosystem. As such the application of spatial analysis may both further knowledge of ecosystem functioning and improve environmental quality.

## Summary

Soil biological processes are spatially variable. Many interacting factors together explain this variability at multiple scales. However, often spatial statistical descriptions of heterogeneity cannot explain the variation. Biological and physicochemical models are used to improve understanding of soil variability. Knowledge of the distribution of soil biological processes at point, field, and landscape scale can support management of environmental quality and ecosystem productivity as well as fundamental research.

*See also:* **Geographical Information Systems**; **Scaling:** Physical Properties and Processes; Transport Processes; **Spatial Variation, Soil Properties**; **Statistics in Soil Science**

## Further Reading

Amador JA, Wang Y, Savin MC, and Görres JH (2000) Fine-scale spatial variability of physical and biological soil properties in Kingston, Rhode Island. *Geoderma* 98: 83–94.

Bonmati M, Ceccanti N, and Nanniperi P (1991) Spatial variability of phospharase, urease, protesase, organic carbon and total nitrogen in soil. *Soil Biology and Biogeochemistry* 23: 391–396.

Bull KR, Brown MJ, Dyke H *et al.* (1995) Critical loads for nitrogen deposition for Great Britain. *Water, Air and Soil Pollution* 85: 2527–2532.

Burrough PA (1983) Multiscale sources of spatial variability in soil. I. The application of fractal concepts to nested levels of soil variation. *Journal of Soil Science* 34: 577–597.

Ferris H, Mullens TA, and Foord KE (1990) Stability and characteristics of spatial description parameters for nematode populations. *Journal of Nematology* 22: 427–439.

Groffman PM and Tiedje JM (1989) Denitrification in north temperate forest soils: relationship between denitrification and environmental factors at the landscape scale. *Soil Biology and Biogeochemistry* 21: 621–626.

Gross KL, Pregitzer KS, and Burton AJ (1995) Changing scales of nitrogen availability in early and late successional plant communities. *Journal of Ecology* 83: 357–367.

Isaaks EH and Srivastava RM (1989) *Applied Geostatistics*. Oxford, UK: Oxford University Press.

Kampichler C (1999) Fractal concepts in studies of soil fauna. *Geoderma* 88: 283–300.

McBratney AB and Pringle MJ (1999) Estimating average and proportional variograms of soil properties and their potential use in precision agriculture. *Precision Agriculture* 1: 125–152.

Parkin TB (1993) Spatial variability of microbial processes in soil – a review. *Journal of Environmental Quality* 22: 409–417.

Robertson GP and Freckman DW (1995) The spatial distribution of nematode trophic groups across a cultivated ecosystem. *Ecology* 76: 1425–1432.

Smith JL, Halvorson JJ, and Bolton H, Jr (1994) Spatial relationship of soil microbial biomass and C and N mineralization in a semi-arid shrub-steppe ecosystem. *Soil Biology and Biogeochemistry* 26: 1151–1159.

Vogel J, Normand P, Thioulouse J, Nesme X, and Grundmann GL (2003) Relationship between spatial and genetic distance on *Agrobacterium* spp. in 1 cubic centimeter of soil. *Applied and Environmental Microbiology* 69: 1482–1487.

Webster R and Oliver MA (2001) *Geostatistics for Environmental Scientists*. Chichester, UK: John Wiley.

Wirth SJ (2001) Regional-scale analysis of soil microbial biomass and soil basal $CO_2$-respiration in Northeastern Germany. In: Stott DE, Mohtar RH, and Steinhardt GC (eds) *Sustaining the Global Farm*, pp. 486–493. 10th International Soil Conservation Organization Meeting, 24–29 May 1999. West Lafayette, IN: Purdue University and the USDA–ARS National Soil Erosion Laboratory.

ISBN 0-12-348530-4

9 780123 485304

ELSEVIER
ACADEMIC
PRESS

elsevier.com

VOLUME 3 ISBN 0-12-348533-9
SET ISBN 0-12-348530-4