



LAND CARBON CYCLE MODELING

Matrix Approach, Data Assimilation,
& Ecological Forecasting

EDITED BY

Yiqi Luo

Benjamin Smith



CRC Press
Taylor & Francis Group

Land Carbon Cycle Modeling



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Land Carbon Cycle Modeling

Matrix Approach, Data Assimilation, & Ecological Forecasting

Edited by
Yiqi Luo
Benjamin Smith



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

First edition published 2023
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

© 2023 Taylor & Francis Group, LLC

Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

The Open Access version of this book, available at www.taylorfrancis.com, has been made available under a Creative Commons Attribution-Non-Commercial-No Derivatives 4.0 license.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Names: Luo, Yiqi, editor. | Smith, Benjamin, editor.

Title: Land carbon cycle modeling : matrix approach, data assimilation, & ecological forecasting / edited by Yiqi Luo and Benjamin Smith.

Description: First edition. | Boca Raton : CRC Press, 2022. | Includes bibliographical references and index. | Summary: "Carbon moves through the atmosphere, through the oceans, onto land, and into and between various ecosystems. This cycling has a large effect on climate - changing geographic patterns of rainfall and the frequency of extreme weather. The impact of changes in global carbon cycling are altered as the use of fossil fuels add carbon to the cycle. This book addresses the crucial question of how to assess, evaluate, and estimate the potential impact of the additional carbon to the global carbon cycle. The contributors describe a set of models for exploring ecological questions regarding changes in carbon cycling; provide background for developing new models; employs data assimilation techniques for model improvement; and do real- or near-time ecological forecasting for decision support. This book strives to balance theoretical considerations, technical details, and applications of ecosystem modeling for research, assessment, and crucial decision making"-- Provided by publisher.

Identifiers: LCCN 2021048046 (print) | LCCN 2021048047 (ebook) | ISBN 9781498737012 (hardback) | ISBN 9780367762421 (paperback) | ISBN 9780429155659 (ebook other)

Subjects: LCSH: Carbon. | Carbon cycle (Biogeochemistry)

Classification: LCC QE516.C37 L66 2022 (print) | LCC QE516.C37 (ebook) | DDC 577/.144--dc23/eng/20211130

LC record available at <https://lcn.loc.gov/2021048046>

LC ebook record available at <https://lcn.loc.gov/2021048047>

ISBN: 978-1-4987-3701-2 (hbk)

ISBN: 978-0-367-76242-1 (pbk)

ISBN: 978-0-429-15565-9 (ebk)

DOI: 10.1201/9780429155659

Typeset in Joanna
by SPi Technologies India Pvt Ltd (Straive)

CONTENTS

Preface / ix
Notes on the Editors / xiii
Contributors / xv

Unit I • Fundamentals of Carbon Cycle Modeling / 1

- 1 • THEORETICAL FOUNDATION OF THE LAND CARBON CYCLE AND MATRIX APPROACH / 3
Yiqi Luo
- 2 • INTRODUCTION TO MODELING / 13
Benjamin Smith
- 3 • FLOW DIAGRAMS AND BALANCE EQUATIONS OF LAND CARBON MODELS / 23
Yuanyuan Huang
- 4 • PRACTICE 1: CARBON FLOW DIAGRAM AND CARBON BALANCE EQUATIONS / 31
Yuanyuan Huang

Unit II • Matrix Representation of Carbon Balance / 35

- 5 • DEVELOPING MATRIX MODELS FOR LAND CARBON MODELS / 37
Yuanyuan Huang
- 6 • COUPLED CARBON-NITROGEN MATRIX MODELS / 45
Zheng Shi and Xingjie Lu

7 • COMPARTMENTAL DYNAMICAL SYSTEMS AND CARBON CYCLE MODELS / 57

Carlos A. Sierra

8 • PRACTICE 2: MATRIX REPRESENTATION OF CARBON BALANCE EQUATIONS AND CODING / 65

Yuanyuan Huang

Unit III • Carbon Cycle Diagnostics for Uncertainty Analysis / 71

9 • UNIFIED DIAGNOSTIC SYSTEM FOR UNCERTAINTY ANALYSIS / 73

Yiqi Luo

10 • SENSITIVITY ANALYSIS WITH MATRIX EQUATIONS: A CASE STUDY WITH ORCHIDEE / 79

Yuanyuan Huang

11 • MATRIX PHOSPHORUS MODEL AND DATA ASSIMILATION / 87

Enqing Hou

12 • PRACTICE 3: DIAGNOSTIC VARIABLES IN MATRIX MODELS / 95

Xingjie Lu

Unit IV • Semi-Analytic Spin-Up (SASU) / 101

13 • NONAUTONOMOUS ODE SYSTEM SOLVER AND STABILITY ANALYSIS / 103

Ying Wang

14 • SEMI-ANALYTIC SPIN-UP (SASU) OF COUPLED CARBON-NITROGEN CYCLE MODELS / 115
Xingjie Lu and Jianyang Xia

15 • TIME CHARACTERISTICS OF COMPARTMENTAL SYSTEMS / 123
Carlos A. Sierra

16 • PRACTICE 4: EFFICIENCY AND CONVERGENCE OF SEMI-ANALYTIC SPIN-UP (SASU) IN TECO / 129
Xingjie Lu

Unit V • Traceability and Benchmark Analysis / 137

17 • OVERVIEW OF TRACEABILITY ANALYSIS / 139
Jianyang Xia

18 • APPLICATIONS OF THE TRANSIENT TRACEABILITY FRAMEWORK / 147
Lifen Jiang

19 • BENCHMARK ANALYSIS / 157
Yiqi Luo and Forrest M. Hoffman

20 • PRACTICE 5: TRACEABILITY ANALYSIS FOR EVALUATING TERRESTRIAL CARBON CYCLE MODELS / 163
Jianyang Xia and Jian Zhou

Unit VI • Introduction to Data Assimilation / 171

21 • DATA ASSIMILATION: INTRODUCTION, PROCEDURE, AND APPLICATIONS / 173
Yiqi Luo

22 • BAYESIAN STATISTICS AND MARKOV CHAIN MONTE CARLO METHOD IN DATA ASSIMILATION / 181
Feng Tao

23 • APPLICATION OF DATA ASSIMILATION TO SOIL INCUBATION DATA / 189
Junyi Liang and Jiang Jiang

24 • PRACTICE 6: THE SEVEN-STEP PROCEDURE FOR DATA ASSIMILATION / 197
Xin Huang

Unit VII • Data Assimilation with Field Measurements and Satellite Data / 207

25 • MODEL-DATA INTEGRATION AT THE SPRUCE EXPERIMENT / 209
Daniel Ricciuto

26 • APPLICATION OF DATA ASSIMILATION TO A PEATLAND METHANE STUDY / 215
Shuang Ma

27 • GLOBAL CARBON CYCLE DATA ASSIMILATION USING EARTH OBSERVATION – THE CARDAMOM APPROACH / 225
Mathew Williams

28 • PRACTICE 7: DATA ASSIMILATION AT THE SPRUCE SITE / 237
Shuang Ma

Unit VIII • Value of Data to Constrain Models and Their Predictions / 243

29 • INFORMATION CONTENTS OF DIFFERENT TYPES OF DATA SETS TO CONSTRAIN PARAMETERS AND PREDICTIONS / 245
Enqing Hou

30 • USING DATA ASSIMILATION TO IDENTIFY MECHANISMS CONTROLLING LAKE CARBON DYNAMICS / 255
Oleksandra (Sasha) Hararuk

31 • DATA-CONSTRAINED UNCERTAINTY ANALYSIS IN GLOBAL SOIL CARBON MODELS / 263
Zheng Shi

32 • PRACTICE 8: INFORMATION CONTENTS OF LAND CARBON POOL AND FLUX MEASUREMENTS TO CONSTRAIN A LAND CARBON MODEL / 273
Enqing Hou

Unit IX • Ecological Forecasting with EcoPAD / 285

33 • INTRODUCTION TO ECOLOGICAL FORECASTING / 287
Yiqi Luo

- 34 • **ECOLOGICAL PLATFORM FOR ASSIMILATING DATA (ECOPAD) FOR ECOLOGICAL FORECASTING / 293**
Yuanyuan Huang
- 35 • **PRACTICE 9: ECOLOGICAL FORECASTING AT THE SPRUCE SITE / 301**
Jiang Jiang

Unit X • Process-based Machine Learning and Data-driven Modeling (PRODA) / 307

- 36 • **INTRODUCTION TO MACHINE LEARNING AND NEURAL NETWORKS / 309**
Toby Dylan Hocking
- 37 • **PROCESS-GUIDED DEEP LEARNING AND DATA-DRIVEN MODELING (PRODA) / 319**
Feng Tao and Yiqi Luo

- 38 • **PRACTICE 10: DEEP LEARNING TO OPTIMIZE PARAMETRIZATION OF CLM5 / 329**
Feng Tao

Appendices

- Appendix 1 Matrix Algebra in Land Carbon Cycle Modeling / 337
Ye Chen
- Appendix 2 Introduction to Programming in Python / 343
Xin Huang
- Appendix 3 CarboTrain User Guide / 353
Yuan Gao
- References / 363
- Index / 373



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

PREFACE

Yiqi Luo, School of Integrative Plant Science, Cornell University, Ithaca, USA

*Benjamin Smith, Hawkesbury Institute for the Environment,
Western Sydney University, Richmond, Australia*

The ecosystems of the vegetated land surface are critical to the health of the planet, its biodiversity, and people, providing benefits, so-called ecosystem services, on which human health, well-being and economic activity rely. The carbon cycle connects ecosystems on land to the atmosphere and the climate system. Rising temperatures, changing rainfall distributions, and more frequent extreme weather impact ecosystem services, often in a negative way. Rising atmospheric carbon dioxide concentrations – the main proximal cause of climate change – also affect ecosystems directly, enhancing photosynthesis and plant water-use efficiency in experimental settings, and almost certainly in nature. Due to a small imbalance between global photosynthesis, absorbing carbon dioxide from the atmosphere, and the return flux (mainly decomposition of litter and soil carbon) from the land to the atmosphere, a sizeable proportion of anthropogenic CO₂ emissions are reabsorbed by ecosystems – the land carbon sink. A key goal of international climate policy, to supplement fossil fuel reductions with negative emissions over a transitional period, largely relies on management interventions to preserve and enhance the land carbon sink. The land carbon cycle, then, is central both to mitigation (emissions reduction) and adaptation (management of climate impacts and risks) responses and policies. Developing effective measures requires predictions of how the system may be expected to respond under various scenarios. For this, of course, we need models.

While relevant to supporting climate science and policy, modeling is also gaining “market share” in environmental and ecological research generally. Several forces are involved in this trend. Ever increasing volumes of readily available data from observational and experimental networks are making it easier to parameterize and robustly validate models. Cheaper, more powerful computers, including cloud computing services, bring complex numerical algorithms and data assimilation methods within reach for many applications. New statistical and optimization methodologies have been developed and made accessible through convenient packages and toolboxes, useful not only to modelers but as platforms for collaboration between modelers and empirical scientists.

This book provides an overview of the current state of the land carbon cycle modeling field, exemplifying recent developments as described above. The book is built upon a summer training course, *New Advances in Land Carbon Cycle Modeling*, held annually since 2018 at Northern Arizona University. Over the four years the course has been offered, attendees from 32 countries in six continents have undertaken the training. The first training course in 2018 attracted about 40 participants, with a similar number in 2019. Due to the pandemic of COVID-19, the in-person course was replaced by an online version in 2020. Originally, we planned to have 25 attendees, but ended up with nearly 85 participants from six continents. This grew further to 150 virtual attendees in 2021.

This book is mainly based on 31 lectures (including pre-training lectures) and ten practices prepared for the *New Advances in Land Carbon Cycle Modeling* training course in 2020 and 2021.

The book offers cutting-edge knowledge and techniques on carbon cycle science and modeling. We have designed ten training units in such a way that everyone can gain regardless of their prior background in modeling. The chapters range from theoretical foundation of land carbon cycling, traceability, and data assimilation to machine learning and ecological forecasting. Overall, four techniques are covered: the matrix approach to land carbon modeling; data assimilation for data-driven modeling; ecological forecasting; and combined machine learning with data assimilation to improve model prediction.

The organization of the book aligns with the training course, which has two blocks. The first block in units 1–5 is about the matrix approach to land carbon cycle modeling. The second block in units 6–10 is on data assimilation, ecological forecasting, and machine learning.

The matrix approach introduced in units 1–5 first describes a matrix equation. It is demonstrated that the matrix equation can unify land carbon cycle models; offer a new theoretical framework to guide carbon cycle research; help accelerate computational efficiency for spin-up; offer new analytics to diagnose model performance; and allow data assimilation of complex models. Five skills are covered in units 1–5, namely: drawing the carbon flow diagram and writing carbon balance equations of a model; developing matrix models from carbon balance equations and coding the matrix model; adding diagnostic variables to matrix models; adding semi-analytic spin-up (SASU) algorithms; and traceability analysis.

The matrix equation can be used to derive three diagnostic variables: carbon input, residence time, and carbon storage potential. The matrix equation can also be used to get an analytic solution of steady-state pool sizes, leading to SASU. The matrix equation is the foundation for traceability analysis. The chapters of units 1–5 explain these new skills.

Units 6–10 cover data assimilation, ecological forecasting, and machine learning. To realistically forecast ecological responses to climate change, three elements all need to be perfectly aligned: model structure, model parameterization, and the external forcing variables. The matrix approach discussed in units 1–5 is about process-based

model structure. Data assimilation and machine learning in units 6–8 and 10 will help improve model parameterization. EcoPAD, a workflow system that is described in unit 9, will link real-time forcing to model forecasting.

Chapters in unit 6 describe the seven-step procedure of data assimilation. The seven steps are: defining a research objective; acquiring data sets; using one model; developing a cost function; minimizing mismatches between modeled and observed values with a global optimization method; estimating parameters; and predicting ecosystem responses. Chapters in units 7 and 8 are about applications of data assimilation to the SPRUCE field experiment and satellite observations, and evaluation of values of different data sets to constrain models and their predictions.

Chapters in unit 9 describe the Ecological Platform for Assimilating Data (EcoPAD) framework, which automatically ingests data into a model through a data assimilation system for ecological forecasting.

Chapters in unit 10 introduce machine learning, a PROcess-guided deep learning combined with DATA-driven modeling (PRODA) approach, and its application to optimize parameterization of the CLM5 land surface model.

Most of the chapters are written in such a way as to be understandable by readers with minimal modeling background. Practices are targeted at a suitable level for such readers. A few chapters may require some mathematical background to be fully understood. The book offers three appendix chapters on, respectively: basic linear algebra; introductory programming with Python; and the Carbon Training (CarboTrain) package we use as a toolbox for the training course and the practice chapter in each unit. Depending on their prior knowledge, readers may choose to read these appendix chapters as optional supporting material to the chapters in units 1–10.

All the chapters are accompanied with pre-recorded lectures or practice instruction. These pre-recorded videos are available at https://www2.nau.edu/luo-lab/download/4th_training_course.php. (Please search for the videos using “ecolab Yiqi Luo” if the website is moved away from NAU.) If you plan to master skills described in the chapters, you may find it useful to read the book chapter; listen to the corresponding video; take the quiz at the end of each chapter after watching the video; and attempt the practice chapter at the end of each unit, following the pre-recorded instruction.

Please be aware that this book does not teach programming or how to code a model, nor does it teach how to do model development or modification. Readers with limited programming experience may, however, find the brief introduction to Python coding in appendix 2 useful.

The open access electronic version of this book has been made available thanks to financial contributions by Northern Arizona University, Lund

University, and Oak Ridge National Laboratory. Finally, we wish to thank all 22 authors who have worked very hard for months to prepare the material for this book. We hope that you, the reader, find it useful and rewarding.

YIQI LUO,
BEN SMITH,
September, 2021.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

NOTES ON THE EDITORS

Yiqi Luo is a Regents Professor at Northern Arizona University, USA. He obtained his PhD degree from the University of California, Davis in 1991 and did postdoctoral research at UCLA and Stanford University from 1991 to 1994 before he worked at the Desert Research Institute as Assistant and Associate Research Professor from 1994 to 1998, and the University of Oklahoma as Associate, Full, and George Lynn Cross Professor from 1999 to 2017. Professor Luo has studied land carbon cycling using empirical and modeling approaches for more than 30 years. His research program has been focused on addressing two key issues: (1) how global change alters the structure and functions of terrestrial ecosystems, and (2) how terrestrial ecosystems regulate climate change. To address these issues, Dr. Luo's laboratory has conducted field global change experiments; developed terrestrial ecosystem models; synthesized extensive data sets using meta-analysis methods; integrated data and models with data assimilation techniques; and carried out theoretical and computational analysis. Particularly, his research team has recently developed a matrix approach to land carbon cycle modeling; applied data assimilation techniques to ecological research; and pioneered in ecological forecasting. Previously he has published two books, 30 book chapters, and more than 500 papers in peer-reviewed journals. He was a Highly Cited Researcher recognized by the Web of Science Group, Clarivate Analytics in 2018–2021. He was elected fellow of the American Association for

the Advancement of Science (AAAS) in 2013; the American Geophysical Union (AGU) in 2016; and the Ecological Society of America (ESA) in 2018.

This book, *Land Carbon Cycle Modeling: Matrix Approach, Data Assimilation, & Ecological Forecasting*, evolved from an international training course, *New Advances in Land Carbon Cycle Modeling*. The training course is supported by the US National Science Foundation and has been held four times from 2018 to 2021. The materials in the book have been used by approximately 300 attendees of the training course. Note that Yiqi Luo has recently moved to Cornell University, USA.

Benjamin Smith is a Professor of Ecosystem Science, based in Sydney, Australia where he is Research Director of Western Sydney University's Hawkesbury Institute for the Environment, a leading international center for global change ecosystem research and innovation. Following undergraduate studies in biology at the University of Tasmania, Australia, Ben relocated to Dunedin, New Zealand, where he obtained his PhD from the University of Otago in 1996. Following postdoctoral posts in Sweden and Germany, he obtained tenure at Lund University in Sweden, transitioning to his current role at Western Sydney University in 2018. Benjamin Smith is known as a pioneer in the dynamic global vegetation modeling field. The original developer of the widely used LPJ-GUESS ecosystem model, he continues to lead the multi-institutional consortium of

developers serving its international user community. As Distinguished Visiting Scientist with the CSIRO Oceans & Atmosphere Flagship, he contributed to the implementation of vegetation demography, disturbance and wildfire dynamics in the Australian Community Land Surface Model, CABLE. He is interested in the role of the biosphere in regional and global climate dynamics, using Earth system models to examine feedback of ecosystems and land surface changes to the

atmosphere and the climate system. He led the development of the first published regional Earth system model, RCA-GUESS, and is active in the pan European consortium developing the global EC-EARTH ESM. An author of several influential papers in biosphere modeling, carbon cycle and ecosystem impact assessment fields, Ben is recognized as being in the top one percent of the world's most cited researchers in the Clarivate Highly Cited Researchers list.

CONTRIBUTORS

YE CHEN

Northern Arizona University
Flagstaff, USA

YUAN GAO

Northern Arizona University
Flagstaff, USA

OLEKSANDRA (SASHA) HARARUK

University of Central Florida
Orlando, USA

TOBY DYLAN HOCKING

Northern Arizona University
Flagstaff, USA

FORREST M. HOFFMAN

Oak Ridge National Laboratory
Oak Ridge, USA

ENQING HOU

Northern Arizona University
Flagstaff, USA

XIN HUANG

Northern Arizona University
Flagstaff, USA

YUANYUAN HUANG

Climate Science Centre, CSIRO
Canberra, Australia

JIANG JIANG

Nanjing Forestry University
Nanjing, China

LIFEN JIANG

Northern Arizona University
Flagstaff, USA

JUNYI LIANG

China Agricultural University
Beijing, China

YIQI LUO

Cornell University
Ithaca, USA

XINGJIE LU

Sun Yat-sen University
Guangzhou, China

SHUANG MA

Northern Arizona University
Flagstaff, USA

DANIEL RICCIUTO

Oak Ridge National Laboratory
Oak Ridge, USA

CARLOS A. SIERRA

Max Planck Institute for Biogeochemistry
Jena, Germany

ZHENG SHI
University of Oklahoma
Norman, USA

BENJAMIN SMITH
Western Sydney University
Richmond, Australia

FENG TAO
Tsinghua University
Beijing, China

YING WANG
University of Oklahoma
Norman, USA

MATHEW WILLIAMS
University of Edinburgh
Edinburgh, UK

JIAN YANG XIA
East China Normal University
Shanghai, China

JIAN ZHOU
East China Normal University
Shanghai, China

UNIT ONE

Fundamentals of Carbon Cycle Modeling



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER ONE

Theoretical Foundation of the Land Carbon Cycle and Matrix Approach

Yiqi Luo

Cornell University, Ithaca, USA

CONTENTS

Convergence of the Land Carbon Cycle / 3
Donor Pool-Dominant Transfer and Other Properties That Govern the Land Carbon Cycle / 4
The Matrix Approach to Model Representation of the Land Carbon Cycle / 6
The Paradox of the Matrix Equation and Nonautonomous Systems / 10
Predictability of the Land Carbon Cycle / 11
Dynamic Disequilibrium of Land Carbon Cycle / 11
Suggested Reading / 12
Quizzes / 12

The land carbon cycle has been extensively studied yet its fundamental properties have not been fully understood. This chapter offers empirical evidence to demonstrate a general dynamic pattern that the land carbon cycle changes in a direction toward a moving attractor in response to global change. This general pattern is captured by a matrix equation. The relatively simple matrix equation can unify land carbon cycle models, accelerate computational efficiency for spin-up, diagnose model performance with new analytics, and enable data assimilation with complex models to improve their predictive skills, and guide carbon cycle research with a new theoretical framework of dynamic disequilibrium.

CONVERGENCE OF THE LAND CARBON CYCLE

In the late 2010s, a deserted village, Houtouwan, on an island off the east coast of China was discovered to be completely overrun by vegetation approximately 20 years after dwellers left (Smithsonian Channel, *The Abandoned Chinese Village that Nature Reclaimed*) (Figure 1.1). What might happen to the place in another 20 years or

even longer? It is likely that trees will gradually take over to form a coastal forest.

In fact, Mother Nature would overrun all urban places in the world if human disturbances were removed. Imagine that we could magically remove humans from any highly commercialized, heavily human-disturbed urban areas, such as Manhattan of New York City or Lujiazui of Shanghai, for 300 years: the place would soon be overrun by plants, animals and microbes. Without any human activities, small trees would grow from cracks in concrete in five years, most of the high-rise buildings would collapse and forests would probably take over in 50 years. In 300 years, Manhattan would most likely be occupied by a deciduous forest similar to those in northeastern USA, and Lujiazui of Shanghai by some lowland forests.

Similarly, it has been repeatedly observed how vegetation takes over landscapes after natural disturbances occur. For example, following the 1988 Yellowstone fires – massive blazes that burned about 1.2 million acres in and around Yellowstone National Park – their size and severity led to a proclamation that Yellowstone had been destroyed. The burned landscape was retaken by thriving young lodgepole



Figure 1.1. Overrunning of the deserted village of Houtouwan by vegetation approximately 20 years after dwellers left the island that is situated on off the east coast of China. Vegetation taking over after anthropogenic and natural disturbances are removed is caused by the internal processes of the carbon cycle that drive an ecosystem toward an attractor.

pine trees 30 years after the fires (Turner 2018). Secondary succession is an ecological term for this entirely natural process. Ecosystem succession has been extensively studied, mainly from the perspectives of species dynamics and community structures.

From a carbon cycle perspective, ecosystems converge toward some attractor states after anthropogenic and natural disturbances are removed. (Note that the states that ecosystems converge toward are moving attractors under global change. This point will be discussed later). This convergence is done by “Mother Nature” and actually results from the internal processes of the land carbon cycle. What, then, are those internal processes? And, how can we mathematically represent them?

DONOR POOL-DOMINANT TRANSFER AND OTHER PROPERTIES THAT GOVERN THE LAND CARBON CYCLE

Before we answer those questions, let us briefly review the land carbon cycle. Carbon enters an ecosystem via photosynthesis. Photosynthetic products are partly allocated for autotrophic respiration and partly for growth of leaf, stem, and root. Once a plant or its parts die, they become litter and enter the litter pools. Litter decomposes, partly released to the atmosphere via heterotrophic

respiration and partly incorporated into soil to become soil organic matter. Soil organic matter goes through decomposition and stabilization over and over again, driving soil carbon cycling. We will examine some of the processes to see what the best equation would be to represent them.

Let us first look at litterfall in which dead leaves fall from the tree canopy to the ground. To make it a subject of study, we define two new terms: donor pool and recipient pool. The donor pool donates litter whereas the recipient pool receives litter (Figure 1.2a). Litterfall is a rate process that moves carbon from the donor pool to the recipient pool. In this case, the rate of litter falling is proportional to the amount of litter in the donor pool while the amount of litter in the recipient pool has nothing to do with the rate of litterfall at all. Thus, the rate of litterfall is controlled by the donor pool.

The rate of litterfall, denoted by $dX(t)/dt$, equals the donor pool size, $X(t)$, times a coefficient (k) as:

$$\frac{dX(t)}{dt} = kX(t) \quad (1.1)$$

This equation describes the *donor pool-dominated carbon transfer* (Figure 1.2a). Litterfall in the real world is also affected by wind and other environmental conditions over seasons. Modelers usually use an environmental scalar, $\xi(t)$, to account for

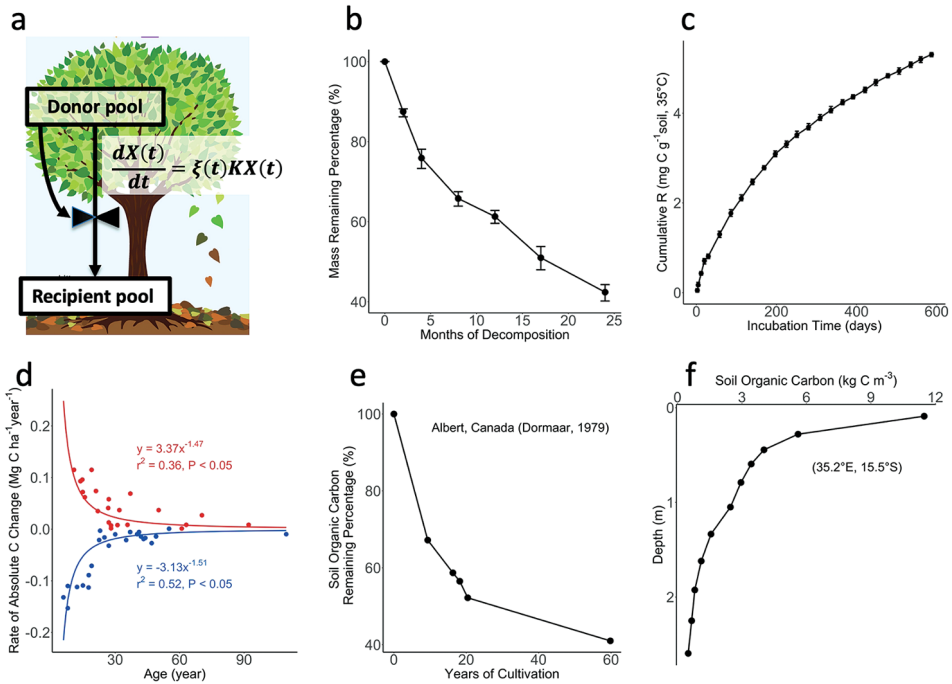


Figure 1.2. Macroscopic patterns of carbon transfer processes and distributions. The carbon transfer processes include (a) litterfall; (b) decomposition of ragweed (*Ambrosia psilostachya*) litter (Cheng et al. 2010); (c) soil organic carbon decomposition from incubation at 35 °C (replotting the data from Haddix et al. 2011); (d) rates of absolute carbon (C) change during the forest succession (Yang et al. 2011); (e) changes in soil organic matter after agricultural cultivation in Alberta, Canada (extracted data from Doomaar 1979); and (f) a vertical distribution of soil carbon with depth in Malawi, Africa, from a soil carbon database. The macroscopic patterns of almost all carbon transfer processes and distributions can be described by the first-order decay function.

the effects of phenology, wind and other environmental factors on litterfall as:

$$\frac{dX(t)}{dt} = \xi(t)kX(t) \quad (1.2)$$

Once litter has fallen on the ground, it decomposes. Litter decomposition is usually studied with litterbags or wood logs on ground or in air. Researchers initiate a study with a certain amount of litter in litterbags, place them in the field, then collect a subset of litterbags once every few weeks, find the dry weight, and calculate the dry weight remaining in comparison with the initial amount. A typical data set of litter decomposition shows that mass remaining becomes less and less as sampling time goes on (Figure 1.2b). This type of data set can usually be fitted by a first-order decay curve as in Equation 1.1. Thus, litter decomposition is often described by the same *donor pool-dominated carbon transfer* equation. In this case, k is often called the litter decay constant. Actually, k varies with litter type and location. Zhang

et al. (2008) synthesized nearly 300 datasets from 70 studies all over the world. The study found that Equation 1.1 fits all data sets very well although the value of k greatly varies with litter types and environment. Cai et al. (2018) synthesized more than 1,600 data sets of straw decomposition for six types of crops. Their study fits a three-exponent equation to all the data sets. Thus, straw decomposition also follows the *donor pool-dominated carbon transfer*.

Another important process of the carbon cycle is soil organic carbon (SOC) decomposition. SOC decomposition is usually studied by soil incubation. That is, researchers collect soil samples from the field and put the samples in jars for a period of time. They then collect gas samples once in a while to measure the amount of carbon released from the soil sample through microbial respiration. Data are usually plotted either by measured CO₂ release or cumulative CO₂ released on the y-axis with time on the x-axis (Figure 1.2c). Almost all data follow a similar pattern. This pattern also can be well described by *donor pool-dominated transfer*.

Schädel et al. (2013) fitted data of SOC decomposition with one-, two-, and three-pool models. They found that two or three-pool models work well for SOC decomposition. Schädel et al. (2014) synthesized more than 120 data sets from permafrost regions. Xu et al. (2016) synthesized nearly 400 data sets from different locations all over the world. Both of the latter studies found that all their data sets (more than 500) can be well fitted by two or three-pool models. This suggests that the donor pool-dominated carbon transfer equation also works for SOC decomposition.

Yang, Luo, and Finzi (2011) synthesized more than 124 studies of soil carbon dynamics at different stages of forest succession. During the course of secondary succession, some forests gain carbon and some lose carbon (Figure 1.2d). In either case, carbon dynamics still can be described by the donor-pool dominated carbon transfer. Moreover, the soil organic matter remaining after long-term cultivation (Figure 1.2e) and the vertical distribution of soil organic carbon with depth (Figure 1.2f) both follow monotonic patterns, consistent with the donor-pool dominated carbon transfer.

So far, we have examined macroscopic patterns of key carbon transfer processes (e.g., litterfall, litter decomposition, and SOC decomposition) and distributions. These macroscopic patterns are very typical as almost ubiquitously revealed by thousands of field and laboratory studies. The macroscopic patterns can be well described by the donor pool-dominated carbon transfer equation.

The donor pool-dominated carbon transfer, then, is one of the four fundamental properties that govern dynamics of the land carbon cycle. The other three properties are (1) photosynthesis as the primary carbon influx pathway, (2) compartmentalization of carbon processes into plant, litter, and soil, and (3) the first-order kinetics of carbon transfer from the donor pool (Luo and Weng 2011) (Equation 1.1). Among the four properties, the donor pool-dominated carbon transfer is the most important property in determining the trajectory of the land carbon cycle (Luo, Keenan, and Smith 2015). If this property is altered in our model, the carbon cycle will not behave as we have observed in the real world.

The four properties fundamentally characterize the internal processes of the land carbon cycle. The internal processes drive the land carbon cycle to converge toward an attractor (Luo et al. 2017). This is the reason why places like Manhattan could become deciduous forests and Liujiazui of

Shanghai could become lowland forests if human disturbances were removed. This convergence is applicable to almost any place on Earth.

Active research is going on to incorporate microbial processes and traits into carbon cycle models to account for the important role of microbes in catalyzing decomposition of soil organic matter. Many of the microbial models were developed on purpose because the responsible researchers suspected that models based on first-order kinetics of carbon transfer from donor pools were too simple. However, these newly developed, microbially-based models are usually nonlinear and do not fit the observed macroscopic patterns of litter and SOC decomposition well (e.g., Liang et al. 2018). It remains challenging that litter and SOC decomposition models not only adequately represent microbial processes but also are consistent with the macroscopic patterns observed from almost all experimental studies.

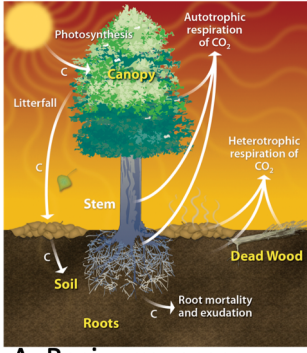
Next, we examine how well these four properties are represented in models.

THE MATRIX APPROACH TO MODEL REPRESENTATION OF THE LAND CARBON CYCLE

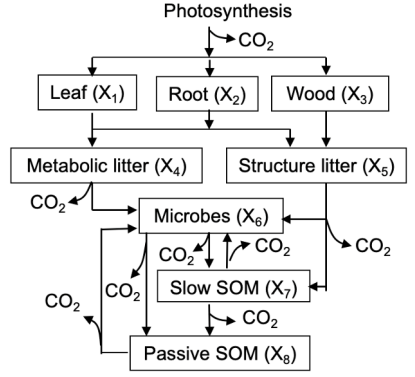
The four properties identified above are all well represented in models as long as the models use a so-called pool-and-flux or box-and-arrow structure. In this structure, we use pools to represent different carbon compartments and fluxes to represent carbon transfer among compartments or carbon input into and output out of the ecosystem. For this structure, we need one carbon balance equation to trace how much carbon gets into one pool and how much carbon leaves the pool. For example, the leaf pool, X_1 , receives carbon from photosynthesis partitioned to leaves and loses carbon by senescence (Figure 1.3). Thus, we need one equation to calculate the amount of carbon resulting from photosynthesis and the amount of carbon lost to litterfall. The equation to describe the dynamics of carbon balance in the leaf pool over time can be represented by:

$$\frac{dX_1(t)}{dt} = b_1\mu(t) - \xi(t)k_1X_1(t) \quad (1.3)$$

where $\mu(t)$ represents the amount of carbon input from net primary production (NPP, photosynthesis



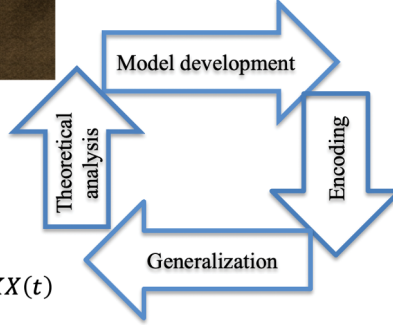
A: Basic processes



B: Shared model structure

D: General model

$$\frac{dX(t)}{dt} = B\mu(t) + A\xi(t)KX(t)$$



C: Similar algorithm

$$\begin{cases} \text{Plant} \begin{cases} dX_1(t)/dt = b_1\mu(t) - \xi(t)k_1X_1(t) \\ dX_2(t)/dt = b_2\mu(t) - \xi(t)k_2X_2(t) \\ dX_3(t)/dt = b_3\mu(t) - \xi(t)k_3X_3(t) \end{cases} \\ \text{Litter} \begin{cases} dX_4(t)/dt = \xi(t)[a_{41}k_1X_1(t) + a_{42}k_2X_2(t) - k_4X_4(t)] \\ dX_5(t)/dt = \xi(t)[a_{51}k_1X_1(t) + a_{52}k_2X_2(t) + k_3X_3(t) - k_5X_5(t)] \end{cases} \\ \text{SOM} \begin{cases} dX_6(t)/dt = \xi(t)[a_{64}k_4X_4(t) + a_{65}k_5X_5(t) + a_{67}k_7X_7(t) + a_{68}k_8X_8(t) - k_6X_6(t)] \\ dX_7(t)/dt = \xi(t)[a_{75}k_5X_5(t) + a_{76}k_6X_6(t) - k_7X_7(t)] \\ dX_8(t)/dt = \xi(t)[a_{86}k_6X_6(t) + a_{87}k_7X_7(t) - k_8X_8(t)] \end{cases} \end{cases}$$

Figure 1.3. A generalized matrix model of the terrestrial carbon cycle. (a) The basic carbon cycle processes are represented by four fundamental properties for all terrestrial ecosystems. (b) The four properties have been incorporated into terrestrial carbon cycle models with a pool-and-flux structure. (c) The structure is typically encoded using a set of balance equations with carbon input into and output from each pool. (d) The balance equations of terrestrial carbon cycle models can be converted to a matrix equation. Thus, the matrix equation can be considered as a general system equation (or a dynamical equation) for the terrestrial carbon cycle.

minus autotrophic respiration), b_1 is the carbon partitioning from NPP to leaves, k_1 is the rate of senescence, and $\xi(t)$ is an environmental modifier. Thus, the change in the carbon pool size in the leaf pool ($\frac{dX_1(t)}{dt}$) equals carbon input to the

leaf pool $b_1\mu(t)$ minus carbon leaving the leaf pool $\xi(t)k_1X_1(t)$.

Extending this idea to all the eight pools of the Terrestrial Ecosystem (TECO) model, we have eight carbon balance equations to track carbon cycling in the ecosystem as:

$$\left\{ \begin{array}{l} dX_1(t)/dt = b_1\mu(t) - \xi(t)k_1X_1(t) \\ dX_2(t)/dt = b_2\mu(t) - \xi(t)k_2X_2(t) \\ dX_3(t)/dt = b_3\mu(t) - \xi(t)k_3X_3(t) \end{array} \right\} \text{plant} \\ \left\{ \begin{array}{l} dX_4(t)/dt = \xi(t)[a_{41}k_1X_1(t) + a_{42}k_2X_2(t) - k_4X_4(t)] \\ dX_5(t)/dt = \xi(t)[a_{51}k_1X_1(t) + a_{52}k_2X_2(t) + k_3X_3(t) - k_5X_5(t)] \end{array} \right\} \text{litter} \\ \left\{ \begin{array}{l} dX_6(t)/dt = \xi(t)[a_{64}k_4X_4(t) + a_{65}k_5X_5(t) + a_{67}k_7X_7(t) + a_{68}k_8X_8(t) - k_6X_6(t)] \\ dX_7(t)/dt = \xi(t)[a_{75}k_5X_5(t) + a_{76}k_6X_6(t) - k_7X_7(t)] \\ dX_8(t)/dt = \xi(t)[a_{86}k_6X_6(t) + a_{87}k_7X_7(t) - k_8X_8(t)] \end{array} \right\} \text{SOM} \quad (1.4)$$

where X_i , $i = 1, 2, \dots, 8$, is the amount of carbon, respectively, in three plant, two litter, and three SOC pools; b_1 , b_2 , and b_3 are plant partitioning coefficients to leaf, root, and wood; k_i , $i = 1, 2, \dots, 8$, is the rate of carbon leaving the eight pools (i.e.,

process rate or exit rate); a_{ij} , $i = 1, 2, \dots, 8$, $j = 1, 2, \dots, 8$, is the transfer coefficient of carbon from pool j to pool i ; and $\xi(t)$ is the environment modifier. The eight carbon balance equations in Equation 1.4 can be reorganized into a matrix form as:

$$\begin{pmatrix} dX_1(t)/dt \\ dX_2(t)/dt \\ dX_3(t)/dt \\ dX_4(t)/dt \\ dX_5(t)/dt \\ dX_6(t)/dt \\ dX_7(t)/dt \\ dX_8(t)/dt \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \mu(t) + \xi(t) \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ a_{41} & a_{42} & 0 & -1 & 0 & 0 & 0 & 0 \\ a_{51} & a_{52} & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_{64} & a_{65} & -1 & a_{67} & a_{68} \\ 0 & 0 & 0 & 0 & a_{75} & a_{76} & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & a_{86} & a_{87} & -1 \end{pmatrix} \times \begin{pmatrix} k_1 \\ & k_2 \\ & & k_3 \\ & & & k_4 \\ & & & & k_5 \\ & & & & & k_6 \\ & & & & & & k_7 \\ & & & & & & & k_8 \end{pmatrix} \begin{pmatrix} X_1(t) \\ X_2(t) \\ X_3(t) \\ X_4(t) \\ X_5(t) \\ X_6(t) \\ X_7(t) \\ X_8(t) \end{pmatrix} \quad (1.5)$$

In this equation, a_{53} equals 1 as all the carbon from the wood pool goes to the structural litter pool. The above matrix equation can be succinctly expressed as:

$$\frac{dX(t)}{dt} = B\mu(t) + A\xi(t)KX(t) \quad (1.6)$$

where $X(t)$ is a vector of pool sizes, B is a vector of partitioning coefficients from carbon input to each of the pools, $\mu(t)$ is the carbon input rate, A is a matrix with -1 in the diagonal and transfer coefficients in the off-diagonal to quantify carbon movement along the pathways (equivalent to microbial carbon use efficiencies for carbon transfer among soil pools), K is a diagonal matrix of process rates (mortality for plant pools and decomposition coefficients for litter and soil pools) from donor pools, $\xi(t)$ can be a scalar or a diagonal matrix of environmental modifiers to represent responses of the carbon cycle to changes

in temperature, moisture, and oxygen. When $\xi(t)$ is a diagonal matrix, the environmental modifiers can be the same for all the pools or different for individual pools.

Equation 1.6 is generalizable to represent land carbon cycle models that follow first-order kinetics. It describes net carbon pool change, dX/dt , as a difference between carbon input $\mu(t)$, distributed to different plant pools via partitioning coefficients B , and carbon loss through the transformation matrices ($A\xi(t)K$) among individual pools $X(t)$.

As Equation 1.6 is generic to unify the land carbon cycle models that follow first-order kinetics, any single model is a special case of the generic equation. For example, the Community Land Model version 4.5 (CLM4.5) incorporates carbon transfer among seven pools per soil layer over 10 layers (Koven et al. 2013). The seven pools in each layer are metabolic litter, cellulose litter, lignin litter, coarse woody debris, fast soil organic matter, slow soil organic matter, and passive soil organic

matter. Carbon vertical transfers mainly occur between one layer and the next layer down. Huang et al. (2018b) converted the soil carbon module of CLM4.5 into one matrix equation as:

$$\frac{dX(t)}{dt} = B\mu(t) + A\xi(t)KX(t) + V(t)X(t) \quad (1.7)$$

The matrix equation above has three items on the right side, i.e., carbon input and plant partitioning, carbon transfers and release, and vertical movement. The dimension of the matrix equation is 70 for 70 pools. Transfer matrix A , for example, is a 7×7 block matrix. Each element of the block matrix, $V(t)$, is a 10×10 matrix. The vertical movement matrix includes carbon transferring to the next layer up and the next layer down.

Equation 1.7 shares similar mathematic properties with Equation 1.6. Equation 1.7 has 70 dimensions whereas Equation 1.6 has eight dimensions. Equation 1.7 describes carbon transfers among 70 pools over 10 soil layers where Equation 1.6 describes carbon transfers among eight pools, including plant, litter, and soil pools without explicitly defined soil layers.

Lu et al. (2020) converted the CLM5 carbon and nitrogen cycles into four matrix equations, two for vegetation carbon and nitrogen cycles and two for soil carbon and nitrogen cycles (see Chapter 6). The vegetation carbon cycle in CLM5, for example, contains 18 pools, including six tissue pools: leaf, fine root, live stem, dead stem, live coarse root, and dead coarse root. Each tissue pool is accompanied by a storage pool and a transfer pool. A crop grain tissue pool, accompanied by a grain storage pool and a grain transfer pool, is added when the crop model is used. Vegetation carbon dynamics are controlled by phenology for the leaf onset and offset dates according to air temperature and soil water conditions. Harvest and fire remove part of plant carbon from ecosystems and part goes to litter pools according to the harvest rate. Natural death moves carbon from plant pools to litter pools at defined mortality rates. The fire module triggers the occurrence of occasional fire events based on the amount of the fuel (i.e., litter) and the soil moisture. When fire occurs, carbon in plant pools is partly released to the atmosphere and partly transferred to the litter pools based on their tissue quality and burned area. Thus, the vegetation carbon dynamics are represented by the following matrix equation:

$$\begin{aligned} \frac{dX_v}{dt} = & B\mu(t) + (A_{phc}(t)K_{phc} + A_{gmc}(t)K_{gmc} \\ & + A_{fic}(t)K_{fic})X_v(t) \end{aligned} \quad (1.8)$$

Here, X_v is an n -entry vector, representing vegetation carbon pool size. K is an $n \times n$ diagonal matrix. Its subscripts phc , gmc , and fic indicate carbon processes related to phenology, gap mortality (i.e., harvest from land use and natural mortality), and fire, respectively. The diagonal entries are the process (or exit) rates of all vegetation carbon pools due to phenology (K_{phc}), gap mortality (K_{gmc}), and fire (K_{fic}). Once converted, the matrix equations make CLM5 more modular, analytically clear, easily diagnosed, and computationally more efficient for spin-up.

Equation 1.8 similarly shares mathematic properties with Equation 1.6. Equation 1.8 uses three processes, phenology, mortality, and fire, to control carbon transfers among 18 vegetation pools whereas Equation 1.6 only uses the environmental scalar $\xi(t)$ to control carbon transfer.

Some other global models, such as CABLE, LPJ-GUESS and ORCHIDEE, have also been converted to matrix equations. Once a model is converted to a matrix equation, it appears quite simple; a first-order differential equation. In fact, the equation is not that simple as it represents a nonautonomous system, which is discussed below. With this unified matrix equation, we can explore the general properties of carbon models and the general behavior of the land carbon cycle. This is what we call the *matrix approach*.

Note that matrix models themselves do not represent much significant advance in carbon cycle research. The matrix models are merely used to represent a set of differential equations in a matrix form to describe carbon cycling among multiple pools of the earth system (Bolin and Eriksson 1958). What we propose is to take the matrix expression to unify the land carbon cycle models, enable data assimilation, and develop a theoretical framework to guide carbon research.

The matrix representation can be adapted to accommodate nonlinear dynamics. For instance, Equation 1.6 can be modified for a nonlinear microbial model as:

$$\begin{aligned} \frac{dX(t)}{dt} = & B(X,t)\mu(X,t) \\ & + A(X,t)\xi(t)K(X,t)X(t) \end{aligned} \quad (1.9)$$

Thus, carbon input U , plant carbon partitioning B , transfer coefficients A , and process rates K all are potentially functions of pool sizes X . Carlos Sierra's group from the Max Planck Institute for Biogeochemistry in Germany have converted many nonlinear microbial models into matrix equations and explored their general behavior (Sierra and Müller 2015, Metzler, Müller, and Sierra 2018).

THE PARADOX OF THE MATRIX EQUATION AND NONAUTONOMOUS SYSTEMS

While we argue that all the land carbon cycle models that follow first-order kinetics can be converted into a unified matrix form, the general equation is mathematically extremely simple (e.g., Equation 1.6). A paradox arises: how can such a simple equation represent the extremely complex phenomena of the carbon cycle observed in the real world?

To explore this paradox, we organized a workshop in 2012, sponsored by US National Institute for Mathematical and Biological Synthesis, NIMBioS. We invited 20 applied mathematicians and 20 ecologists to explore the paradox: why is the matrix equation (Equation 1.6) extremely simple whereas carbon cycle phenomena observed in the real world can be very complex? We presented this paradox to the workshop participants. In the first couple of days, it was very difficult to convince the group about this issue. Some ecologists said that ecosystems are complex and cannot be described by such a simple equation. They urged us to re-examine this mathematical expression. We told them that we have used thousands of data sets to verify the equation. The applied mathematicians told us that this equation is too simple to be interesting enough for them to do any study with. They suggested that we add a nonlinear term in the equation. We told them that if a nonlinear term is added to the equation, it no longer describes the land carbon cycle as revealed by data. We were stubborn and kept pointing out the paradox. After one and a half days, Dr. James Cushing, an applied mathematician from the University of Arizona, told us that the system we were studying is probably a nonautonomous system.

Nonautonomous systems have been a subject of mathematic research in recent decades. They comprise dynamical systems with input and parameters being time dependent. The matrix equation of the land carbon cycle as expressed by Equation

1.6, indeed, has its inputs and parameters being time dependent.

With this new insight, we organized a working group to study the nonautonomous system, sponsored by US NIMBioS again. The working group consisted of a few applied mathematicians and a few ecologists. One of the group members, Dr. Martin Rasmussen, a professor from Imperial College London, had coauthored two books on nonautonomous systems. He taught us a lot about how to study nonautonomous systems. We worked together over three years and had four meetings.

The working group gained a few key findings. First, it found that nonlinear models of soil carbon decomposition (e.g. Equation 1.9) generate unrealistic responses to small perturbations and carbon input (Wang et al. 2014, Wang et al. 2016). A stability analysis and numerical simulations were conducted for two nonlinear microbial models (a two-pool model and a three-pool model) of soil carbon decomposition. Both models exhibit dampening oscillatory responses to small perturbations. In addition, the equilibrium pool sizes of litter or soil carbon are insensitive to carbon inputs in the nonlinear microbial models. This oscillatory behavior and insensitivity of soil carbon to carbon input exhibited by the nonlinear models have not been observed in the real world.

Second, we identified a mathematical foundation through proof of theorems on exponential stability to explain observed convergence of the land carbon cycle (Rasmussen et al. 2016). The land carbon cycle can be considered as a linear nonautonomous compartmental system described by a dynamical equation (i.e., Equation 1.6). The equation can be rewritten as:

$$\frac{dX(t)}{dt} = B\mu(t) + G(t)X(t) \quad (1.10)$$

where

$$G(t) = A\xi(t)K$$

Matrix G is invertible. The entries $g_{ij}(t)$ of G satisfy three conditions. First, there is always carbon to leave any individual pool as $g_{ii}(t) < 0$ for all i . Second, there is either no carbon flow pathway or some amount of carbon moving from pool j to pool i (i.e., $g_{ij}(t) \geq 0$ for all $i \neq j$). Third, carbon

exiting pool i is either all going to other pools or some going to other pools and some being lost from the system via respiration (i.e., $\sum_{i=1}^d g_{ij}(t) \leq 0$

for all j). As the land carbon cycle satisfies the three conditions, the system is exponentially converging to a time-dependent attractor, which we define as carbon storage capacity. The storage capacity in a nonautonomous system is not a static concept but varies with time due to the time-dependence of carbon input and residence time.

The working group also found a mathematical representation for a transient dynamical equation of the land carbon cycle (Luo et al. 2017), which is discussed below.

PREDICTABILITY OF THE LAND CARBON CYCLE

In the very beginning of this chapter, we showed convergence of the land carbon cycle toward some attractor states by “Mother Nature”. This “Mother Nature” is the exponential stability of the carbon cycle equation (Rasmussen et al. 2016). The certainty of any ecosystems converging to some attractor states after human and natural disturbances reflects the high predictability of land carbon cycle (Figure 1.4).

Given one type of forcing, patterns of carbon cycle dynamics are usually highly predictable (Luo et al. 2015). For example, periodic climate forcing over seasons usually influences the carbon cycle system to generate periodicity in carbon fluxes. Similarly, one fire disturbance event generates a pulse release of carbon from land to

the atmosphere followed by gradual recovery. The gradual recovery is well illustrated by vegetation overrunning the deserted village in China (Figure 1.1) and thriving young lodgepole pine trees recolonizing the burned landscape after the 1988 Yellowstone fires (Turner 2018) as described in the first section of this chapter.

Global change factors, such as rising atmospheric CO_2 concentration and warming, usually generate gradual but directional changes in carbon cycle processes. Shifts in fire regimes usually lead to disequilibrium in the carbon cycle. And ecosystem state changes, due, for example, to land use change from forests to croplands, usually result in abrupt changes in ecosystem properties, such as photosynthetic capacity and carbon processes that depend on it. While many processes of the carbon cycle are highly predictable, precisely predicting carbon storage changes under climate change requires lots of data to constrain model parameterization.

DYNAMIC DISEQUILIBRIUM OF LAND CARBON CYCLE

It is well known that the carbon cycle dynamics at steady state can be described by two terms: carbon input and residence time. The product of these two terms determines the carbon storage capacity. However, directional climate change pushes the land carbon cycle out of steady state towards dynamic disequilibrium (Luo and Weng 2011). To quantify the dynamic disequilibrium, we need a third term, in addition to carbon input and residence time, to represent the transient dynamics of the land carbon cycle. The working group

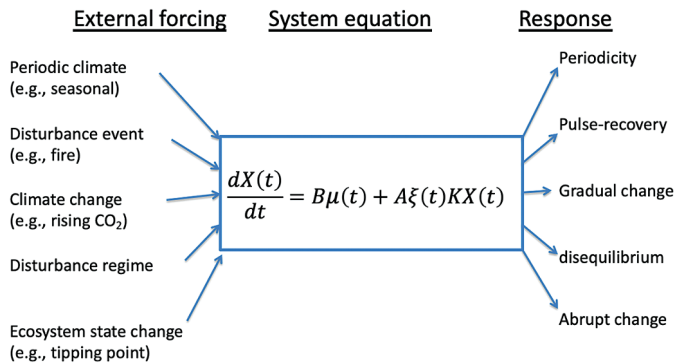


Figure 1.4. Predictability of the terrestrial carbon cycle. Responses of the carbon cycle to a given type of external forcing are highly predictable.

supported by NIMBioS worked very hard to find a mathematical expression of the transient dynamics. We derived a variety of mathematical formulations for the transient dynamics of the carbon cycle. One day we came up with the third term to describe the disequilibrium of the land carbon cycle. The original derivation was very complicated. After a few rounds of re-organization, we can now explain the mathematical derivation in a very simple way. Basically, we can multiply each term of Equation 1.6 by the inverse matrix of $A\xi(t)K$, $(A\xi(t)K)^{-1}$, and then re-organize the equation to get an equation of the form:

$$\begin{aligned} X(t) = & (-A\xi(t)K)^{-1} B\mu(t) \\ & - (-A\xi(t)K)^{-1} X'(t) \end{aligned} \quad (1.11)$$

Then, the transient dynamics of the land carbon cycle can be expressed by:

$$X(t) = \tau_E \mu(t) - X_p(t) \quad (1.12)$$

where τ_E is ecosystem residence time as defined by

$$\tau_E = (-A\xi(t)K)^{-1} B \quad (1.13)$$

and X_p is carbon storage potential. It represents the disequilibrium term of carbon cycle as:

$$X_p(t) = (-A\xi(t)K)^{-1} X'(t) \quad (1.14)$$

The first part in the right side of Equation 1.12 is the carbon storage capacity $X_c(t)$ as:

$$X_c(t) = \tau_E \mu(t)$$

This transient equation is also the dynamical equation of the land carbon cycle. The term

dynamical equation is often used in mathematics to describe how the state of a system changes over time. Although the formulation of the carbon dynamical equation is simple, it represents a nonautonomous system, which is influenced by five categories of external forcing variables (i.e., periodic climate, disturbance event, disturbance regime shift, climate change, and ecosystem state change) (Figure 1.4) (Luo et al. 2017). Those external forcing variables are superimposed on each other to influence carbon cycle dynamics, generating complex phenomena. That is why we observe complex phenomena of carbon cycle dynamics in the real world. Behind the phenomena is a relatively simple dynamical equation to govern the trajectory of the carbon cycle.

In the coming five units of this book, we will show you how the matrix approach can unify land carbon cycle models, accelerate spin-up, offer new analytics for model diagnostics, and facilitate data assimilation to improve the fit of models to observations.

SUGGESTED READING

Luo YQ, Weng ES. (2011) Dynamic disequilibrium of the terrestrial carbon cycle under global change. *Trends in Ecology & Evolution*, **26**, 96–104.

QUIZZES

1. What are the four fundamental properties of internal land carbon cycle processes?
2. What are the five categories of external forcing to influence the carbon cycle?
3. Briefly describe the donor-pool dominated carbon transfer.
4. How do external forcing variables interact with internal processes to create the complex phenomena of the land carbon cycle?

CHAPTER TWO

Introduction to Modeling

Benjamin Smith

Western Sydney University, Richmond, Australia

CONTENTS

What is a Model? / 13
Models in Research / 14
Ways of Using Models / 15
System Dynamics / 16
Types of Land Carbon Cycle Models / 16
Modeling Workflow / 18
Specify the Question or Hypothesis and Identify How Modeling Can Help / 18
Choose a Model / 19
Verify that the Model Works / 19
Calibrate the Model / 20
Validate the Model / 20
Design the Model Experiment / 21
Summary / 21
Suggested Reading / 21
Quizzes / 21

This chapter introduces the concept of a model, and its role within modern research methodology and the scientific method. Some typical characteristics of the system dynamics models used in carbon cycle studies are described, alongside examples of different ways they can be applied in ecosystem and Earth system research. With reference to in-depth discussion and examples throughout the book, we introduce the workflow of six steps you would typically follow when integrating a model within a robust research study design.

WHAT IS A MODEL?

The dictionary definition of a *model* is an *idealized, simplified or down-sized representation* of something, the purpose of which is to *describe, explain or depict* that something that the model represents, which we may refer to as its object. The model encodes information about certain, well-chosen aspects

of the object, and its purpose is to convey that information to the viewer or user of the model in a clear, concise, and potentially useful way. Compared to the phenomenon or notion it represents, the model may constitute a simpler, more lucid, or even an exaggerated representation of those aspects it is designed to convey – and that is exactly the purpose (Figure 2.1). Models induce us to perceive – and may sometimes help us to understand – something significant about the object by discarding extraneous detail and focusing on the essential. Thus, models are designed to *simplify, explain, and communicate*, and these aspects are interdependent: a simple explanation of a complex idea makes it easier to understand and convey to others. Similarly, models are used in learning, research, and academic exchange as a way of simplifying, synthesizing, and communicating knowledge, data and ideas, allowing us to put them into practice in useful ways.



Figure 2.1. Models are common in everyday life, and serve a similar purpose as in learning, research, and academic interchange, namely to simplify complex information and facilitate communication and understanding.

MODELS IN RESEARCH

Models in a broad sense are fundamental to modern research methodology. In its simplest form, a model may be a mental construct describing a perceived pattern or relationship; for example, the simple observation that an increase in one variable tends to be consistently associated with an increase or decrease in another. The existence of patterns and relationships in nature has surely been noted by people throughout time. The fact that certain types of edible plants were often found growing in a particular physical situation, or associated with certain conspicuous species, would have been important practical knowledge in hunter-gatherer societies. From the beginnings of agriculture, farmers have taken note of environmental events that signaled propitious conditions for sowing and harvesting their crops, or that warned of hazards such as droughts, freezing conditions, or insect plagues. Philosophers of ancient Greece pioneered the documentation of patterns and relationships in nature in a systematic way, reasoning about their wider significance. In modern terminology, they could be said to be the first to employ modeling as part of the research process. The ‘models’ of this era were qualitative, mental models, describing and potentially suggesting explanations for observed patterns and relationships, and expressed in words, illustrations, or allegories.

Today, models, whether mental, conceptual, or mathematically formalized (Figure 2.2), feature in most scientific research. Oreskes (1994) stated: “Numerical models are a form of highly complex scientific hypothesis”. The hypothesis, of course, is an element or ‘step’ in the empirical scientific method, pioneered by Galileo in the 17th century. The core of this method is the experiment, by which

observations from the real world are collected and examined in such a way as to shed light on the validity of a hypothesis. Often, several or many empirical studies addressing the same or related hypotheses are needed before scientists achieve consensus and the former hypothesis becomes an element of a consensus theory or ‘settled science’. The hypothesis behind a given study is rarely entirely novel, but represents a step beyond the existing frontier of knowledge in the research area or field. We can think of the hypothesis as an ‘informed guess’ given what we already know thanks to the consensus knowledge and theory of the field, accrued over earlier studies and scientific discourse.

Similarly, the models we commonly use in research have the character that they bring together elements of established knowledge with ideas or informed guesses about things we do not yet know with certainty, or lack data to express in a precise, quantitative way. In this sense, a model provides a context or frame for integrating knowledge and observations to pose questions about the real world in the form of a testable hypothesis. A model is not necessarily ‘true’ or proven, but can stand as a formalized or explicit hypothesis of how the system under study works.

Many different kinds of models are applied within the environmental and earth sciences. This book focuses mainly on numerical simulation models, implemented as software algorithms and executed on a computer. A familiar class of simulation models are the numerical weather prediction (NWP) models used by weather service agencies (like the US National Weather Service) to produce daily and longer-term weather forecasts. These models depend for the (relative) accuracy of their predictions on knowledge of the physical processes that govern the dynamics of weather, and

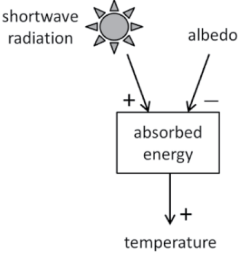
Mental	Conceptual	Mathematical
<p>“the darker a surface, the more sunlight it absorbs and the warmer it gets”</p>		$T = \sqrt[4]{\frac{E(1-\alpha)}{\sigma}}$ <p>where: T = temperature of surface in Kelvin; E = incoming solar radiation flux in Wm^{-2}; σ = Stefan – Boltzmann constant ($5.670373 \times 10^{-8} \text{ Wm}^{-2}\text{K}^4$); α = albedo.</p>

Figure 2.2. Three increasingly formal and precisely specified models that express a common hypothesis.

on extensive observations (such as measurements from weather stations, balloons and satellites) both to ‘train’ the model, improving its fit with the measurements, and to fix the state or ‘initial conditions’. Integrating observations in this way is called ‘constraining’ the model.

NWP models are an example of the class of models denoted as mechanistic or process-based. Many of the core equations of an NWP model express known physical relationships or laws governing energy balance and motion, applied to the three-dimensional atmosphere. Others, such as equations governing the formation and behavior of clouds, are parameterized, i.e., fitted to observations, but informed by physics. This combination of mechanistic (process) knowledge and empirical fitting is characteristic for most process-based models.

WAYS OF USING MODELS

Perhaps because weather prediction models and forecasts are so familiar from daily life, many people, including scientists who do not habitually work with numerical models as part of their methodological toolbox, tend to think of such models primarily as tools for prediction forward in time. Certainly, extrapolating beyond the range of observations (forward in time, or in any temporal or spatial context for which observations are sparse or lacking) can be one useful way to apply some types of models (though not all models are suitable for this). However, this is far from being the only way modeling can form part of research methodology, and is perhaps that with the least relevance to science’s central endeavor to advance fundamental knowledge about the natural world. In general, the potential for new discoveries is greatest at the interface between modeling and observation, where information on real-world phenomena encoded in

data meets the potential of modeling to explain or decode patterns in the data in terms of underlying mechanisms of cause and effect.

Box 2.1 shows some ways in which process-based models of the kind covered by this book

BOX 2.1 Modes of model application

- project future changes and impacts, extrapolate beyond current data (e.g., Chapter 18 – C cycle transient responses to future climate change)
- scale-up findings from local studies to regional or global scale (e.g., Chapter 27 – model-data fusion of sub-continental GPP)
- characterise uncertainty, identify robust responses and relationships (e.g., Chapter 17 – tracing uncertainty sources in land carbon models)
- attribute observed relationships and patterns to underlying drivers and mechanisms (e.g., Chapter 30 – mechanisms controlling lake C dynamics)
- synthesise the state of knowledge, identify gaps, generate hypotheses to guide empirical studies (e.g., Chapter 33 – ecological forecasting of field experiments)
- communicate scientific evidence to societal end-users and decision-makers

increasing potential for integration with empirical studies

can be applied within research on the land carbon cycle. In general, the potential for integration between modeling and empirical approaches increases as you go down through the list. We will see examples of most of these modes of applying models in the course of this book.

SYSTEM DYNAMICS

Land carbon cycle models are at their core system dynamic models. Systems are fundamental to the organization and processes of human society and our daily lives. Think of a ‘political system’, ‘educational system’ or ‘energy system’ as examples. Fundamentally, systems are a mental framework through which we may view, understand or organize complex activities or ideas that involve or depend on linkages between different interacting elements or parts. The presence of interlinked elements, constituting a ‘whole that is greater than the sum of its parts’ is characteristic for a system. ‘Systems thinking’ can also be applied to nature and has had a powerful influence on the development of quantitative techniques in ecology and numerous other fields. General systems theory, formalized in the 1940s, depicts an entity under study as a network of interrelated elements whose properties and linkages influence the behavior of the system and its evolution over time. Systems theory provides a framework for formalizing a conceptual or hypothetical understanding of an object of study in terms of drivers, responding processes and networks of interacting elements, unified through the consideration of flows – for example energy, matter, or capital – across the network.

When we model the carbon cycle, the system in focus is an ecosystem, or more broadly the Earth system (constituting the Earth’s interacting ‘spheres’ – the atmosphere, hydrosphere, biosphere etc.). The origins of the ecosystem concept can be traced to the British botanist Arthur Tansley, writing in 1935, who argued for the need to consider organisms and their environment as part of a unified whole in order to understand patterns and changes in nature. Tansley noted that organisms are locked into constant interactions with their immediate physical environment, and that the interactions between individuals and species are mediated by the changes they impose on physical (and chemical) factors of the microenvironment through their growth and function. This notion that *interactions* between organisms and the air, soil, and water in which they occur

govern pattern and process in nature is central to systems thinking in ecology.

TYPES OF LAND CARBON CYCLE MODELS

Today’s ecosystem and Earth system models (ESMs) can be considered the end branches of a family tree having its roots in the work of a family of brothers, Eugene and Howard, and their textbook, *Fundamentals of Ecology*, first published in 1953. They combined concepts from system dynamics theory with analogies from electrical engineering and thermodynamics to frame ecological systems as networks in which interactions result from energy flows between trophic species levels, linking organisms with their abiotic environment via biological analogues of concepts from electrical networks such as voltage, capacitance and resistance. Modern biogeochemical models build on similar foundations to Howard Odum’s depiction of energy flow for the Silver Springs aquatic ecosystem in Florida, USA (Odum 1971; Figure 2.3). Energy that enters the system as sunlight, powering the photosynthesis of producers (algae and aquatic plants), is used to drive the metabolism of the producers and the consumers that depend on them, up through the trophic chain. At each trophic step, energy is lost through respiration, returning heat to the environment. The Silver Springs ecosystem is here represented as a system of compartments (the trophic groups) that exchange energy with each other and with the external environment. Each compartment has a store of biochemical energy, linked to the abundance of organisms in that trophic group, and the size of this store changes over time as energy is gained and lost through processes like photosynthesis, herbivory, predation and detritus production. All flows into, out of and between compartments of the system are mirrored by changes in the sizes of the compartments themselves – the system is said to uphold *mass balance* (here expressed in units of energy). We can state that the Silver Springs ecosystem is here depicted as a *compartmental dynamic system*. Such a representation has some very useful mathematical properties for simulation and analysis of the system, as we shall see elsewhere in this book, and as discussed in detail in Chapter 7.

An alternative ‘currency’ for this model, instead of energy, would be the carbon that enters the system as CO₂ assimilated through photosynthesis, is transferred between trophic levels through

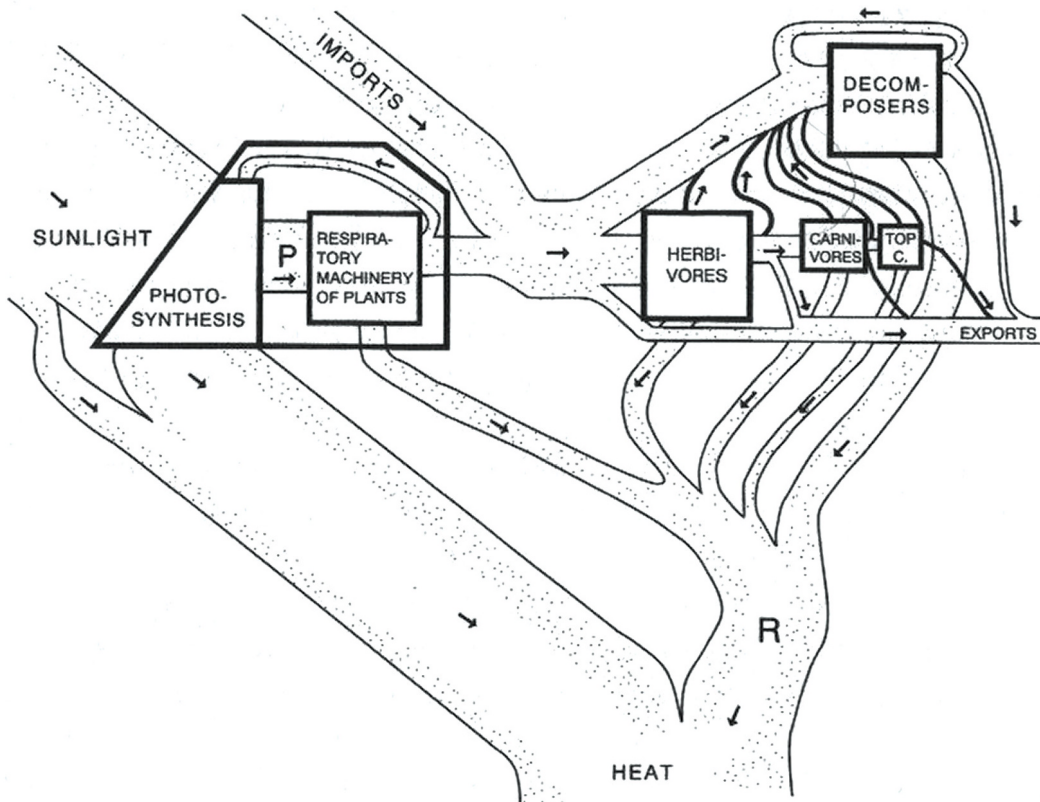


Figure 2.3. Energy flow diagram for the Silver Springs ecosystem in Florida. Adapted by S. Maud from Odum, H.T. (1971) *Environment, Power, and Society*. Wiley-Interscience, New York.

herbivory, production of detritus, or predation, and is lost to the system as CO_2 released through autotrophic or heterotrophic respiration (in this aquatic system, there are also imports and exports due to streamflow and movement of organisms in and out of the study area). Similarly, ecosystem models in use today typically adopt a carbon-based representation of the ecosystem and its network of pools and fluxes as their basic framework. An example of such a framework was presented for the TECO model in Chapter 1. Superimposed on this, many models incorporate representations of hydrological, nutrient and energy cycles that interact with carbon cycle processes and states. In Chapter 6 we shall see how a nitrogen cycle may be added to the carbon-only version of TECO, to account for progressive nitrogen limitation under elevated atmospheric CO_2 .

Some of the main ‘families’ of modern land models incorporating representations of carbon processes are detailed in Table 2.1. These groupings and their defining characteristics as shown in the

table are not clear-cut. On the contrary, each model family has evolved over time as investigators have enhanced and adapted existing tools for application to novel research questions and management problems. Each type of model was originally developed with a certain research goal in mind. For example, Soil-Vegetation-Atmosphere Transfer (SVAT) schemes, first developed in the 1980s, focus on land surface hydrology and energy balance, as important exchange processes that affect the dynamics of the atmosphere, driving weather and climate. SVAT models like the revised version of the Simple Biosphere Model (SiB2; Sellers et al. 1996), which originally did not incorporate explicit carbon processes, were extended to incorporate a representation of canopy CO_2 exchange because an estimate of photosynthesis rate was required as a control on stomatal conductance and evapotranspiration, in turn affecting the partitioning of the vegetation-atmosphere energy flux into sensible and latent heat components. Modern land surface models (LSMs) coupled to ESM frameworks

TABLE 2.1
Model types used in contemporary land carbon cycle research

	Soil-vegetation-atmosphere transfer scheme (SVAT)	Forest growth model	Terrestrial biogeochemistry model	Dynamic global vegetation model (DGVM)	Biogeochemical Land surface model (LSM)
Example	SiB2	3PG, 4C, SORTIE	CENTURY, TECO	IBIS, LPJ-GUESS, ED2	CABLE, ORCHIDEE, CLM4.5
Energy cycling/balance	✓	—	sometimes	sometimes	✓
Hydrological cycling/balance	✓	✓	✓	✓	✓
Canopy physiology/CO ₂ exchange	✓	✓	✓	✓	✓
Plant C dynamics	—	✓	✓	✓	✓
Belowground C dynamics	—	sometimes	✓	✓	✓
Nutrient (N, P) dynamics	—	sometimes	✓	sometimes	sometimes
Plant functional types (PFTs)	static	static/dynamic	static	dynamic	usually static
Stand dynamics	—	sometimes	—	sometimes	—
Typical application	local/global	local	local	regional/global	regional/global

and used for global climate change simulations have evolved from SVAT models, progressively incorporating more processes and interactions, often adopted from schemes first developed for the other model families like forest growth models (used to simulate the growth and dynamics of trees and forest stands), terrestrial biogeochemistry models (adding belowground carbon and nutrient dynamics) and dynamic global vegetation models (DGVMs; adding competition between plant functional types or PFTs). In fact, current representatives of many of the model families shown in Table 2.1 have converged to incorporate many of the same algorithms, functionality, and underlying theory across model families. The frameworks contributing to global studies and assessments, such as annual updates of the global carbon budget (Sitch et al. 2015), and to account for biogeochemical and biophysical feedbacks to climate change in IPCC climate projections, are collectively termed terrestrial biosphere models (TBMs). Fisher et al. (2018) provides a useful account of the current status of, and research front for, TBMs incorporated within ESM frameworks.

MODELING WORKFLOW

When using models as part of a robust research study design, there are a number of steps to go through. Depending on the study and the prior work we are building on, by ourselves or others, we may leave out some steps, change the sequence, or iterate more than once over a given set of steps. However, most modelers would agree, in principle, that a study should feature the following steps, and that these should be systematically described in publications discussing the work, allowing others to fully understand and reproduce it.

Specify the Question or Hypothesis and Identify How Modeling Can Help

It is rather elementary that any scientific study begins with a question, and linked to this, a testable hypothesis or hypotheses. If the hypothesis can be expressed or captured in the form of a model, and there are data available to compare and contrast to the model results, modeling may be valid to consider as part of the study design.

Choose a Model

This step could also be labelled ‘develop a model’ or ‘adapt a model’. Textbooks on modeling often advocate for building the model from first principles, starting from a conceptual diagram that captures the hypotheses of the study, to be reflected in the model. In an ideal world, this ensures the chosen model is targeted specifically to the question of the study, no more and no less. This principle is captured by the popular modeler saying “the model should be as simple as possible, but no simpler”.

In practice, the carbon cycle literature contains only rare examples of studies using completely novel model structures and codes. Many current model frameworks have evolved over decades, involving many person-years of development, coding, evaluation, and application. Studies employing the model along the way turn up issues (scientific, as well as programming errors) focusing attention on needed revisions and improvements, but also increase confidence in the model where they demonstrate that it is able to reproduce patterns or relationships seen in observations, or in independent studies using alternative approaches. A typical ecosystem model is a sophisticated software application comprising hundreds to thousands of lines of computer code. Even though it could be argued that many models have grown overly complicated, to at least some degree this reflects the complexity of real ecosystems, the range of biotic and abiotic factors and interactions that govern their dynamics, and the steadily advancing research front in understanding and modeling natural systems.

Thus, in carbon cycle science, it is seldom efficient or realistic for an investigator to develop a new model completely from scratch, even if we accept that this would be the ideal. Rather, it is a question of choosing a model framework that fits the research question and study system in terms of criteria such as: model versus system complexity, scaling assumptions (temporal/spatial), available evidence of model skill (e.g., past published studies on similar systems or questions), configurability of parameters, input files, and source code, and availability of code and documentation.

A common pitfall is to select a model that has convenient technical features but is poorly matched to the target study in terms of spatial and temporal scale. There are two essential aspects to this. One is that the same process can exhibit different

sensitivity depending on the scale of observation. For example, due to structural, functional and compositional heterogeneity, photosynthesis measured at the leaf scale in a forest will show a different relationship to drivers such as temperature, insolation or CO₂ concentration compared to the canopy, stand or landscape scale. Decomposition of soil organic matter will show a different apparent sensitivity to temperature (Q_{10}) in a chamber measurement over an hour, compared to an incubation experiment over a year, or a decades-long soil inventory dataset. This means that the structure and parameters of the most suitable model to study variations in photosynthesis or decomposition will differ depending on the temporal and spatial context of the system under study and the research question in focus. Second, different processes and entities of the ecosystem are important in controlling variations at different scales. For example, seasonal cycles of leaf production and shedding (phenology) are important for the productivity of nemoral forests at annual and longer time scales, but leaf area index could be prescribed (specified as a fixed value) when modeling the gas exchange of a nemoral tree over a diurnal cycle. Changes in the sizes of ‘slow’ and ‘passive’ soil organic matter (SOM) pools tend to dominate soil carbon dynamics on scales of centuries and millennia, but have no impact on variation in CO₂ flux in short-term chamber measurements. This implies that a single-pool SOM model may be suitable for the analysis of data from soil chambers, whereas a model used to analyze changes in global soil carbon under IPCC future climate projections needs to distinguish multiple SOM pools of different lability (this is further discussed in Chapter 23). Ensuring that the chosen model accounts for the processes and entities relevant to the scale and question of the study is a particularly important consideration in carbon cycle studies.

Verify that the Model Works

Modelers distinguish between ‘verification’ and ‘validation’. The former seeks confirmation that the model implementation behaves as expected in a purely technical sense. When we run a numerical model we are ‘solving’ the model given the input data. Verification entails ensuring that the output data from the model matches the solution of the set of difference (or balance) equations the model (typically) encodes (see Chapter 3). In practice we

seldom have the true (analytical) solution of these equations available to compare with the output of the model, but, based on our scientific knowledge of the system we are simulating, we can usually tell whether the output is reasonable or 'sane' given the forcing. For example, carbon pool sizes should not normally go negative, while fluxes such as GPP, autotrophic and heterotrophic respiration, or CH₄ emission, should be within a certain range.

Calibrate the Model

Some though not all models can be calibrated against different types of observational or measurement data. For a system dynamic model, calibration may involve tuning parameters of individual equations or processes based on measurements or estimates of those processes. For example, key parameters of a biochemical photosynthesis model may be calibrated based on gas exchange measurements of leaves, yielding a so-called *A-c*_i curve. When you choose an existing model framework for your study, this kind of process-level calibration will likely already have been done, though you may have good reason to perform your own calibration if you have relevant measurements from your system.

A greater challenge is performing calibration on the overall output of the model, where this emerges from interactions between different processes, drivers and the evolving system state. For example, net biome production (NBP) in a global carbon cycle model may be the emergent balance between uptake (photosynthesis/GPP) and multiple release fluxes (autotrophic and heterotrophic respiration, emissions from wildfires) integrated across the land ecosystems of the world. The release fluxes in particular depend not only on current environmental drivers, but on the cumulative sizes of source pools such as SOM pools of different quality/labidity in different climates. Because of the long response lags, or spin-up effect, in such a model, and due to spatial heterogeneity, errors in individual processes can rapidly accumulate and combine to generate large bias in NBP, even if individual processes are well calibrated to measurement data, where such are available.

This book offers a number of the best available solutions to the considerable challenge of calibrating the emergent output of an ecosystem model, in all its complexity. In essence, these involve

identifying the parameters across different process formulations of the model to which the model is most sensitive (with respect to a particular output variable, such as NBP), specifying the potential real-world range of each parameter, and searching the hyperdimensional space of this parameter set (within the realistic range) to find a combination of values that yields the best fit between the model output, given those parameter values, and observational data on the same variable. This approach, termed data assimilation or model-data fusion, is introduced in Chapter 21, and further elaborated and exemplified in subsequent chapters of Units 6–8.

Validate the Model

In the introduction to this chapter it was noted that all models, by definition, are a simplified depiction of the real thing they represent. This is not a failing or shortcoming but the very essence of a model. However, being simpler also implies that even a good model can never be expected to replicate the behavior of a real system exactly, given that some processes and interactions involved in the behavior of the real system are missing. Thus, model error is inherent in the very concept and purpose of a model. Box (1979) stated: "all models are wrong, but some are useful".

The simple observation that a model can never be perfectly 'true' has led some writers (e.g., Oreskes (1994)) to suggest a model cannot be validated, in the sense that validation implies confirmation of truth. This argument is largely semantic however, and in practice we are of course very interested in knowing the extent to which the model behaves in a similar way to what we know, from observation and theory, of the real system under study. Canham (2003) stated: "The process of evaluating model structure is clearly critical enough to warrant a specific term, and 'validation' appears to be the best candidate".

A robust validation strategy focuses not only on the emergent output of the model (for example, NBP), but on the behavior of individual process representations and the assumptions they entail. This is because compensating errors and biases in different processes, or for example across the grid of a spatially-distributed model, can coincidentally produce the 'right result for the wrong reason', when focusing only on an emergent or

integrated output such as NBP. This issue is formally termed *equifinality*, the potential to obtain similar results with different model structures and parameterizations (Luo et al. 2009). The Free-Air CO₂ Enrichment Model Data Synthesis (FACE-MDS) initiative have pioneered best practice in validation of complex ecosystem models using datasets from field experiments through an ‘assumption-centered’ approach (Medlyn et al. 2015). The alternative, perhaps more established and less labor-intensive, approach of *benchmarking* uses data on multiple output variables to simultaneously query the skill of the model in different dimensions, reflecting different processes and feedbacks (Chapter 19).

Design the Model Experiment

The model experiment, sometimes also termed the simulation protocol, tailors the configuration, forcing and output data from the model simulation so as to shed light on the research question of the study. As discussed in the introduction to this chapter, the model typically stands for, and incorporates, our hypothesis or hypotheses about the system and its responses to influences such as a shift in climate or a management intervention. Through the model experiment, we wish to probe whether, or under what assumptions and conditions, the model reproduces salient observations of the system, thereby testing our hypothesis as encapsulated by the model, and allowing us to draw inferences about the behavior of the system.

There are several elements to designing a robust model experiment and some of these are challenging in the case of carbon cycle studies. One challenge concerns the initialization or *spin-up* of the model state. In general, the spin-up strives to attain a steady state for ecosystem carbon pools ahead of the main part of the model simulation, avoiding drift in the model output. Some challenges and solutions are presented in Chapter 13 and exemplified in Chapter 14. An alternative to the steady state assumption is discussed in Chapter 27.

The book contains many examples of model experiments tailored to different research questions and problems. Some examples were listed in Box 2.1.

SUMMARY

We have seen that the model, whether conceptual or formally specified, is an inherent component of research methodology and the scientific method, integrating the knowledge we have with hypotheses we may wish to pose about the system under study. Several different types of models are applied within carbon cycle research, but this book focuses on numerical simulation models encapsulating networks of plant and soil compartments and the processes that govern the flows of carbon and other elements across the network, influenced by environmental drivers. The workflow of six steps, further elaborated in other parts of the book, serves as a guide to how we may integrate modeling in the design of a robust scientific study.

SUGGESTED READING

Canham, C.D., Cole, J.J. & Lauenroth, W.K. 2003. *Models in Ecosystem Science*. Princeton University Press, Princeton, NJ.

QUIZZES

1. “The model should be as simple as possible, but no simpler” – Discuss!
2. List four ways in which models and observations can be combined within ecosystem studies. What is the role of the model versus the observational data in each case?
3. The carbon cycle of an ecosystem can be depicted as a compartmental dynamic system. List four properties of such a system.
4. How is the spatial and temporal scale of a study relevant in selecting a suitable model?
5. Define these terms: calibration, verification, validation, benchmarking.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER THREE

Flow Diagrams and Balance Equations of Land Carbon Models

Yuanyuan Huang

Climate Science Centre, CSIRO, Canberra, Australia

CONTENTS

Carbon Flow Diagram / 23
Carbon Balance Equations / 25
From Flow Diagram to Carbon Balance Equations / 27
Suggested Reading / 30
Quizzes / 30

Great oaks from little acorns grow. This chapter offers basic concepts and tools for building land carbon models. If we build on them step-by-step, we can end up understanding complex land carbon models. The goal of this chapter is to understand carbon flow diagrams and balance equations, to be able to derive generic carbon balance equations from carbon flow diagrams, and *vice versa*.

CARBON FLOW DIAGRAM

Flow diagrams are not new to us. We use flow diagrams in different forms, complexities and for different purposes. For example, we could draw a simple flow diagram to help us diagnose what might be the issue and what to do when a computer stops working (Figure 3.1b). Or you could find a complex flow diagram like the schematic of carbon, nitrogen and phosphorus cycles considered in ORCHIDEE-CNP (Figure 3.1b), which illustrates sophisticated interactions among carbon, nitrogen and phosphorus dynamics. A classic carbon flow diagram from IPCC reports is shown in Figure 3.1c. When it comes to carbon, we are generally interested in tracking the amount of carbon in space and time. In Figure 3.1c, you see different compartments or pools with different amount of carbon. You also see flows of carbon from one compartment to another.

We use land carbon models to track temporal-spatial dynamics of carbon, to answer questions like: Where is carbon stored? How long does carbon stay in one place? When, how, where, and how much carbon is transferred? Depending on the scientific questions, we may also use land carbon models to track water, nutrients, and energy as they are parts of the critical environment that drives and interacts with carbon dynamics. Earth is a system comprised of numerous interacting processes. Here, we focus mainly on carbon dynamics, but a lot of principles and methods introduced here are also applicable to water, nutrients, and energy. A flow diagram is very helpful in conceptualizing our issues, clarifying study boundaries, and disentangling complex interactions.

We use the terms stock (or storage, pool) and flow (or flux) very frequently in carbon studies. Suppose that you have a bank account. The total amount of money in your bank account is the stock. Every month, you deposit a certain amount of money, for example, from your salary. You also withdraw some money each month for your everyday expenses. This deposit and withdrawal are the flows into and out of your account. The same concept applies to carbon. We could take total carbon in the biosphere, for example, as our stock, and we track carbon flows into (deposits) and out of (withdrawals) the biosphere to understand the dynamics of carbon stock in the biosphere.

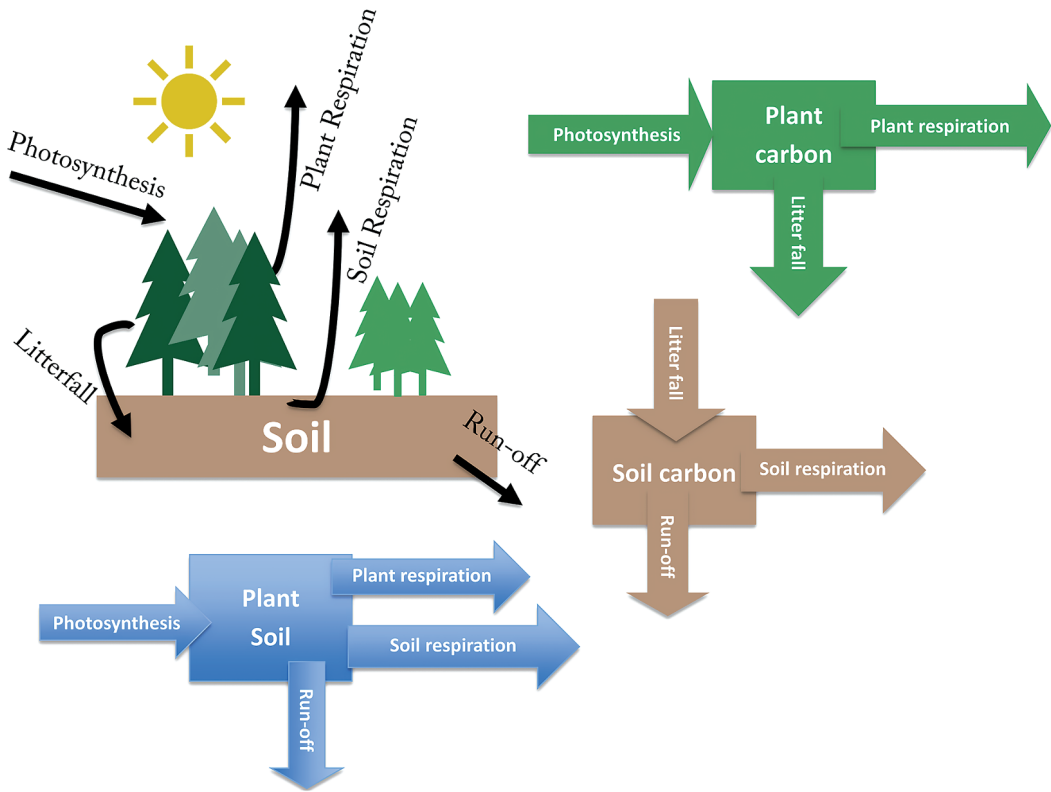


Figure 3.2. An idealized illustration of terrestrial processes of carbon flow.

In the field, soil respiration generally comprises two main components: the autotrophic respiration that takes place in plant roots and the heterotrophic respiration due to the breakdown of soil organic matter by microbes. If, instead, we take plant and soil together as our system, the carbon stock is the total carbon in plant and soil. The incoming carbon flux is the carbon flow through photosynthesis and the outgoing fluxes are plant respiration, soil respiration and carbon lost during runoff. Litterfall is no longer an incoming or outgoing flow, but an internal flow linking the plant and soil carbon stocks.

CARBON BALANCE EQUATIONS

The principle for tracking carbon dynamics is the law of conservation of mass. Matter can neither be

created nor destroyed, but chemical structure and physical form can change.

If we know the initial carbon pool size, and the values of carbon fluxes going into and out of the pool, we can track the dynamics of the carbon pool through time. This follows the law of conservation of mass. In terms of carbon, the change in its pool size is always equal to the mass of total incoming minus total outgoing flows. Suppose we have a carbon pool (I call it x here) that weighs 30,000 grams (gC) initially, i.e., at time, $t = 0$ (Figure 3.3). You could think about it as if you have \$30,000 in your bank account at the start of the year. Let's say the incoming flux is 2,000 gC year⁻¹. We call it the input rate I . In our monetary analogy, it corresponds to your annual salary. Let's say your income is relatively stable, then I does not change with time. For the outgoing flow, the donor pool-dominant



Figure 3.3. A conceptual one-pool carbon model.

transfer is a common example from the land carbon cycle. For example, if x represents the carbon contained in the biomass of a plant, respiration represents an outgoing flow that depends on the size of this pool: if we have a bigger plant, we would end up with more respiration. Back to our analogy: the higher your income, the more tax you pay. Suppose 20% of carbon in the plant pool is used up through respiration every year. This can be expressed as a turnover rate: a proportion of the pool that is used up and converted to an outgoing flux per unit time. Let's give it the symbol k . Here, $k = 0.2 \text{ year}^{-1}$. So how much carbon will we have at the end of the year or at the beginning of the next year?

Here our time step is 1 year. When $t = 0$, we have $x = 30,000 \text{ gC}$ (initial pool size), $I = 2000 \text{ gC year}^{-1}$, $k = 0.2 \text{ year}^{-1}$.

When $t = 1$, our pool size $x(t = 1) = x(t = 0) + \text{input} - \text{output} = 30,000 + 2,000 \times 1 - 0.2 \times 30,000 \times 1 = 26,000$.

What if we want to know x when $t = 2$? It is the same logic. $x(t = 2) = x(t = 1) + \text{input} - \text{output} = 26,000 + 2,000 \times 1 - 0.2 \times 26,000 \times 1 = 22,800$. Similarly, we could derive the carbon stock in the third, fourth, fifth year, and so on. The mathematic equation for x could be summarized as:

$$x(t+1) = x(t) + I \times \Delta t - k \times x(t) \times \Delta t \quad (3.1)$$

which says x at the next time step equals x at the current time step, plus the change of x . The time step (1 year) is denoted by Δt . The change of x is expressed as the input flow minus the output flow. This is the basis for the Euler method, a numerical method we sometimes apply to solve the differential equation to be introduced below.

Equation 3.1 can be rewritten as,

$$\frac{x(t+1) - x(t)}{\Delta t} = I - kx(t) \quad (3.2)$$

When $\Delta t \rightarrow 0$, Equation 3.2 is the same as,

$$\frac{dx}{dt} = I - kx \quad (3.3)$$

Equation 3.2 is the difference version of the carbon balance equation, and Equation 3.3 is the differential version. Both equations embody the law of conservation of mass. Another way of putting this is, they uphold mass balance: the rate of change of a carbon pool equals the input rate minus the output rate.

If we have two carbon pools in our system, sometimes the outgoing flux of one pool might become the incoming flux of the other (Figure 3.4). For example, in our plant-soil system in Figure 3.2, the litterfall is an outgoing flow of the plant pool and an incoming flow into the soil pool. For the plant, the carbon balance equation can be written the same as in the above one-pool case. The carbon balance equation of the second (soil) pool also expresses the mass balance. Here the incoming rate is a fraction, say f_{21} , of the outgoing flux from pool x_1 . The outgoing flux from pool x_1 is a constant annual proportion of the plant pool: $k_1 x_1$. This outgoing flux comprises the fraction that goes into soil (e.g., through litterfall) and the fraction that does not enter soil (e.g., plant respiration). Combining these terms, we have the incoming flux of pool x_2 as $f_{21} k_1 x_1$. The outgoing flux from x_2 is also pool-size dependent: $k_2 x_2$. The carbon balance equations for this two-pool system are given by Equations 3.4 and 3.5:

$$\frac{dx_1}{dt} = I - k_1 x_1 \quad (3.4)$$

$$\frac{dx_2}{dt} = f_{21} k_1 x_1 - k_2 x_2 \quad (3.5)$$

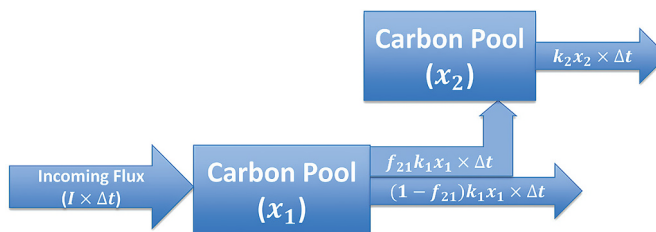


Figure 3.4. A conceptual two-pool carbon model.

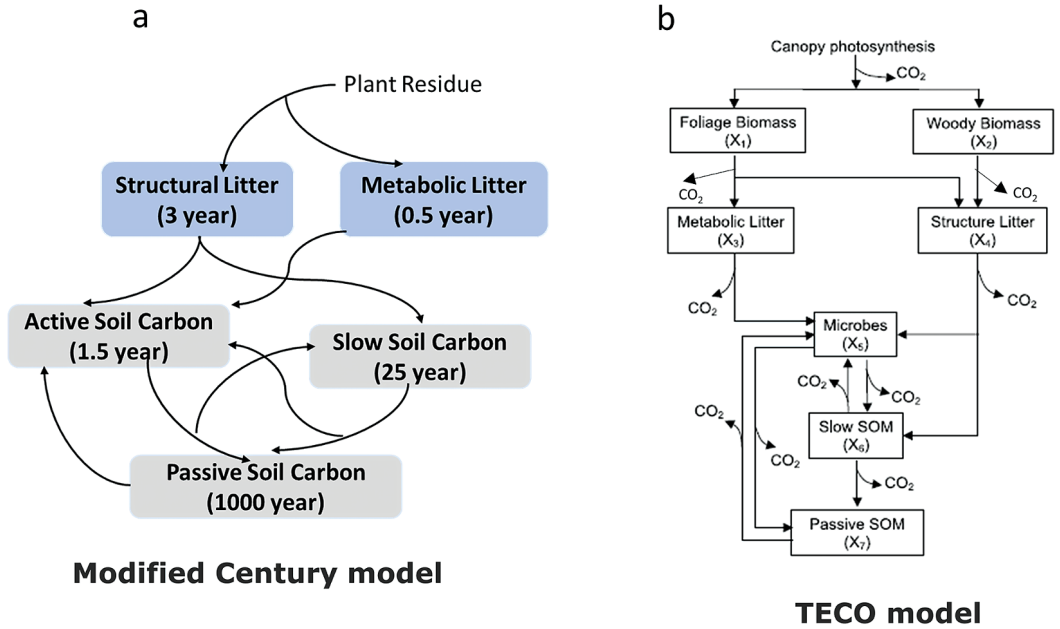


Figure 3.5. Flow diagrams of two carbon models: (a) CENTURY; (b) TECO. (a: adapted from Parton et al., 1988; b: adapted from Xu et al. 2006).

We may extend this same logic to models with any number of pools. For each pool, we need to know what are its incoming and what are the outgoing fluxes. Figure 3.5 shows the carbon flow diagram for two well-known carbon models. The left one is the classic soil carbon model, CENTURY (simplified, Parton, et al., 1988). The right is the TECO model (Xu et al., 2006). Both models conceptualize our natural carbon cycles into different pools or ‘compartments’, represented by boxes in the diagram, and track the transfers of carbon among different pools, represented as arrows. The transfers could be parameterized as one constant turnover rate relative to the size of the donor (source) pool, or could be modeled by functions with multiple terms, parameters, and dependencies, capturing complex interactions with climate, soil, and other factors.

FROM FLOW DIAGRAM TO CARBON BALANCE EQUATIONS

We take the TECO model as an example to illustrate how to derive a system of carbon balance equations from the carbon flow diagram. In Figure 3.5, we see that TECO has seven carbon pools: two plant biomass carbon pools (foliage, woody biomass), two litter carbon pools (metabolic, structure), and

three soil organic carbon pools (microbes, slow soil organic matter (SOM) and passive SOM). We use x_1 to x_7 to denote these pools. For each carbon pool, we need to specify the fluxes that enter and leave the pool. Carbon flux through photosynthesis is the external carbon input into our system of seven pools. We will denote this input rate by I . Photosynthetic carbon input is allocated into foliage and woody biomass, in relative proportions denoted by the allocation coefficients β_1 and β_2 . The carbon balance equation for foliage pool (x_1) and woody pool go as Equations 3.6 and 3.7.

$$\frac{dx_1}{dt} = I\beta_1 - k_1x_1 \quad (3.6)$$

$$\frac{dx_2}{dt} = I\beta_2 - k_2x_2 \quad (3.7)$$

The structure litter pool (x_4) receives carbon input from pools x_1 and x_2 . So the carbon balance equation of x_4 has two input fluxes. One is a fraction (f_{41}) of the outgoing carbon flux from x_1 (k_1x_1). The other is a fraction of the outgoing carbon flux from x_2 . This incoming flux could be written as $f_{42}k_2x_2$. The outgoing flux for pool x_4 is dependent on the pool size and its turnover rate, k_4 . We write this flux as k_4x_4 . Equation 3.9 is the

resulting carbon balance equation of the structure litter pool, x_4 .

$$\frac{dx_4}{dt} = f_{41}k_1x_1 + f_{42}k_2x_2 - k_4x_4 \quad (3.8)$$

The same procedure applies for the other carbon pools. We count how many arrows go into a pool and how many go out. One arrow normally corresponds to one flux. The boundary of the system in this example encompasses the seven organic matter pools. The atmosphere is not part of the system and we do not track further the fluxes that go into the atmosphere, i.e., respiration represented by arrows labelled 'CO₂' in Figure 3.5. That being said, how much carbon goes into the atmosphere as carbon dioxide is an important topic as it is closely linked to climate change. Carbon balance equations for x_3 , x_5 , x_6 , x_7 are given below:

$$\frac{dx_3}{dt} = f_{31}k_1x_1 - k_3x_3 \quad (3.9)$$

$$\frac{dx_5}{dt} = f_{53}k_3x_3 + f_{54}k_4x_4 + f_{56}k_6x_6 + f_{57}k_7x_7 - k_5x_5 \quad (3.10)$$

$$\frac{dx_6}{dt} = f_{64}k_4x_4 + f_{65}k_5x_5 - k_6x_6 \quad (3.11)$$

$$\frac{dx_7}{dt} = f_{75}k_5x_5 + f_{76}k_6x_6 - k_7x_7 \quad (3.12)$$

A further example of a multiple carbon pool is the ORCHIDEE model. Figure 3.6 shows the litter and soil organic carbon (SOC) component of ORCHIDEE. As a rule of thumb, around 50% of soil organic matter is SOC. Later versions of this model separate SOC into different layers according to soil depths. However, the version depicted in Figure 3.6 tracks seven pools – four litter pools and three SOC pools – with different layers lumped together. Therefore, we have seven carbon balance equations. The idea is the same as what we have already worked through for the TECO model. We track the arrows that go into and out of each pool. For example, for the active SOC pool, x_5 , we have six arrows that go into this pool. Each arrow represents an incoming flux from a different pool, either litter or SOC. We will use f_{51} , for example, to represent the fraction of the carbon flux goes from the 1st pool (aboveground metabolic litter) to the fifth pool (active SOC).

When we have a small number of carbon pools, it is not difficult to write down the carbon balance equations one-by-one. It would be very tedious to write down all the equations if we have many pools, for example, 100 for the version of ORCHIDEE that simulates SOC dynamics at different soil depths.

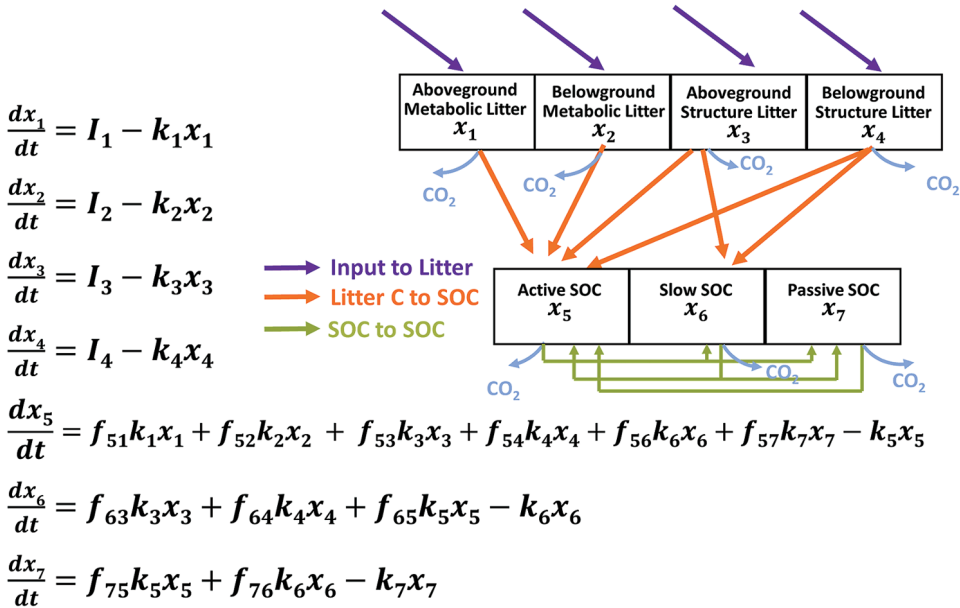


Figure 3.6. Flowchart and carbon balance equations of the ORCHIDEE model.

This version, known as ORCHIDEE-MICT, tracks seven types of organic carbon (four litter compartments and three SOC compartments), similarly to the version depicted in Figure 3.6. ORCHIDEE-MICT has 32 soil layers (Huang, et al., 2018a). For each soil layer, the model has active, slow and passive SOC pools. Such a vertical soil discretization is especially helpful and more realistic in modeling soil carbon dynamics in permafrost regions. In total, the model has $32 \times 3 + 4 = 100$ pools. Instead of writing down the carbon balance equation one-by-one, we can put them together into one matrix equation, as shown in Figure 3.7. The matrix form of the carbon balance equations says the same thing: the rate of change in carbon pool

size equals the input minus the output. Now it is not that obvious to spot out the input fluxes and the output fluxes. They are folded into the matrices and depend on the sign of the elements in these matrices. The first item on the right side of the matrix equation in Figure 3.7 summarizes the external carbon input into the system. The second block of matrices summarizes the turnovers and transfers of fluxes among different organic carbon pools. The third block of matrices captures vertical processes between the adjacent soil layers, such as bioturbation, diffusion, and advection.

Variations in these matrices reflect structural or parametric differences among models. For example, the left side of Figure 3.8 shows the

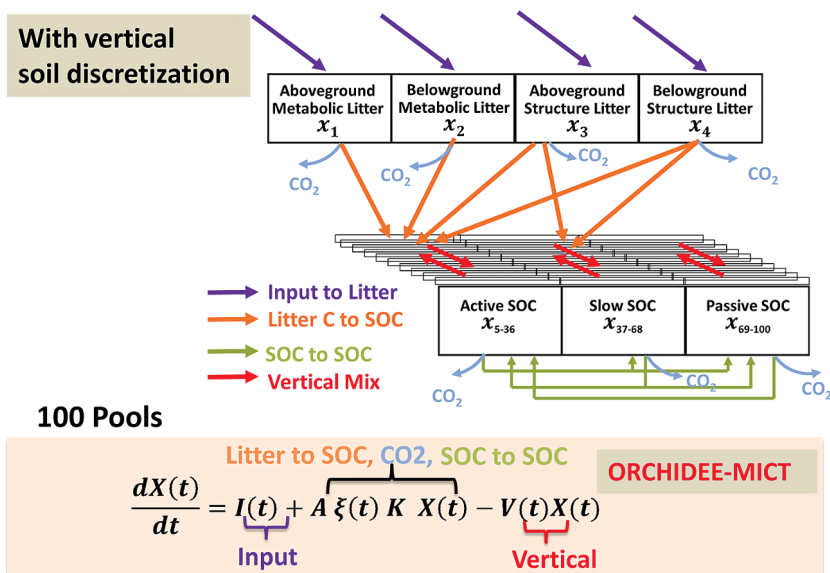


Figure 3.7. Flowchart and matrix-form carbon balance equation of the ORCHIDEE-MICT model with vertical soil discretization.

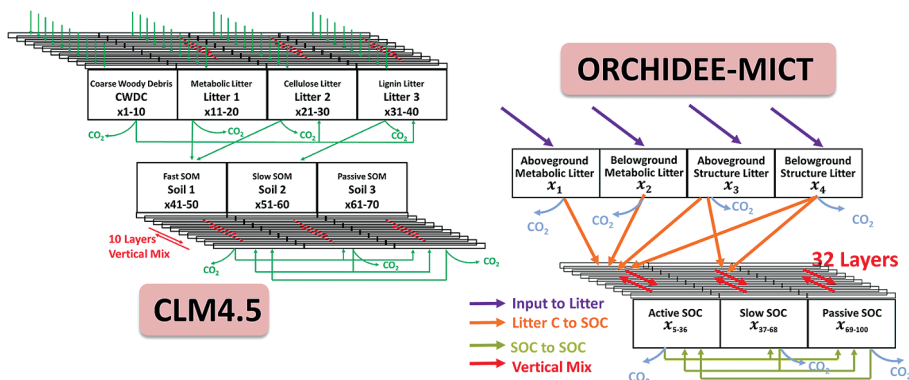


Figure 3.8. Flowcharts of the litter and soil organic carbon components of CLM4.5 and ORCHIDEE-MICT.

flow diagram of the litter and SOM component of the CLM4.5 model. CLM4.5 tracks coarse woody debris, metabolic litter, cellulose litter, lignin litter, fast, slow, and passive SOM in ten soil layers (Huang et al., 2018b). Different from ORCHIDEE-MICT, CLM4.5 also tracks the vertical distribution of litter. The matrix equation here takes a similar general form. Elements in each matrix are different. The dimension of the CLM4.5 matrices are 70×70 corresponding to 70 organic matter pools, while ORCHIDEE-MICT matrices are 100×100 .

We will explore matrix representations in more detail in the following chapter. If you have understood the concepts introduced here, then big congratulations. These complex land carbon models are built upon these basics step by step.

SUGGESTED READING

Luo YQ, Weng ES. (2011) Dynamic disequilibrium of the terrestrial carbon cycle under global change. *Trends in Ecology & Evolution*, **26**, 96–104.

QUIZZES

1. What does a pool stand for?
2. What does a flux mean?
3. What is the principle of writing carbon balance equation?
4. Does burning of the organic matter by fire violate the carbon balance?

CHAPTER FOUR

Practice 1

CARBON FLOW DIAGRAM AND CARBON BALANCE EQUATIONS

Yuanyuan Huang

Climate Science Centre, CSIRO, Canberra, Australia

CONTENT

Introduction / 31

The exercises are designed to help you practice writing carbon balance equations from a carbon flow diagram and *vice versa*. The goal is to lead you to understand the concepts of flow, pool, and mass balance. Please refer to Chapter 3 if you have difficulties with the exercises.

INTRODUCTION

Land biogeochemical models simulate carbon cycling through different pools in an ecosystem all according to carbon balance equations, regardless of model structures. The balance equations are based on the law of conservation of mass which tells us that the change of one carbon pool size is always equal to the net difference between the incoming and outgoing carbon fluxes. Different pools in the ecosystem are connected via carbon transfer between them. To track carbon flows among different carbon pools, scientists usually

develop a carbon flow diagram that uses boxes to indicate pools and arrows to indicate fluxes. Thus, the carbon flow diagram is often called a box-arrow diagram. The models that track carbon flows among pools are often called pool-flux models. The matrix form of land carbon cycle models is built upon the carbon balance equations that are connected via carbon transfer among pools. To learn the matrix approach to land carbon cycle modeling, it is essential to understand the carbon flow diagram and carbon balance equation.

The first exercise of this unit is focused on writing carbon balance equations from a carbon flow diagram, whereas the second exercise is about drawing the carbon flow diagram according to carbon balance equations. Both drawing the carbon flow diagram from the carbon balance equations and writing the equations from the diagram assist us in understanding land carbon dynamics and developing matrix models.

EXERCISE 1: Writing carbon balance equations for the CENTURY model

Figure 4.1 is the carbon flow diagram of the CENTURY (simplified) model. Plant residue is divided into structural and metabolic litter

according to its lignin and nitrogen content. Structural and metabolic litters are decomposed by microbes. The resulting microbial products become the substrate for the formation of soil organic matter in three soil pools (i.e., active, slow, and passive). For the decomposition flux

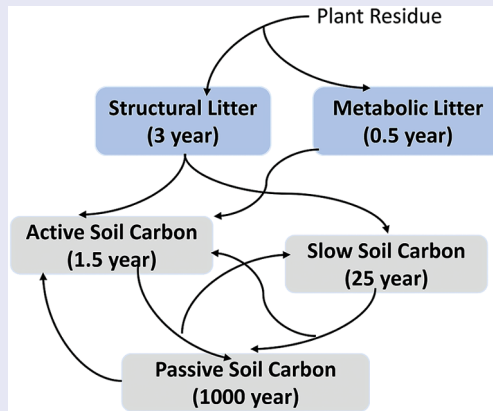


Figure 4.1. Flow diagram for the carbon flows in the CENTURY model (simplified). Reproduced from Parton et al. (1988).

of the structural litter, a fraction (A) is incorporated into the slow soil organic carbon with a turnover time of 25 years. We assume some of this remaining fraction goes as microbial respiration, and the other part contributes to the active soil organic C. The flow diagram also tells us the transfer fluxes among different soil carbon pools.

Please write down the carbon balance equations based on the carbon flow diagram (Figure 4.1). Note that some details in the diagram such as L/N , $1-A$, $F(T)$ and CO_2 fluxes to the atmosphere are not shown as in the original paper by Parton et al. (1988). You could neglect information such as $(3y)$ in the box for structural C and other similar boxes when you do this exercise. You might read the paper by Parton et al. (1988) if you are interested in this model.

To write the balance equations, you could use symbols as in Chapter 3 or develop your own symbol system. Fundamentally, you need to figure out the amount of carbon moving into one pool, the amount of carbon moving out of the pool, and changes in the amount of carbon in the pool (i.e., pool size change) in one unit of time. Here are a few tips. First, you could define state variables, such as x_1 to represent the pool of structural C, x_2 to represent the pool of metabolic C, etc. Then, the size change in the pool of structural C can be expressed by dx_1/dt . You could similarly write changes in carbon amounts in other pools. Second, you need to define rate variables to represent rates of carbon moving out of pools, such as k_1 for a rate of carbon moving out of the structural C

pool, k_2 for a rate of carbon moving out of the metabolic C pool, etc. With the rate of carbon moving out of a pool and a state variable of the pool, you might figure out how to calculate the amount of carbon leaving the pool. Then, you need to figure out the amount of carbon moving to a pool. The third step is to define the total amount of carbon input, let's say using a symbol I , from plant residue. Fourth, you need to figure out a fraction of the incoming carbon that enters a particular pool. For example, carbon input from plant residue partly goes to the structure C pool and partly goes to the metabolic pool. You need to define a symbol, b_1 , as a fraction of incoming carbon going to the structural C pool and b_2 to the metabolic pool. Then the amount of carbon moving to the metabolic C pool is $b_2 \times I$. Similarly, you can figure out the amount of carbon moving to the structural C pool. However, it becomes much more complicated to figure out the amount of carbon that moves to the active soil C pool. The pool receives carbon from all the other four pools. Thus, it has four terms, which are: $f_{31} * k_{1x1} + f_{32} * k_{2x2} + f_{34} * k_{4x4} + f_{35} * k_{5x5}$. Can you figure out how to get each of the terms? As a note, the symbol f_{31} describes the fraction of the carbon that leaves pool 1 and enters pool 3. If you can go over these steps successfully, you may obtain one carbon balance equation, $\frac{dx_4}{dt} = f_{41} * k_{1x1} + f_{43} * k_{3x3} - k_{4x4}$, for the slow soil C pool. What are the carbon balance equations for the other pools?

Exercise 2: Drawing a carbon flow diagram for the ReSOM model based on its carbon Balance equations

This exercise aims to develop your ability to draw a carbon flow diagram from carbon balance equations using the ReSOM model. ReSOM is a REaction-network-based model of Soil Organic Matter and microbes (Tang and Riley, 2013; Tang and Riley, 2015). The model simulates depolymerization of polymers,

mineralization and sorption of monomers by microbes via extracellular enzymes and microbial assimilation, microbial death, and necromass decomposition.

Key equations that govern exchange of carbon among pools over time in the ReSOM model are listed in the following Table 4.1 together with definitions of symbols and parameters. All pools are in units of carbon mass per soil volume (g C m^{-3}).

TABLE 4.1
Governing equations of RaSOM, pools and parameters and their definitions

Pool	Description	Differential equation
S	polymeric organic carbon	$\frac{dS}{dt} = I_S - F_S + \gamma_{B1}B + f_E\gamma_E E$ (4.1)
D	monomeric organic carbon	$\frac{dD}{dt} = I_D + F_S - F_D + \gamma_{B1}X + (1 - f_E)\gamma_E E$ (4.2)
X	reserve microbial biomass	$\frac{dX}{dt} = Y_X F_D - (\kappa - g + \gamma_{B1})X$ (4.3)
B	structural microbial biomass	$\frac{dB}{dt} = mX - (\gamma_{B1} + p_E)B$ (4.4)
E	extracellular enzymes	$\frac{dE}{dt} = p_E B - \gamma_E E$ (4.5)

where:

I_S : polymeric input flux ($\text{g C m}^{-3} \text{ d}^{-1}$)

I_D : monomeric input flux ($\text{g C m}^{-3} \text{ d}^{-1}$)

F_S : polymeric depolymerization flux ($\text{g C m}^{-3} \text{ d}^{-1}$)

F_D : monomeric uptake flux ($\text{g C m}^{-3} \text{ d}^{-1}$)

Y_X : yield coefficient for reserve biomass (unitless)

f_E : fraction of decayed extracellular enzymes contributing to the polymer pool (unitless)

γ_{B1} : microbial mortality rate (d^{-1})

γ_E : enzyme turnover rate (d^{-1})

κ : metabolic turnover rate (d^{-1})

g : growth rate (d^{-1})

p_E : enzyme production rate (d^{-1})

m : structural microbial biomass formation rate (d^{-1})

Modified from Tang and Riley (2015) by Rose Z. Abramoff.

To draw a carbon flow diagram of the ReSOM model from its carbon balance equations, you need to figure out the following items: (1) the number of pools and what they are; (2) carbon transfer between any of the two pools; (3) carbon input from outside of the system; and (4) carbon

loss from the system. Here is a hint to figure out carbon transfer between two pools from a carbon balance equation: a positive sign before a flux (e.g., F_S in equation 4.2) indicates carbon entering the pool whereas a negative sign before a flux (e.g., F_D in equation 4.2) indicates carbon leaving the pool.

UNIT TWO

Matrix Representation of Carbon Balance



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER FIVE

Developing Matrix Models for Land Carbon Models

Yuanyuan Huang

Climate Science Centre, CSIRO, Canberra, Australia

CONTENTS

What is the Matrix Version of the Carbon Balance Equation? / 37

How to Derive the Matrix Equation? / 39

Suggested Reading / 42

Quizzes / 43

The goal of this chapter is to understand how the carbon balance of a system could be represented in matrix form, to be able to write a matrix carbon model and, for advanced participants, to think about potential applications of the matrix model for your own research.

WHAT IS THE MATRIX VERSION OF THE CARBON BALANCE EQUATION?

We briefly introduced the matrix representations of the carbon balance of a multipool system in Chapter 3. Instead of writing down the carbon balance equations one-by-one for each pool, we can integrate all carbon balance equations within one matrix equation. This matrix equation captures the carbon balance for each carbon pool and its linkages to other carbon pools of the system via the flows (fluxes) between them. One thing to keep in mind is that the matrix version of the carbon model, though it uses a different notation, is mathematically identical to the original carbon model that has one carbon balance equation for each carbon pool. Benefits of the matrix version of carbon models include simplicity in model structure, high modularity in coding, clarity in diagnostics, and computational efficiency in spin-up. These properties are discussed in detail in the following Chapters: 6, 9, 14, 17, and 18.

The matrix version of many carbon models could be written as:

$$\frac{dX}{dt} = B * I + A * K * X \quad (5.1)$$

Each of these matrices could be time-dependent. For simplicity, we neglect the time dependence for now. The left-hand side of equation 5.1 depicts the change rates of carbon pool sizes in our system. If we have, for example, seven pools, X here is a column vector with seven rows, one for each pool, which are the state variables of the system. The right-hand side of equation 5.1 captures the incoming and outgoing terms, i.e., the fluxes entering or leaving the pools of the system. B is the allocation matrix, which functions to partition the external carbon input I into different carbon pools. In the plant-soil system of Chapter 3 (Figure 3.2), you could think of I as the photosynthetic carbon input, and B captures the fraction of plant-assimilated carbon that is allocated into different plant organs. K is the turnover rate matrix. A is the $n \times n$ dimensional transfer matrix, where n is the number of carbon pools. A captures the fraction of outgoing carbon flow from one pool that goes into a second pool. From the location and sign of elements in A , we could reconstruct the network of carbon transfers among different carbon pools.

The flow diagram and carbon balance equations of the TECO model were presented in Chapter 3 (Figure 3.5b, equations 3.6–3.12). In the case of TECO, which has seven carbon pools, B and X are 7×1 vectors, while A and K are 7×7 matrices. K is a diagonal matrix with each diagonal element k corresponding to the turnover rate of a carbon pool. All other elements in K are set to zero. The diagonal elements of A are all -1 . Non-zero elements of the off-diagonal parts of A correspond to the

fraction of carbon fluxes transferred from the i^{th} to the j^{th} pool, where i is the column number and j the row number, starting at $(1, 1)$ from the top-left corner of the matrix. For example, f_{41} in the first column and fourth row tells us there are carbon fluxes transferred from the first to the fourth carbon pool, that is, from foliage biomass to structure litter. All transfer fluxes in the carbon flow diagram can be found in the A matrix. In the case of TECO, equation 5.1 can be expanded as:

$$\begin{pmatrix} dx_1(t)/dt \\ dx_2(t)/dt \\ dx_3(t)/dt \\ dx_4(t)/dt \\ dx_5(t)/dt \\ dx_6(t)/dt \\ dx_7(t)/dt \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ f_{31} & 0 & -1 & 0 & 0 & 0 & 0 \\ f_{41} & f_{42} & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & f_{53} & f_{54} & -1 & f_{56} & f_{57} \\ 0 & 0 & 0 & f_{64} & f_{65} & -1 & 0 \\ 0 & 0 & 0 & 0 & f_{75} & f_{76} & -1 \end{pmatrix} \times \begin{pmatrix} k_1 & & & & & & \\ & k_2 & & & & & \\ & & k_3 & & & & \\ & & & k_4 & & & \\ & & & & k_5 & & \\ & & & & & k_6 & \\ & & & & & & k_7 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix}$$

This matrix version of the carbon model is especially handy when we have many carbon pools. We will take the litter and soil organic carbon components of CLM 4.5 (70 pools) and ORCHIDEE-MICT (100 pools) as examples. The matrix equation for CLM4.5 and ORCHIDEE-MICT look similar to the matrix equation for TECO. In Figure 5.1 we have one more item: the V matrix, which captures the vertical transfers of carbon between adjacent soil layers. The A matrix is more complex than the earlier case. Both CLM4.5 and ORCHIDEE-MICT assume the fraction of carbon transferred between different organic carbon pools is the same for different soil depths. Here, A has the form of a block matrix, which means A is made up by smaller matrices, i.e., blocks. For CLM4.5, each block, for example, A_{31} , is a 10×10 matrix, corresponding to 10 soil layers. The diagonal elements

in A_{31} are the same since the model assumes the fraction of transfers do not change with depth. For ORCHIDEE-MICT, it is a little more complex as litter pools are not separated by soil depth. So transfers of litter carbon fluxes to soil carbon fluxes (for example, A_{51}) are 32×1 vectors and zeros. For transfers among soil organic carbon pools, for example A_{65} , the matrix has a dimension of 32×32 , with each element representing the transfer from one of the 32 active soil organic carbon pools to one of the 32 slow soil organic carbon pools. ξ is the matrix of environmental scalars that quantify the deviation of the actual decomposition rate from the potential rate due to the non-optimal environmental conditions. We could add it to the TECO model also. Please check in Huang et al. (2018b) and Huang et al. (2018a) for more details if you are interested.

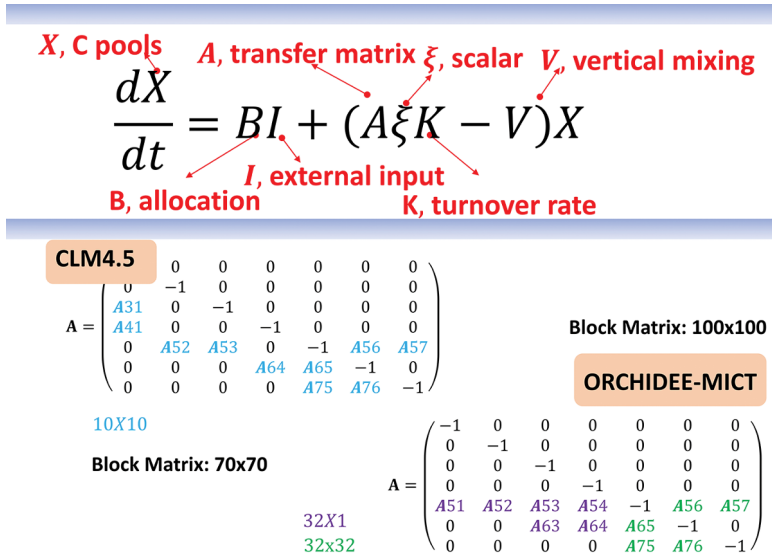


Figure 5.1. Matrix equation of CLM4.5 and ORCHIDEE-MICT.

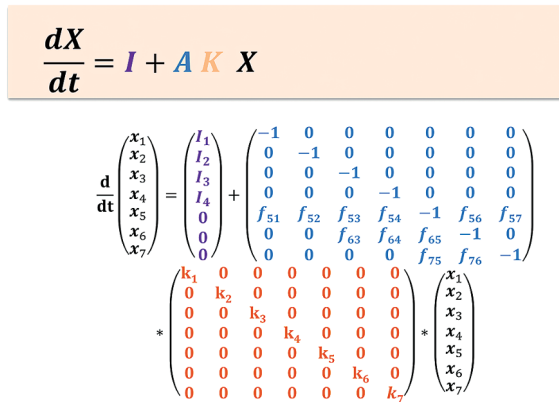


Figure 5.2. Matrix equation of the litter and soil organic carbon component of the ORCHIDEE model. The flow diagram and carbon balance equations are shown in Figure 3.6, Chapter 3.

Based on these examples, we see that the matrix equations for different carbon models may take a similar form. But the structure of each matrix might be different. How each matrix is organized depends on the structure of the original model.

HOW TO DERIVE THE MATRIX EQUATION?

After we know what the matrix looks like, the next step is to derive the matrix model from the carbon balance equations.

Here we take the ORCHIDEE model in Figure 3.6 of Chapter 3 as an example. We assume we have

I_1, I_2, I_3 and I_4 as inputs to the respective litter pools. Carbon from the litter pools is transferred into different soil organic carbon pools. Additionally, there are internal carbon transfers between different soil organic carbon pools. Figure 5.2 shows the matrix equation of this model. X and I are column vectors, and A and K are 2-dimensional matrices. The dimensions of X , I , A and K are determined by the number of carbon pools. X is easy to construct. We put the seven carbon pools x_1, x_2, \dots, x_7 in a column. I is a column vector with I_1 to I_4 as the input rates for litter carbon pools and zero for soil organic carbon pools as these pools receive no external

How to construct the transfer matrix A?

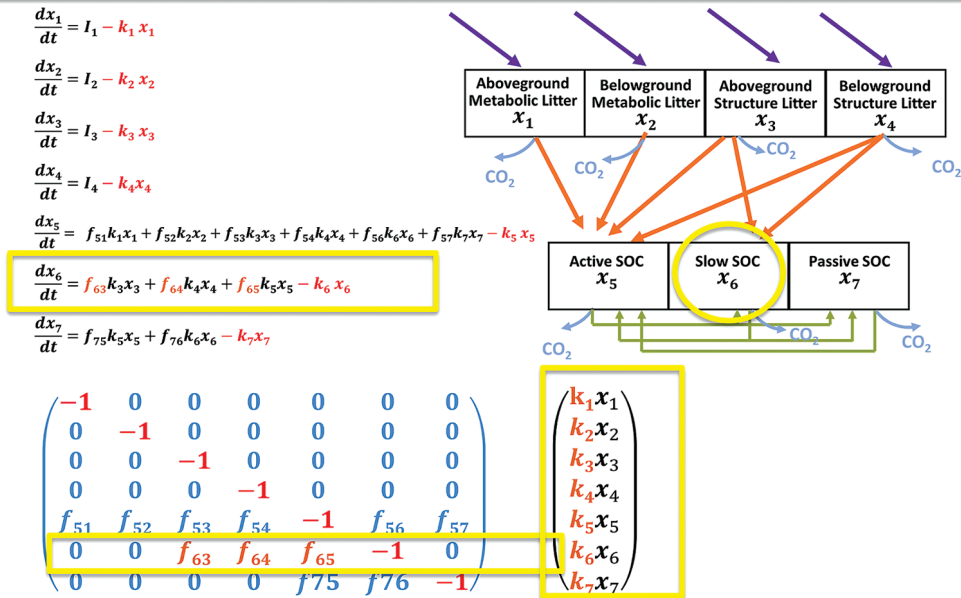


Figure 5.3. Illustration of the function of the transfer matrix A.

carbon inputs. K is a diagonal matrix with each diagonal element corresponding to the turnover rate of one of the carbon pools.

Mathematically, a $n \times n$ matrix (A matrix) multiplied by a $n \times 1$ vector (X) will produce a $n \times 1$ vector. The value of i^{th} row of this multiplication equals the sum of the products between elements from the i^{th} row with the column vector. The multiplication of K and X will give us a column vector with elements $k_1x_1, k_2x_2, \dots, k_7x_7$. For the multiplication of A by KX , we take the sixth row as an example (Figure 5.3). We have 0 multiply k_1x_1 , 0 multiply k_2x_2 , f_{63} multiply k_3x_3 , f_{64} multiply k_4x_4 , f_{65} multiply k_5x_5 , and -1 multiply k_7x_7 , which is exactly the right side of the carbon balance equation for the sixth carbon pool. The diagonal elements of A are all -1. There are many 0 elements in A . These 0s indicate there are no carbon transfers from the i^{th} (column number) to j^{th} (row number) carbon pools.

If we look at each column of the A matrix, it summarizes all the fluxes that go out of the i^{th} carbon pool (Figure 5.4). For example, after reading the fourth column of the A matrix, we know there is carbon transferred from the fourth to the fifth and sixth carbon pool if f_{54} and f_{64} are not equal to 0.

This is exactly what we see in the carbon flow diagram (Figure 3.6 of Chapter 3). If we sum up all the off-diagonal elements in one column and the value is smaller than 1, for example, $f_{54} + f_{64} < 1$, this means there is carbon leaving the system which we do not track further. For example, if our model is focused on the storages of organic carbon in soil, we may not care about carbon that leaves the soil system in the form of CO_2 flux to the atmosphere.

Each row of the matrix A summarizes all the carbon fluxes that are transferred into the i^{th} carbon pool. For example, if we look at the sixth row, we have f_{63}, f_{64}, f_{65} as non-zero, off-diagonal elements. This tells us there are fluxes from the third, fourth and fifth carbon pools being transferred into this sixth carbon pool. The relevant transfers are highlighted in yellow in the carbon flow diagram of Figure 5.5.

We have seen that a single matrix equation can encompass the details of multiple processes simulated by a carbon model. Detailed information about the carbon dynamics is folded into different components of the matrices. If we expand the matrix calculation by row, we should get the exact set of carbon balance equations, one equation for each pool.

$$\begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ f_{51} & f_{52} & f_{53} & f_{54} & -1 & f_{56} & f_{57} \\ 0 & 0 & f_{63} & f_{64} & f_{65} & -1 & 0 \\ 0 & 0 & 0 & 0 & f_{75} & f_{76} & -1 \end{pmatrix}$$

A matrix by column:
summarize flux transferred **out from** the *i*th carbon pool

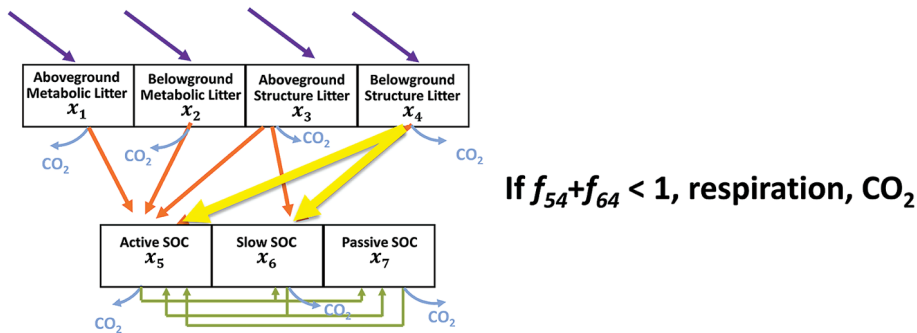


Figure 5.4. Illustration of the function of the column of the transfer matrix (A matrix).

$$\begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ f_{51} & f_{52} & f_{53} & f_{54} & -1 & f_{56} & f_{57} \\ 0 & 0 & f_{63} & f_{64} & f_{65} & -1 & 0 \\ 0 & 0 & 0 & 0 & f_{75} & f_{76} & -1 \end{pmatrix}$$

A matrix by row:
summarize fluxes transferred **into** the *i*th carbon pool

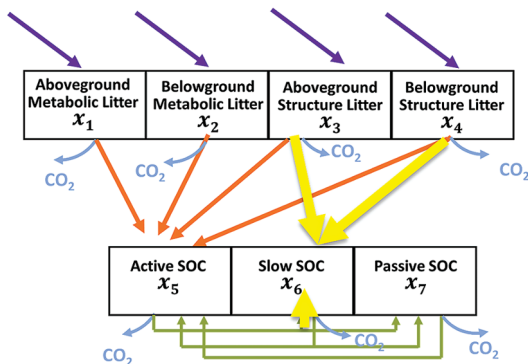


Figure 5.5. Illustration of the function of the row of the transfer matrix (A matrix).

$$\frac{dX}{dt} = I + AKX - VX$$

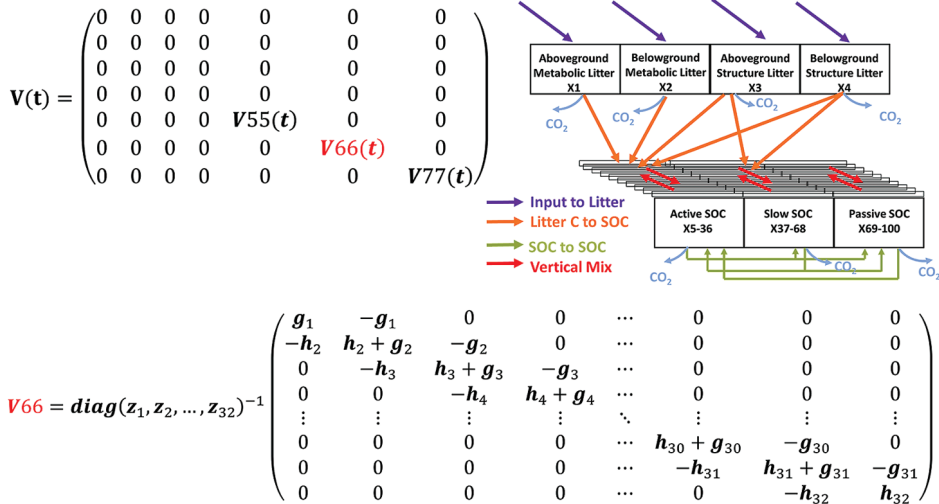


Figure 5.6. Illustration of the vertical transfer matrix V of the ORCHIDEE-MICT model.

Next we are going to explore a little more on the matrix equation for models with vertical soil layers. For such a model, we require one more vertical transfer matrix, V (Figure 5.6). The X , I and K matrices are generally constructed in a same way as in the earlier examples above. The transfer matrix A , with 100×100 dimension, is folded into the block matrix. Each of the smaller blocks in black, for example $A51$, is a depth-dependent vector, tracking the fraction of litter carbon transferred into SOC in different soil depths. The matrices in red are 32×32 diagonal matrices capturing transfers between different soil organic carbon categories in the same soil layer. Each item of the diagonal takes the same value, as ORCHIDEE-MICT assumes the transfer fractions are not depth-dependent.

The vertical transfer matrix can be represented by the block matrix shown in Figure 5.6. Most of its components are 0 except for the active, slow, and passive SOC pools. Each diagonal block is a tridiagonal matrix that describes vertical redistribution of corresponding carbon pools among different soil layers. As the vertical transfer rates are not differentiated among different types of carbon pools, $V55(t)$, $V66(t)$, and $V77(t)$ are identical. The subscript numbers indicate soil layers; h and g correspond respectively to the mixing rate between

the current soil layer and the one above it, and the current soil layer and the one below it; z_i indicates the depth of soil layer i . Detailed information is available in Huang et al. (2018a).

In this chapter we have learned how to derive the matrix equation from the carbon balance equations. It would also be possible to derive the matrix equation directly from the carbon flow diagram. From the carbon flow diagram, we can determine the dimension of matrices (or vectors) from the number of carbon pools. The diagram also tells us the external carbon inputs, the incoming and outgoing fluxes of each pool, and the direction of these fluxes. These are basically what we need to construct the matrix equation.

SUGGESTED READING

- Huang, Y. Y., Lu, X. J. et al. 2018. Matrix approach to land carbon cycle modeling: A case study with the Community Land Model. *Global Change Biology* **24**, 1394–1404, doi:10.1111/gcb.13948.
- Huang, Y. Y., Zhu, D. et al. 2018. Matrix-Based Sensitivity Assessment of Soil Organic Carbon Storage: A Case Study from the ORCHIDEE-MICT Model. *Journal of Advances in Modeling Earth Systems* **10**, 1790–1808, doi:10.1029/2017ms001237.

QUIZZES

1. Is the sum of each row of the transfer matrix A always not bigger than 0?
2. Is the sum of each column of the transfer matrix A always not bigger than 0?
3. How would you add a new carbon pool into the matrix equation?
4. Why is a matrix equation exactly the same as a carbon balance equation in terms of describing the carbon cycle?



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER SIX

Coupled Carbon-Nitrogen Matrix Models

Zheng Shi

University of Oklahoma, Norman, USA

Xingjie Lu

Sun Yat-sen University, Guangzhou, China

CONTENTS

Introduction / 45

Matrix Representation of C-N Coupling in Terrestrial Ecosystem (TECO) Model / 45

Application of Matrix Representation of C-N Coupled Model / 47

Matrix Representation of C-N Coupling in CLM5 / 47

Global Validation of the CLM5 Matrix Model for C and N Simulations / 55

Suggested Reading / 56

Quizzes / 56

Carbon sequestration in terrestrial ecosystems is strongly regulated by nitrogen processes. Many global land models now simulate the carbon and nitrogen interaction. The goal of this chapter is to understand how coupled carbon and nitrogen models at ecosystem and global scales may be represented in the matrix form. Basically, nitrogen transfers among all the organic nitrogen pools can be represented in one matrix equation that is equivalent to the carbon matrix equation. The carbon and nitrogen matrix equations are coupled through the C:N ratio. Mineral nitrogen dynamics, determined by nitrogen input, mineralization, plant uptake and leaching, can also be described by one equation.

INTRODUCTION

Rising atmospheric carbon dioxide (CO₂) concentration tends to induce carbon (C) sequestration in terrestrial ecosystems. The conceptual framework of progressive nitrogen (N) limitation has predicted N limitation on future C sequestration in terrestrial ecosystems in response to rising atmospheric CO₂. The N limitation may become

progressively stronger over time unless N fixation is stimulated and/or N losses are reduced, leading to increased N capital (Luo et al., 2004). In addition, the degree of N regulation on terrestrial C sequestration depends on changes in several C-N coupling parameters, such as the stoichiometric flexibility of C:N ratios of biomass compartments, changes in plant N uptake via soil exploration, and N redistribution from soil to vegetation. Encoding what we know about how C and N flow within ecosystems, and how these flows are coupled, can help to fully understand the strength of N regulation on C sequestration in terrestrial ecosystems.

In this chapter, we will show how C and N cycling are coupled in an ecosystem model and a global land model, and how the coupled processes can be represented in the matrix form.

MATRIX REPRESENTATION OF C-N COUPLING IN TERRESTRIAL ECOSYSTEM (TECO) MODEL

One of the coupled C and N models presented in this chapter is developed from the terrestrial ecosystem (TECO) model (Weng & Luo, 2011). TECO

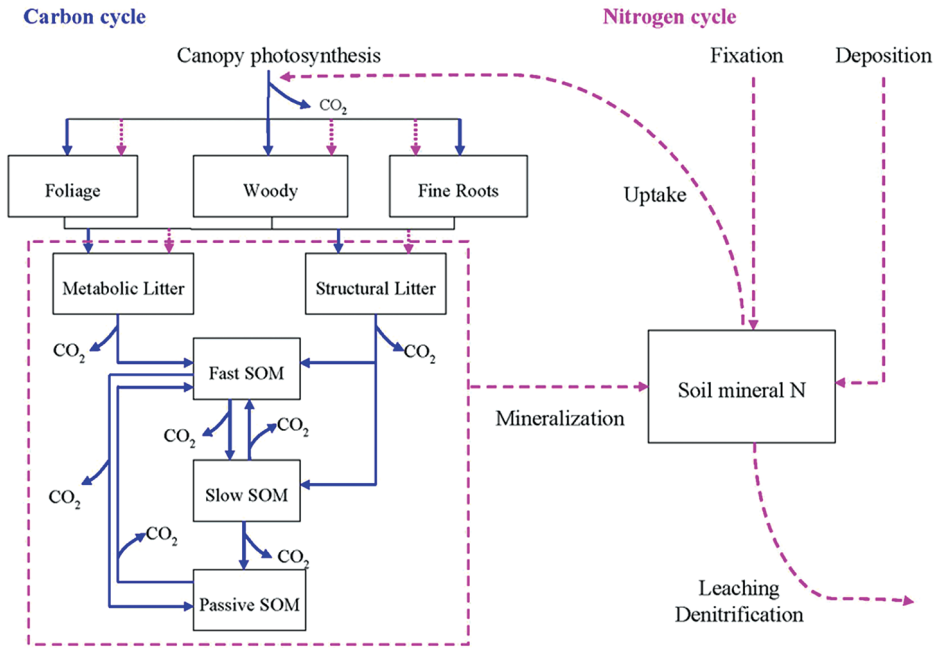


Figure 6.1. Carbon and nitrogen pools and pathways of carbon and nitrogen fluxes in the TECO-CN model. Blue arrows show carbon cycling processes, while pink arrows indicate nitrogen cycling processes. SOM = soil organic matter.

was first presented in Chapter 1, and its matrix representation introduced in Chapter 5. The coupled C-N version we will discuss here, called TECO-CN, has eight C and N pools in addition to one mineral N pool. C enters the ecosystem through canopy photosynthesis and is then allocated into foliage (X_1), wood biomass (X_2) and fine roots (X_3) (Figure 6.1). Similarly, N is absorbed by plants from mineral soil, and then partitioned among leaf (N_1), woody tissues (N_2) and fine roots (N_3). Detritus from plant biomass turnover is transferred to metabolic litter (X_4) and structure litter (X_5) pools, where it is decomposed by microbes (X_6). The structure litter (X_5) is partly respired while partly converted into fast (X_6) and slow soil organic matter (SOM, X_7). During these C cycling

processes, N in plant detritus is also transferred among the same set of ecosystem pools (i.e., litter, fast, slow and passive SOM). Mathematically, these C processes may be described by the following first-order ordinary differential equation:

$$\frac{dX}{dt} = BI + A \xi(t) K X(t) \quad (6.1)$$

where $X = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$, represents C pools in leaf, wood, fine roots, metabolic litter, structure litter, microbe, slow and passive SOM, respectively. $\xi(t)$ is a vector of environmental scalars accounting for temperature and moisture effects on all C decomposition. A describes C transformation among various ecosystem compartments, given as:

$$A = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ f_{4,1} & f_{4,2} & f_{4,3} & -1 & 0 & 0 & 0 & 0 \\ 1-f_{4,1} & 1-f_{4,2} & 1-f_{4,3} & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & f_{6,4} & f_{6,5} & -1 & f_{6,7} & f_{6,8} \\ 0 & 0 & 0 & 0 & f_{7,5} & f_{7,6} & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & f_{8,6} & f_{8,7} & -1 \end{pmatrix}$$

The non-zero elements ($f_{i,j}$) in matrix A describe C transfer coefficients (i.e., the fractions of the C entering the i^{th} pool from the j^{th} pool), while the zero elements indicate no C flows between these two pools. K is an 8×8 diagonal matrix with diagonal entries given by vector $K = (k_1, k_2, k_3, k_4, k_5, k_6, k_7, k_8)$ is baseline turnover rate of carbon pools (i.e., the amounts of C per unit mass leaving each pool per day). I is the input and $B = (b_1, b_2, b_3, 0, 0, 0, 0, 0)$ is the allocations of input into ecosystem carbon pools.

The N processes can be described by:

$$\frac{dN}{dt} = A\xi(t)KR^{-1}X(t) + \kappa_u N_{\min}(t)\Pi \quad (6.2)$$

where $N = (n_1, n_2, n_3, n_4, n_5, n_6, n_7, n_8)$, represents N pools in leaf, wood, fine roots, metabolic litter, structure litter, microbe, slow and passive SOM, respectively. R is an 8×8 diagonal matrix with diagonal entries given by vector $R = (r_1, r_2, r_3, r_4, r_5, r_6, r_7, r_8)$, representing C:N ratios in the eight organic N pools. $\Pi = (\pi_1, \pi_2, 1 - \pi_1 - \pi_2, 0, 0, 0, 0, 0)$, is a vector of allocation coefficients expressing the proportion of plant N uptake that is allocated to leaf, wood and fine root biomass pools. κ_u is the rate of plant N uptake per time step (year), expressed as a proportion of $N_{\min}(t)$, the amount of plant available N (mineral N) in soil at time t . The dynamics of the mineral N pool are determined by balance between N input (i.e., N mineralization, biological fixation and atmospheric deposition) and output through plant N uptake and N loss (i.e., leaching and gaseous N fluxes), which can be expressed by:

$$\begin{aligned} \frac{d}{dt} N_{\min}(t) = & -(\kappa_u + \kappa_L)N_{\min}(t) \\ & + \xi(t)\varphi_1^* AKR^{-1}X(t) + F(t) \end{aligned} \quad (6.3)$$

where κ_u and κ_L are rates of N uptake and loss, respectively. The second term on the right side of Equation 6.3 describes the amount of N released during mineralization. φ_1^* is the proportion for mineral N production. The C and N cycles are coupled through the parameter R which is C:N ratio. $F(t)$ represents N input through biological fixation and atmospheric deposition.

APPLICATION OF MATRIX REPRESENTATION OF C-N COUPLED MODEL

To illustrate the application of matrix forms of a C-N coupled model, we designed a case study to

examine changes in C-N coupling parameters under CO_2 enrichment using a data assimilation approach. Based on measurements of C and N pools in various ecosystem compartments (i.e., foliage, woody tissues, fine roots, microbe, forest floor soil inorganic N, and mineral soil) and fluxes (i.e., litterfall, soil respiration and mineralization, and plant N uptake, N input from biological fixation and atmospheric deposition) obtained from the Duke Forest Free-Air CO_2 Enrichment (FACE) experiment in North Carolina, USA, during the period 1996–2005. Key parameters of TECO-CN (i.e., C:N ratio, N uptake, N allocation coefficient, N input, N loss and the initial value of mineral soil N pool) were estimated through a Bayesian probabilistic inversion. The inversion was done separately for plots with ambient versus elevated CO_2 treatments, yielding one set of parameters for each (Shi et al., 2016).

Comparison of parameter distributions showed that plant N uptake, C:N ratios in foliage, fine root, metabolic and structural litter were significantly higher under elevated than ambient CO_2 , whereas CO_2 enrichment did not exert significant effects on C:N ratios in wood tissues and SOM. Moreover, elevated CO_2 led to decrease of C exit rates in foliage, woody biomass, structural litter and passive SOM, indicating an increase of C residence time in these ecosystem compartments. By contrast, elevated CO_2 resulted in the increase of C exit rate in fine roots, demonstrating faster fine root turnover under CO_2 enrichment. In addition, C allocation to the foliage became smaller under elevated CO_2 , while C allocation to fine roots tended to be larger under CO_2 enrichment. The estimated parameters were then used for a forward analysis to examine ecosystem C and N dynamics under ambient and elevated CO_2 conditions at Duke Forest. Our results demonstrated that modeled N pools in foliage, woody tissues, fine roots, and forest floor closely matched with the corresponding measurements for both ambient and elevated CO_2 scenarios (Figure 6.2). However, TECO-CN could not capture the observed declining trend of microbial N content with time. In addition, the trained model did not simulate N dynamics in mineral soil well, partly due to the large variations in SOM measurements among different years.

MATRIX REPRESENTATION OF C-N COUPLING IN CLM5

The other C-N coupled model we will examine in this chapter is the Community Land Model version

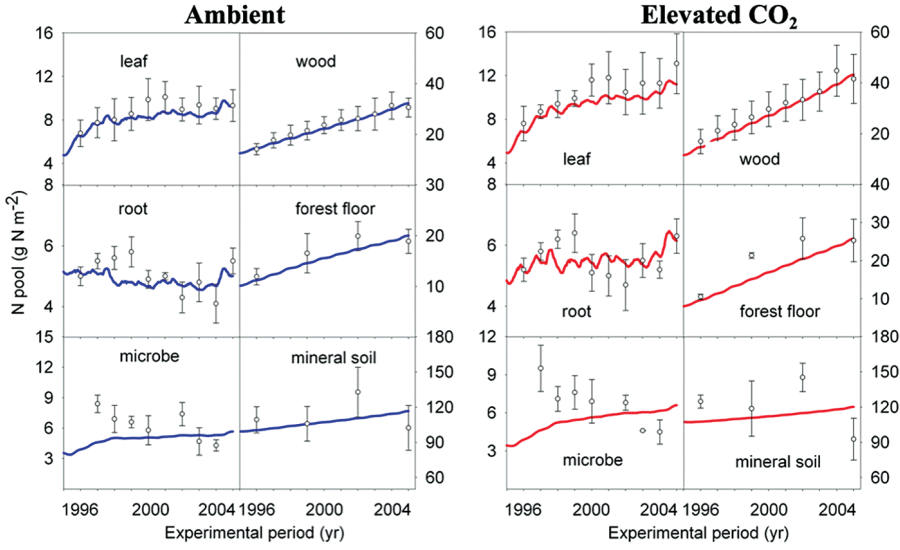


Figure 6.2. The comparisons of modeled v. measured nitrogen pools in various ecosystem compartments under ambient CO_2 (blue lines) and elevated CO_2 (red lines).

5 (CLM5, Figure 6.3). To extend the ecosystem scale to global scale, CLM5 simulations represent various ecosystems on a grid covering the global land area. Like TECO-CN, the CLM5 biogeochemistry module includes carbon and nitrogen cycles for aboveground and belowground processes (Lawrence et al., 2019). The vegetation modules simulate the biogeochemical transfers among 18 carbon pools and 19 nitrogen pools, representing vegetation and soil compartments of a grid cell or tile. Tissue pools, storage pools and transfer pools are included, respectively, in leaf, fine root, live stem, dead stem, live coarse root, and dead coarse root compartments. In addition to C pools, N pools include one more retranslocation pool temporarily storing N from litter fall. The soil biogeochemistry module has 20 soil layers as a default setting. Each layer contains seven pools for organic C and organic N, respectively, in metabolic litter, cellulose litter, lignin litter, coarse woody debris, fast soil organic matter, slow soil organic matter, and passive soil organic matter. Inorganic N pools, such as ammonium and nitrate, interact among each other, or with organic C and organic N pools, or with the environment through multiple nitrogen processes such as nitrification, denitrification, leaching, atmospheric N deposition, and biological N fixation. All these biogeochemical processes and pools in both vegetation and soil modules

can be formulated as carbon or nitrogen balance equations.

We may reorganize the vegetation carbon and nitrogen balance equations of CLM5 into two matrix equations:

$$\frac{dC_{\text{veg}}}{dt} = BI_{\text{Cin}} + \left(A_{\text{phc}}(t)K_{\text{phc}} + A_{\text{gmc}}(t)K_{\text{gmc}} + A_{\text{fic}}(t)K_{\text{fic}} \right) C_{\text{veg}}(t) \quad (6.4)$$

$$\frac{dN_{\text{veg}}}{dt} = BI_{\text{Nin}} + \left(A_{\text{phn}}(t)K_{\text{phn}} + A_{\text{fin}}(t)K_{\text{fin}} \right) N_{\text{veg}}(t) \quad (6.5)$$

C_{veg} and N_{veg} are time-dependent state variables, which are vectors, each entry representing its respective vegetation pool size (gC m^{-2} and gN m^{-2}). I_{Cin} and I_{Nin} are scalars for C and N input, respectively. C input is the net primary productivity, which is the difference between gross primary productivity (i.e., plant photosynthesis) and autotrophic respiration. N input to the vegetation N cycle includes both biological N fixation and N uptake from roots. B_C and B_N are also vectors, representing allocation fractions of plant C and N input to individual pools. The K matrices are $n \times n$ diagonal matrices whose diagonal elements

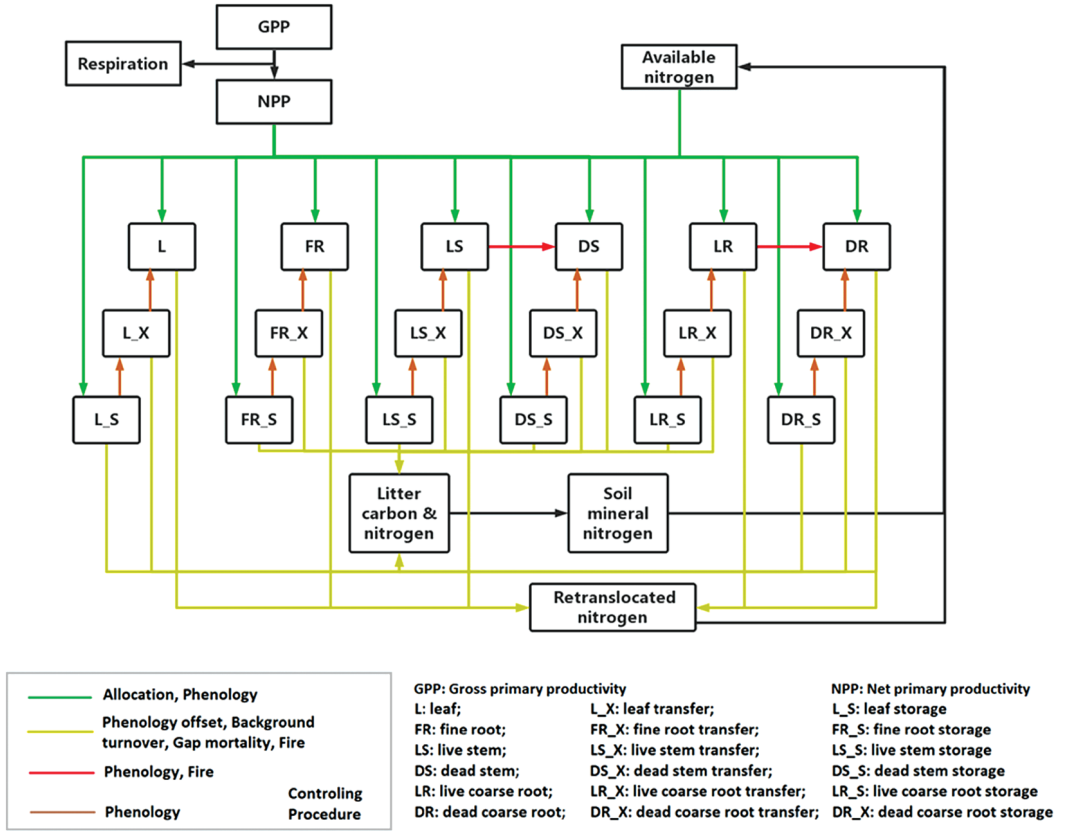


Figure 6.3. Carbon and nitrogen flow diagram of vegetation biogeochemical cycle in CLM5. GPP = gross primary production; NPP = net primary production.

represent turnover rates (pool fraction per annual time step) due to different plant and vegetation processes: subscripts ph , gm , and fi indicate phenology, gap mortality (i.e., harvest from land use and natural mortality), and fire processes, respectively, as described by:

$$K_{ph} = \begin{pmatrix} k_{p1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & k_{pn} \end{pmatrix}$$

$$K_{gm} = \begin{pmatrix} k_{n1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & k_{nn} \end{pmatrix}$$

$$K_{fi} = \begin{pmatrix} k_{f1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & k_{fn} \end{pmatrix}$$

The turnover rates in plant phenology matrix K indicate the leaf, root, live stem, and dead stem turnover due to phenology processes. The exit rates in gap mortality matrix K include harvest rates from land use plus natural mortality. The exit rates in fire matrix K represent the plant C loss rate due to fires.

A is a transfer coefficient matrix, representing C and N transfer among pools as specified in equation 6.6–6.11 below for CLM5. Subscript c and n respectively denote carbon and nitrogen.

The off-diagonal entry, $a_{i,j}$, for matrix A represents a fraction of C or N leaving pool j that goes to pool i . The diagonal entries are set to -1 to represent that all the exiting C leaves pool j . The pool names referred by the subscript i or j can be found in Figure 6.3.

Interactions between vegetation C and N cycles in the original CLM5 are fully preserved in this matrix version of the CLM5 model. The original CLM5 has two modules to regulate C and N interactions for vegetation. First, the photosynthetic capacity, an important variable driving the carbon cycle, interacts with the nitrogen cycle in the LUNA module. LUNA optimizes N allocation to maximize the daily net photosynthetic carbon gain. Second, plant N uptake interacts with the carbon assimilation in the FUN module. FUN is based on the concept that N uptake requires the expenditure of energy in exchange for carbon. Both the modules are fully preserved to drive the C and N interactions in the matrix version of the model.

The implementation of soil C and N cycling extends the maximum soil layers from a fixed value of 10 in CLM4.5 (see Chapter 5) to a default value of 20 with flexibility to change in CLM5.

N balance equations in the original CLM5 are likewise amenable to a matrix representation. The soil organic C and N transfer among soil pools is formalized by the following matrix equations:

$$\frac{dC_{\text{soil}}}{dt} = I_{C_{\text{soil}}} + (A_{\text{hc}}\xi(t)K_{\text{h}} + V(t) + K_{\text{f}}(t))C_{\text{soil}}(t) \quad (6.6)$$

$$A_{ij} = \begin{cases} \begin{bmatrix} -1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -1 \end{bmatrix} & i = j; \text{C and N cycles} \\ \begin{bmatrix} (1-r_{ij})T_{ij} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (1-r_{ij})T_{ij} \end{bmatrix} & i \neq j; \text{C cycle } (A_{\text{hc}}) \\ \begin{bmatrix} (1-r_{ij})T_{ij} \frac{CN_{j,1}}{CN_{i,1}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (1-r_{ij})T_{ij} \frac{CN_{j,20}}{CN_{i,20}} \end{bmatrix} & i \neq j; \text{N cycle } (A_{\text{hn}}) \end{cases}$$

$$\frac{dN_{\text{soil}}}{dt} = I_{N_{\text{soil}}} + (A_{\text{hn}}\xi(t)K_{\text{h}} + V(t) + K_{\text{f}}(t))N_{\text{soil}}(t) \quad (6.7)$$

As with vegetation, C_{soil} and N_{soil} are vectors of state variables, representing soil organic C and N pool sizes in gC m^{-3} and gN m^{-3} , respectively. $I_{C_{\text{soil}}}$ and $I_{N_{\text{soil}}}$ are vectors representing plant litterfall into different litter C and N pools, respectively. A_{hc} and A_{hn} represent the horizontal transfers of C and N, respectively, which means transfers among pools within one soil layer. V stands for the rate of vertical mixing between the same types of pools across soil layers, which is the same for C and N. K_{f} is the rate of fire-induced litter loss, which is the same for C and N.

Matrix A_{h} , including both A_{hc} and A_{hn} , is a block matrix constituted by several matrices as:

$$A_{\text{hcOr } A_{\text{hn}}} = \begin{pmatrix} A_{11} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_{22} & 0 & 0 & 0 & 0 & 0 \\ A_{31} & 0 & A_{33} & 0 & 0 & 0 & 0 \\ A_{41} & 0 & 0 & A_{44} & 0 & 0 & 0 \\ 0 & A_{52} & A_{53} & 0 & A_{55} & A_{11} & A_{11} \\ 0 & 0 & 0 & A_{64} & A_{65} & A_{66} & 0 \\ 0 & 0 & 0 & 0 & A_{75} & A_{76} & A_{77} \end{pmatrix}$$

Each of the matrices A_{ij} within the block matrix A_{hc} or A_{hn} is 20×20 , corresponding to 20 soil layers. The non-zero, off-diagonal matrices A_{ij} , $i \neq j$, indicate C and N transfer among pools within one soil layer as:

Each of the diagonal matrices A_{ij} is a negative identity matrix (i.e. a 20×20 matrix with diagonal elements equal to -1 and all other elements zero). The off-diagonal matrices A_{ij} have non-zero diagonal values, $(1 - r_{ij})T_{ij}$ for carbon and $(1 - r_{ij})T_{ij} \frac{CN_{j,k}}{CN_{i,k}}$ for nitrogen. A_{ij} represents transfer coefficients. r_{ij} is the respired fraction of C along the transfer pathway from pool j to i . T_{ij} represents a pathway fraction of C going to pool i from that leaving the j^{th} pool due to decomposition. $CN_{j,k}$ represents C:N ratio of pool j in layer k .

The diagonal matrices K_h and K_f indicate the turnover rates, respectively, due to decompositionⁿ (horizontal transfer) and fire, at different layers:

$$K_h = \begin{pmatrix} k_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & k_n \end{pmatrix}$$

$$K_f = \begin{pmatrix} k_{f1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & k_{fn} \end{pmatrix}$$

$$v = \text{diag}(dz_1, dz_2, \dots, dz_{20})^{-1} \begin{pmatrix} g_1 & -g_1 & 0 & 0 & 0 & 0 \\ -h_2 & h_2 + g_2 & -g_2 & \cdots & 0 & 0 \\ 0 & -h_3 & h_3 + g_3 & & 0 & 0 \\ & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & & h_{18} + g_{18} & -g_{18} \\ 0 & 0 & 0 & \cdots & -h_{19} & h_{19} + g_{19} \\ 0 & 0 & 0 & & 0 & -h_{20} & h_{20} \end{pmatrix}$$

The subscripts represent the soil layer; g and h are vertical mixing rates related to upward and downward transfers.

As for the vegetation component of CLM5, interactions between C and N cycles in the soil from the original CLM5 model are fully preserved in the matrix version. In the original CLM5, nitrogen limits soil organic C decomposition. The N limitation is represented by the ratio between available mineral N and the total soil N demand. The soil N demand includes soil immobilization during soil decomposition and plant uptake demand. The dynamics of mineral N processes that are involved

Environmental scalars ξ are time-dependent variables and are the product of temperature scalar ξ_T , water scalar ξ_W , oxygen scalar ξ_O , depth scalar ξ_D , and nitrogen scalar ξ_N :

$$\xi(t) = \xi_T(t) \xi_W(t) \xi_O \xi_D \xi_N(t)$$

The vertical mixing coefficient matrix V is made up of 6 identical matrices v :

$$V(t) = \begin{pmatrix} v & 0 & 0 & 0 & 0 & 0 \\ 0 & v & 0 & 0 & 0 & 0 \\ 0 & 0 & v & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & v & 0 \\ 0 & 0 & 0 & 0 & 0 & v \end{pmatrix}$$

Note that vertical mixing of coarse woody debris (CWD) is not allowed in CLM5; therefore, the corresponding vertical mixing matrix is 0 for CWD. The matrix v is a tridiagonal matrix, indicating the vertical mixing only transfers between adjacent layers:

in C and N interactions can be represented by one equation as fully described in Shi et al. (2016). The equation on mineral N dynamics can be coupled with the matrix equations on organic C and N processes to analytically or semi-analytically explore their interactions. The matrix form recoding simplifies their interactions. The matrix representation in complex C and N transform network of a biogeochemistry model like CLM5. Traceable components of the C and N cycles can thereby be abstracted and help build up surrogate models with less computational cost than the original models. Additionally, more robust diagnostic capability is brought out by the matrix form.

GLOBAL VALIDATION OF THE CLM5 MATRIX MODEL FOR C AND N SIMULATIONS

The temporal dynamics of C and N storage simulated by the matrix modules was compared with the dynamics simulated by the original versions of these modules. The CLM5 matrix model and the

original CLM5 use the same default initial size of vegetation and soil C and N storage without spin-up for the transition simulation. Modeled C and N storage in total ecosystem, vegetation, and soil organic matter from the matrix model matched with those from the original CLM5 (Figure 6.4). It is worth mentioning that the matrix module is

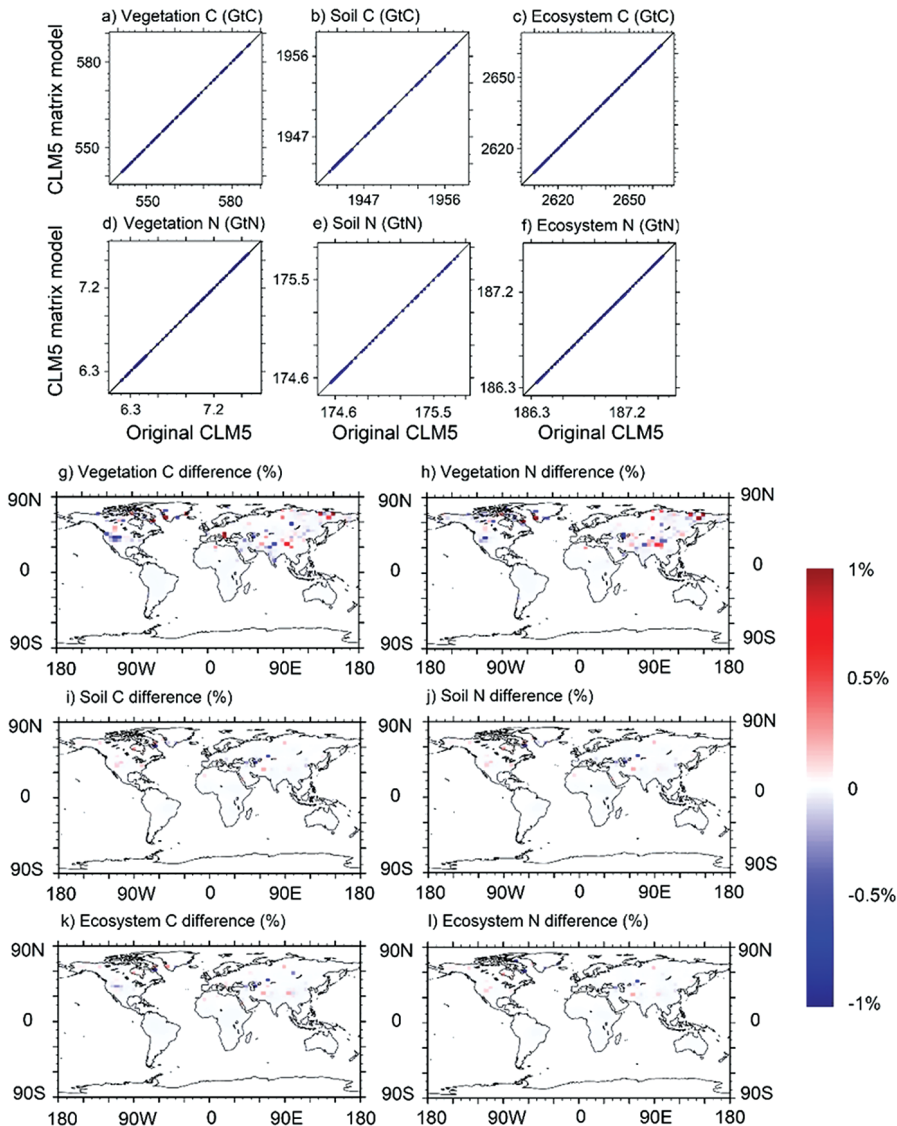


Figure 6.4. Comparison of simulated global carbon (C) and nitrogen (N) stocks from 1901 to 2014 between the CLM5 matrix model and the original CLM5. (a) Vegetation C storage, (b) soil C storage, (c) ecosystem C storage, (d) vegetation N storage, (e) soil N storage, and (f) ecosystem N storage are summed up over all land grid cells each year for comparison. Relative differences averaged over last four year (2011–2014) are calculated as: $\frac{X_{matrix} - X_{original}}{X_{original}} * 100$. X_{matrix} represents (g) vegetation C storage, (h) vegetation N storage, (i) soil C storage, (j) soil N storage, (k) ecosystem C storage, and (l) ecosystem N storage from the CLM5 matrix model. $X_{original}$ is the counterpart from the original CLM5.

not an exact representation of the original model because different processes in the original model are updated stepwise, while all processes in the matrix model are updated simultaneously. However, our validation results show that the relative differences of soil C and N are less than 1% for most grid cells. Only 0.4% and 0.5% of the grid cells in vegetation C and N storage, respectively, diverged by more than 1% relative difference (Figure 6.4g–h). Only two of 1466 global grid cells in ecosystem C and N storage diverged by more than 1%. Global annual means in all six C and N storages over 115 years perfectly lined up on the 1:1 line, signifying exact agreement (Figure 6.4a–f). The good agreements demonstrate that the matrix model can faithfully reproduce both temporal dynamics and spatial patterns of C and N states from the original model.

As mentioned before, minor divergence between the two model versions could be explained by the difference between the simultaneous nature of the calculations encapsulated by the matrix equations, versus the stepwise nature of the original model's algorithms. The matrix modules update C and N state variables only once within each time step, whereas the original modules update these state variables one after the other within each time step. This means that updates in a state variable calculated early in the sequence can affect the calculation for a subsequent state variable. For example, leaf turnover is updated by both phenology and fire in sequence in the original model. Thus, the turnover in leaf C from fire is proportional to the pool size after being updated by phenology in the original model, while the turnover in leaf C from fire is proportional to the pool size before phenology in the matrix model. As a consequence, simulated dynamics may slightly differ between the two model realizations. Nevertheless, the differences in modeled state variables due to different update methods of the two models were small enough not to generate notable differences as shown in global simulation of the C and N storage.

To summarize, we have shown in this chapter how matrix equations may be used to represent C–N coupled models for application at ecosystem and global scales. The new representation as a matrix equation for C–N coupled models has several advantages. The matrix form could help build up a surrogate model for parameter inversion and parameter sensitivity analysis, especially for understanding the role of different C–N

interactions for N limitation of CO₂ fertilization. Previously, parameter inversions or parameter sensitivity analysis of terrestrial biogeochemical models like CLM5 have been prohibitively expensive computationally. The surrogate model in matrix form greatly saves the computational cost (Huang et al., 2018b; Tao et al., 2020). Another advantage of the matrix form is that it enables a strong diagnostic approach, the traceability analysis, which attributes the differences in simulated C and N storage into several traceable components (see Chapter 17). These traceable components convey critical ecological or biogeochemical meanings which can help us understand the key drivers of the spatial and temporal variability in terrestrial C and N cycles emerging from model simulations. For example, water scalars have been identified as the most significant traceable component to explain wide divergence between estimates of permafrost carbon storages driven by two reanalysis meteorological datasets, GSWP3 and CRUNCEP (Lu et al., 2020). The success in capturing the dynamics of biogeochemical cycling of a complex model such as CLM5 indicates the feasibility of implementing matrix form in other terrestrial biogeochemistry models.

SUGGESTED READING

- Shi, Z., Yang, Y., Zhou, X., Weng, E., Finzi, A. C., & Luo, Y. (2016). Inverse analysis of coupled carbon–nitrogen cycles against multiple datasets at ambient and elevated CO₂. *Journal of Plant Ecology*, 9(3), 285–295.
- Xingjie Lu, Zhenggang Du, Yuanyuan Huang, David Lawrence, Erik Kluzek, Nathan Collier, Danica Lombardozzi, Negin Sobhani, Edward A. G. Schuur, Yiqi Luo. 2020. Full implementation of matrix approach to biogeochemistry module of Community Land Model version 5 (CLM5). *Journal of Advances in Modeling Earth Systems*. In press. <https://doi.org/10.1029/2020MS002105>

QUIZZES

1. What is the matrix form of ecosystem nitrogen cycling?
2. What are the differences between carbon and nitrogen cycles in the matrix form?
3. What are the key parameters to couple carbon and nitrogen cycles?
4. What drives the dynamic of the mineral nitrogen pool?

CHAPTER SEVEN

Compartmental Dynamical Systems and Carbon Cycle Models

Carlos A. Sierra

Max Planck Institute for Biogeochemistry, Jena, Germany

CONTENTS

Introduction / 57
Definition of Compartmental Systems / 58
Classification of Compartmental Systems / 59
Autonomous Versus Nonautonomous Systems / 59
Linear Versus Nonlinear Systems / 60
Properties and Long-Term Behavior of Autonomous Compartmental Systems / 60
Linear Systems / 60
Nonlinear Systems / 61
Stability Analysis Near Equilibria / 61
Linear Systems / 61
Nonlinear Systems / 62
Properties and Long-Term Behavior of Nonautonomous Systems / 62
Linear Systems / 62
Nonlinear Systems / 63
Final Remarks / 64
Suggested Reading / 64
Quizzes / 64

Models of the terrestrial carbon cycle are particular cases of compartmental dynamical systems, which are systems of differential equations that must conserve mass. This chapter introduces the main mathematical properties of compartmental dynamical systems and proposes a classification scheme that is useful for the analysis of carbon cycle models. This classification scheme distinguishes between models where carbon inputs and rates change over time or remain constant (nonautonomous versus autonomous models), and between models in which the amount of mass in compartments interact with mass in other compartments (nonlinearity). We show that simple concepts such as steady state may not be well defined for some groups of models, and present alternative concepts such as the pullback attractor for the analysis of models with no steady state. In addition, this chapter

introduces the theoretical basis for the mathematical analysis of models written in matrix form.

INTRODUCTION

The matrix representation of models has emerged as a very general representation of ecosystem models, particularly models that track the movement of carbon, nitrogen, and other elements inside vegetation and soil pools (Mulholland and Keener 1974; Matis et al. 1979; Bolin 1981; Luo and Weng 2011; Xia et al. 2013; Luo et al. 2017). For soil organic matter models, some of the first representations in matrix form were the models of Bolker et al. (1998), Baisden and Amundson (2003), and Tuomi et al. (2009). For these authors, the matrix representation helped them organize the set of differential equations that resulted in their model

in a more manageable and compact form. This is also the case in other fields of science such as biology or chemistry, where large sets of differential equations can be organized using this compact representation.

In fact, any model that represents the mass balance of a quantity such as atoms or molecules, can be represented in this form. Compared to other systems of differential equations, mass-balanced systems are special in the sense that all quantities are generally non-negative; i.e., the information that is fed into the model, and the predictions it produces can only exist inside the domain of the positive real numbers. Furthermore, the mass balance constraint leads to a special type of dynamical system known as a compartmental system.

We will now introduce the mathematical concept of compartmental systems, and will show that models written in compartmental form have a specific set of mathematical properties. These properties however, depend on the specific structure analyzed, mostly on the time dependence of the elements of the model and intrinsic nonlinearities.

DEFINITION OF COMPARTMENTAL SYSTEMS

We start by the defining a *compartment* as an amount of material that is kinetically homogeneous and that follows the law of mass balance. The meaning of ‘mass balance’ is elaborated below. A compartmental system therefore, is a set of compartments that exchange mass with each other and with the external environment. This implies that a compartmental system is an open system with an observer defined boundary (Anderson 1983; Jacquez and Simon 1993).

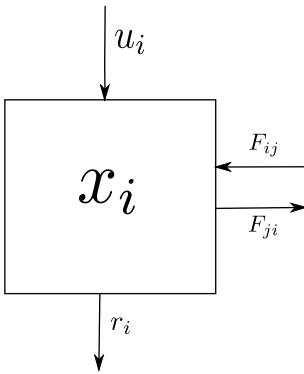


Figure 7.1. The mass balance of a single compartment.

Let’s consider the mass stored in the compartment i , denoted by x_i , as the balance between (Figure 7.1):

- $u_i \geq 0$ inflow (uptake) from outside the system,
- $r_i \geq 0$ outflow (release) to outside the system,
- $F_{ji} \geq 0$ flow transfers from compartment i to compartment j ,
- $F_{ij} \geq 0$ flow transfers from compartment j to compartment i .

The change in mass over time of this compartment, $\frac{dx_i}{dt} = \dot{x}_i$, must be balanced according to the equation:

$$\dot{x}_i = \sum_{j \neq i} (-F_{ji} + F_{ij}) + u_i - r_i,$$

where the constraints $F_{ij} \geq 0$, $u_i \geq 0$, and $r_i \geq 0$ must be met for all i, j , and t . The time dependence is omitted in the notation for simplicity, but all masses and flows may change over time.

An additional constraint for the system is that if the compartment is empty, no mass can flow out of it; i.e., if $x_i = 0$, then $r_i = 0$ and $F_{ji} = 0$ for all j , so that $\dot{x}_i \geq 0$.

If the flows F are continuously differentiable, i.e., they change smoothly over time without sudden jumps, we can define the flows as (Jacquez and Simon 1993):

$$F_{ji}(\mathbf{x}) \equiv b_{ji}(\mathbf{x}) \cdot x_i.$$

Therefore, we can write the mass balance equation for compartment i as:

$$\dot{x}_i = - \left(b_{0i} + \sum_{j \neq i} b_{ji} \right) x_i + \sum_{j \neq i} b_{ij} x_j + u_i.$$

The total outputs from compartment i can be expressed as $b_{ii} \equiv - \left(b_{0i} + \sum_{j \neq i} b_{ji} \right)$, then a general expression for each compartment satisfies the expression:

$$\dot{x}_i = \sum_j b_{ij} x_j + u_i.$$

A general expression for the entire system can be written as:

$$\dot{\mathbf{x}} = \mathbf{B}\mathbf{x} + \mathbf{u}, \quad (7.1)$$

where the elements may be time-dependent and the matrix \mathbf{B} and vector \mathbf{u} depend on the vector of states \mathbf{x} . Notice that in contrast to other chapters, we follow here a different notation and use \mathbf{B} to denote a matrix. The system of Equation 7.1 is a compartmental or reservoir system, and the matrix \mathbf{B} is called the compartmental matrix.

For any compartmental system, the compartmental matrix \mathbf{B} has three properties:

- $b_{ii} \leq 0$ for all i , $t \geq 0$,
- $b_{ij} \geq 0$ for all $i \neq j$, $t \geq 0$,
- $\sum_{i=1}^n b_{ij} = \sum_{i \neq j} b_{ij} + b_{jj} = -z_j \leq 0$ for all j , $t \geq 0$.

In words, the compartmental matrix \mathbf{B} must always meet the requirement that all its diagonal entries are non-positive, its off-diagonal entries non-negative, and the sum of all elements inside each column must be non-positive. This column sum represents the fraction of matter that is released from the system, and it is called the *fractional release coefficient* z_j because it can be used to compute the amount of material that is released to the external environment from each pool j . The total release from the system can be obtained with the expression:

$$\mathbf{r} = \mathbf{z} \circ \mathbf{x},$$

where \mathbf{z} is the vector of fractional release coefficients and \circ is the entry-wise product between the two vectors.

The property $-z_j \leq 0$ implies that \mathbf{B} is a diagonally dominant matrix, which means that each element in the diagonal is greater than or equal to the column sum for this entry. Mathematically, \mathbf{B}

is diagonally dominant if $|b_{ii}| \geq \sum_{j \neq i} |b_{ij}|$, and strictly diagonally dominant if $|b_{ii}| > \sum_{j \neq i} |b_{ij}|$.

One important property of strictly diagonally dominant matrices is that they are invertible (Taussky 1949); i.e., there exists an inverse matrix \mathbf{B}^{-1} such that $\mathbf{B} \cdot \mathbf{B}^{-1} = \mathbf{I}$, where \mathbf{I} is the identity matrix. Compartmental systems that meet this property contain no traps (Jacquez and Simon 1993); i.e., all mass that enters the system eventually leaves from any of the output flows.

CLASSIFICATION OF COMPARTMENTAL SYSTEMS

In the derivation of the compartmental system (Equation 7.1), the explicit representation of time dependencies and nonlinearities was omitted. We will now introduce a classification scheme for compartmental systems based on these two properties, time dependencies (autonomy), and interaction among state variables (linearity). We call a model linear when the vector of inputs and the compartmental matrix are not dependent on the vector of states, and nonlinear otherwise. Similarly, we call a model autonomous when the mass inputs and the compartmental matrix are not explicitly time dependent, and nonautonomous otherwise (Table 7.1).

This classification scheme leads to four distinct groups of compartmental systems, each with specific mathematical properties that we will explore in the following sections.

Autonomous Versus Nonautonomous Systems

In the autonomous case (Table 7.1), mass inputs and process rates in the system are constant. This implies that the external environment (e.g., solar radiation, air temperature, water content) are assumed constant. Although ecosystems are far from being

TABLE 7.1

Classification of carbon cycle models according to their dependence on the vector of states (linearity), and on time (autonomy). Table cells are expressions for the differential equation describing $\dot{\mathbf{x}}(t)$ that captures the change of mass contents with respect to time

\mathbf{x} -dependence	Autonomous	Nonautonomous
Linear	$\mathbf{u} + \mathbf{B} \cdot \mathbf{x}(t)$	$\mathbf{u}(t) + \mathbf{B}(t) \cdot \mathbf{x}(t)$
Nonlinear	$\mathbf{u}(\mathbf{x}) + \mathbf{B}(\mathbf{x}) \cdot \mathbf{x}(t)$	$\mathbf{u}(\mathbf{x}, t) + \mathbf{B}(\mathbf{x}, t) \cdot \mathbf{x}(t)$

surrounded by a constant environment, this assumption is sometimes useful to study basic properties of a system such as its long-term behavior.

However, it is important not to mix up concepts that belong to autonomous systems with concepts that do not apply for nonautonomous systems. For instance, an autonomous compartmental system generally converges to a steady state in the long term where the mass of each compartment does not change with time. In contrast, a nonautonomous system does not reach such a steady state because, by definition, the system is changing all the time. Therefore, it is wrong to talk about the steady state of a nonautonomous system (for additional details see Sierra et al. 2018).

Linear Versus Nonlinear Systems

In the linear case, the contents of compartments do not influence the rates at which mass flows into the system from the external environment, and do not influence the rates at which mass flows out of the compartments (Table 7.1). In other words, there are no feedbacks among compartment contents. However, nonlinear behavior can occur in ecosystems, for instance, when the amount of photosynthesis in the leaves depends on the amount of nonstructural carbohydrates or in fine roots.

Nonlinear compartmental systems can show a very rich set of qualitative behaviors (Jacquez and Simon 1993; Anderson and Roller 1991), which for nonlinear autonomous systems range from sustained oscillations to catastrophic shifts to alternate states (Wang et al. 2014). In the nonlinear nonautonomous case, the time-dependent signals that affect the system introduce an even larger degree of complexity, which complicates the behavior of these systems further (Müller and Sierra 2017).

PROPERTIES AND LONG-TERM BEHAVIOR OF AUTONOMOUS COMPARTMENTAL SYSTEMS

Even though the assumption of a constant environment is unrealistic, autonomous models can be very useful in illustrating potential behavior of compartmental systems. In the following, we will present a few properties of autonomous systems that are useful for many applications, which include: long-term behavior of stocks and fluxes, behavior in the neighborhood of the steady state after a perturbation, the age structure of the

compartments and the release flux, and the behavior of an impulsive tracer.

Linear Systems

We will consider first linear autonomous compartmental systems of the form

$$\dot{\mathbf{x}}(t) = \mathbf{u} + \mathbf{B} \cdot \mathbf{x}(t), \quad (7.2)$$

with \mathbf{B} invertible and some initial conditions at $t = 0$

$$\mathbf{x}(0) = \mathbf{x}_0.$$

One advantage of systems of the form of Equation 7.2 compared to the other systems in Table 7.1, is that it is possible to compute their analytical solution. The general solution of this model is given by:

$$\mathbf{x}(t) = e^{\mathbf{B}t} \mathbf{x}_0 + \left(\int_0^t e^{\mathbf{B}(t-\tau)} d\tau \right) \mathbf{u}, \quad (7.3)$$

where $e^{\mathbf{B}}$ is the matrix exponential.

Equation 7.3 shows that the solution of the system is composed of two terms. The first term accounts for the decomposition of the mass initially stored in the system at time zero. The second term accounts for decomposition of the inputs that entered the system until time t . At any given time, the mass stored in the system is the sum of both the remaining of the initial mass present at time $t = 0$ and all the un-decomposed mass that entered until time t .

The release of mass from the system is computed by multiplying the fractional release coefficients z_j by the amount of carbon stored in each pool as:

$$\begin{aligned} \mathbf{r}(t) &= (z_j \cdot x_j(t))_{j=1, \dots, n}, \\ &= \mathbf{z} \circ \mathbf{x}(t) \end{aligned}$$

If the system runs for a very long time, it eventually reaches a point called the *steady state* where all inputs are equal to the outputs, and there are no changes in mass within the system. Technically, as $t \rightarrow +\infty$, $\mathbf{x}(t) \rightarrow \mathbf{x}^*$, where:

$$\mathbf{x}^* = -\mathbf{B}^{-1} \cdot \mathbf{u}, \quad (7.4)$$

and

$$\begin{aligned} \mathbf{r}^* &= \left(z_j \cdot x_j^* \right)_{j=1, \dots, n}, \\ &= \mathbf{z} \circ \mathbf{x}^* \end{aligned}$$

Notice that the steady state does not depend on the initial conditions. It only depends on the compartmental matrix and the vector of external inputs, and represents the equilibrium point where the total amount of matter in the system and in the individual pools do not change, i.e., $\sum \dot{\mathbf{x}} = 0$, and $\dot{\mathbf{x}} = 0$, respectively.

Nonlinear Systems

In contrast to linear systems, nonlinear compartmental systems have no general explicit analytical solution. However, it is always possible to obtain a numerical solution of the system using any suitable numerical method (LeVeque 2007).

In most applications, we are interested in observing how the system evolves over time and eventually reaches a steady state. Therefore, it is of interest to find an equilibrium solution for the system:

$$\dot{\mathbf{x}} = \mathbf{u}(\mathbf{x}) + \mathbf{B}(\mathbf{x}) \cdot \mathbf{x}, \quad (7.5)$$

such that:

$$\mathbf{0} = \mathbf{u}(\mathbf{x}) + \mathbf{B}(\mathbf{x}) \cdot \mathbf{x}. \quad (7.6)$$

However, it is not certain that a specific nonlinear system has an equilibrium solution, or in case there is one, that this equilibrium is unique. Anderson and Roller (1991) show special cases of nonlinear compartmental systems with constant inputs that have unique solutions, but these cases are too specific for our purposes here.

Certain combinations of parameter values and pool sizes may lead to the situation in which the matrix $\mathbf{B}(\mathbf{x})$ is not compartmental, and therefore the system may not be mass balanced. For this reason, it is useful to define a space in which a nonlinear system is well defined. Following Anderson and Roller (1991), we define $\mathbb{R}_+^n := \{ \mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq \mathbf{0} \}$ as the set of all non-negative real numbers in an n -dimensional space.

Since the mass in all compartments is always non-negative, the solutions of the system can only occupy this space. Now we define the space within \mathbb{R}_+^n where all solutions of the system obey mass balance constraints as:

$$\Omega := \{ \mathbf{x} \in \mathbb{R}_+^n : \mathbf{B}(\mathbf{x}) \text{ is a compartmental matrix} \}.$$

The space Ω is the set of all possible states the system can take without violating mass balance. One important use of Ω is that it can be used to test whether a particular nonlinear model does not violate mass balance for any value of \mathbf{x} and t .

For the case of constant inputs, i.e., \mathbf{u} , Anderson and Roller (1991) propose an iteration strategy to find a steady-state solution for a nonlinear autonomous system. It consists of applying the formula:

$$\mathbf{x}^{p+1} = -\mathbf{B}(\mathbf{x}^p)^{-1} \cdot \mathbf{u}, \quad p = 0, 1, 2, \dots,$$

until $\mathbf{x}^{p+1} \approx \mathbf{x}^p$. Notice that for this method to work, the compartmental matrix must be invertible. Also, the existence of one equilibrium point is not a guarantee that it is unique: other equilibria may exist as well. The choice of the starting $\mathbf{x}^{p=0}$ may determine what equilibrium point the method will find.

Stability Analysis Near Equilibria

In many applications, it is of interest to study the behavior of a system as it approaches an equilibrium point, or the behavior of the system when it is slightly perturbed from this equilibrium. The study of these behaviors usually falls under the label *stability analysis*. Again, the stability analysis would differ depending on whether the autonomous system is linear or nonlinear.

Linear Systems

For linear autonomous compartmental systems (Equation 7.2), their long-term behavior can be studied by analyzing the eigenvalues and eigenvectors of the compartmental matrix \mathbf{B} . It is well established that a compartmental matrix with constant coefficients has no eigenvalues with positive real part, which means that the mass inside the compartments never grows exponentially as long as inputs and rates are kept constant. This is

ensured by the diagonally dominant property of the compartmental matrix.

In most applications, the eigenvalues of the linear autonomous compartmental matrix have a negative real part. In these cases, it is said that the compartmental system is asymptotically stable because all solutions converge in the long-term to the steady state of Equation 7.4. If the eigenvalues also contain a complex part, then the solution will approach the steady state through oscillations. If the eigenvalues contain no complex part, then the system approaches the steady state in the direction given by the eigenvector of the eigenvalue with the smallest absolute value of the real part.

A third possibility is that the compartmental matrix contains at least m eigenvalues with a real part equal to zero. In this case, it is said that the compartmental system contains m traps (Jacquez and Simon 1993). A trap is a compartment, or a set of connected compartments, where mass may flow in but cannot flow out. In this case, the system contains no equilibrium since \mathbf{B} is not invertible and Equation 7.4 cannot be solved. The system therefore, will grow proportionally to the amount of mass entering the m traps.

Nonlinear Systems

For nonlinear systems, it is common to study the behavior of the system in the neighborhood of one or multiple equilibrium points. For compartmental systems, we are only interested in equilibria that reside in the space Ω , since they are the only ones that have a physical and biological interpretation.

We assume that the nonlinear autonomous system of Equation 7.5 has at least one equilibrium point in Ω , then we are interested in calculating the Jacobian matrix, defined as:

$$\mathbf{J}(\mathbf{x}) = \frac{\partial(\mathbf{B}(\mathbf{x}) \cdot \mathbf{x})}{\partial \mathbf{x}},$$

at an equilibrium point $\mathbf{x} = \mathbf{x}^* \in \Omega$. This Jacobian matrix tells us about the behavior of trajectories that are close to the steady state, which is a point in the phase plane. Then, the properties of the Jacobian matrix, particularly its eigenvalues, tell us about the stability of the system in the neighborhood of the equilibrium (Guckenheimer and Holmes 1983). It is possible to treat the

nonlinear system as a linear system in the neighborhood of the equilibrium, and for this reason one can perform the same analysis of eigenvalues as in the linear case (Guckenheimer and Holmes 1983).

If there are eigenvalues with positive real part, trajectories are repelled away from the equilibrium point, which is considered *unstable* (Strogatz 1994). The existence of unstable equilibria is an indication of possible tipping points and alternative states for the system (Scheffer et al. 2001). However, it is often the case that the Jacobian matrix of a compartmental system is also a compartmental matrix, in which case the existence of unstable equilibria is excluded.

When this Jacobian matrix has a compartmental structure, the system is said to be *cooperative*, which means that if the mass of one compartment increases, the fluxes to other compartments also increase (Jacquez and Simon 1993). In this case, trajectories close to the equilibrium point are attracted to it, and in some particular cases this equilibrium may be unique (Jacquez and Simon 1993; Bastin and Guffens 2006). This particular case of a unique equilibrium point means that the system is *global asymptotically stable* or GAS (Müller and Sierra 2017).

PROPERTIES AND LONG-TERM BEHAVIOR OF NONAUTONOMOUS SYSTEMS

Nonautonomous compartmental systems behave in a completely different way to autonomous systems. Since the mass inputs and the rates change with time, it is not possible for them to converge to a fixed point in the state space. Also, the stability analysis tools for autonomous systems are of little use for nonautonomous systems. Methods to analyze nonautonomous systems are relatively new, and they are currently an active branch of mathematical research (Rasmussen 2007; Kloeden and Rasmussen 2011). Concepts from control engineering can also be very useful to study nonautonomous systems, particularly nonlinear ones (Sontag 1998). Again, we will split the concepts for linear versus nonlinear nonautonomous systems in the sections below.

Linear Systems

We will consider two cases for linear autonomous compartmental systems: (1) the case of

time-dependent inputs and constant rates, and (2) the case of time-dependent inputs and rates.

The first case is given by a system of the form:

$$\dot{\mathbf{x}}(t) = \mathbf{u}(t) + \mathbf{B} \cdot \mathbf{x}(t),$$

with initial condition $\mathbf{x}(0) = \mathbf{x}_0$. If the vector-valued function $\mathbf{u}(t)$ is known, an analytical solution can be obtained as:

$$\mathbf{x}(t) = e^{\mathbf{B}t} \mathbf{x}_0 + \int_0^t e^{\mathbf{B}(t-\tau)} \cdot \mathbf{u}(\tau) d\tau,$$

which is a general form for the linear autonomous solution of Equation 7.3. This analytical solution is only possible to compute because the rates in the compartmental matrix \mathbf{B} are constant for all times, and therefore one can take advantage of the analytical properties of the matrix exponential.

For the second case, when both mass inputs and rates are time dependent, the system is expressed as:

$$\dot{\mathbf{x}}(t) = \mathbf{u}(t) + \mathbf{B}(t) \cdot \mathbf{x}(t), \quad (7.7)$$

for which an analytical solution cannot be computed. However, a semi-explicit solution for Equation 7.7 can be expressed in terms of the state transition operator $\Phi(t, t_0)$, which is a matrix whose product with the state vector at an initial time t_0 gives $\mathbf{x}(t)$ at a later time t . In other words, $\Phi(t, t_0) \cdot \mathbf{x}_0$ is the solution to the homogeneous equation $\dot{\mathbf{x}} = \mathbf{B}(t) \cdot \mathbf{x}$.

The semi-explicit solution of the linear nonautonomous system of Equation 7.7 can be expressed as:

$$\mathbf{x}(t, t_0, \mathbf{x}_0) = \Phi(t, t_0) \cdot \mathbf{x}_0 + \int_{t_0}^t \Phi(t, \tau) \cdot \mathbf{u}(\tau) d\tau. \quad (7.8)$$

This solution explicitly depends on the initial conditions since for a nonautonomous system, where mass inputs and rates constantly change with time, the exact time and state when the system starts is of fundamental importance to compute a unique solution. In the autonomous case, solutions only depend on the time elapsed $t - t_0$, while in the nonautonomous case the solutions depend separately on the actual time t and the starting time t_0 (Kloeden and Rasmussen 2011).

Rasmussen et al. (2016) presents a sufficient condition for the global exponential stability of the nonautonomous linear compartmental system. If the compartmental matrix \mathbf{B} of the homogeneous system $\dot{\mathbf{x}} = \mathbf{B}(t) \cdot \mathbf{x}$ is strictly diagonally dominant for all t , then this system is exponentially stable. This means that there is a minimal rate at which the initial mass in the system decays. Now, for the inhomogeneous case (Equation 7.7), we can think of two solutions $s_1(t, t_1, \mathbf{x}_1)$ and $s_2(t, t_2, \mathbf{x}_2)$ that have different initial conditions. As a consequence of the exponential stability property, the two solutions are said to be *forward attracting*, i.e. they get close to each other as $t \rightarrow +\infty$.

Rasmussen et al. (2016) also showed that for linear nonautonomous compartmental systems that meet the sufficient condition for exponential stability, there exists a unique *pullback attracting solution* or *pullback attractor* which all solutions are attracted to. It is defined as:

$$\mathbf{v}(t) := \int_{-\infty}^t \Phi(t, \tau) \cdot \mathbf{u}(\tau) d\tau,$$

and can be interpreted as the solution that has no influence whatsoever from the initial conditions (Kloeden and Rasmussen 2011). Therefore, the pullback attractor is the nonautonomous equivalent of the steady-state concept for autonomous systems (Carvalho et al. 2013).

A particular case is the linear nonautonomous system in which the mass inputs and the process rates are periodic. For example, this is the case of seasonal systems without noise in which the same periodic pattern for the mass inputs and for the process rates is repeated every year. More precisely, a periodic linear compartmental system is one in which $\mathbf{u}(t + T) = \mathbf{u}(t)$ and $\mathbf{B}(t + T) = \mathbf{B}(t)$ for a fixed period T and for all t . Mulholland and Keener (1974) showed that these types of systems have periodic solutions for which $\mathbf{x}(t + T) = \mathbf{x}(t)$. This periodic solution can be interpreted as a pullback attractor because it has no influence on the initial conditions.

Nonlinear Systems

Nonlinear nonautonomous compartmental systems are the most complex cases for their study and analysis. It is not possible in general to obtain analytical solutions, and, contrary to the autonomous

case, it is not possible to study an equilibrium point for these systems because, by definition, compartment contents are always changing and they never reach a constant value.

As mass inputs, and process rates change in a nonlinear nonautonomous compartmental system, it is possible that specific combinations of parameter values and compartment sizes lead the system outside the space Ω where mass balance consideration must be met. Therefore, it is always important to check that solutions for these systems are always inside this space; i.e. $\mathbf{x}(t, t_0, \mathbf{x}_0) \in \Omega$ for all t , where $\mathbf{x}(t, t_0, \mathbf{x}_0)$ is a solution trajectory of the nonlinear nonautonomous compartmental system of the form:

$$\dot{\mathbf{x}}(t) = \mathbf{u}(\mathbf{x}(t), t) + \mathbf{B}(\mathbf{x}(t), t) \cdot \mathbf{x}(t). \quad (7.9)$$

Concepts from control theory could be used to ensure that solutions are well behaved and inside Ω , and more importantly, within certain ‘regions of stability’ that solutions are attracted to (Müller and Sierra 2017; Kloeden and Rasmussen 2011).

Input-to-state stability (ISS) is a concept from the field of control theory that can be used to determine whether a nonlinear nonautonomous compartmental system meets stability properties. We say that a dynamical system is ISS if it is globally asymptotically stable in the absence of time-dependent perturbations, and if its trajectories are bounded by a function of the size of the input for all sufficiently large times (Sontag 1998; Müller and Sierra 2017). Therefore, we can expect the trajectories of an ISS system to remain within a certain region as long as the initial mass decays over time, and the mass inputs stay bounded within a certain limit.

We expect that for most applications, nonlinear nonautonomous compartmental systems meet the properties of ISS systems. However, mathematically showing that a system is ISS is not trivial, and this

should be studied on a case-by-case basis (Sierra and Müller 2015; Müller and Sierra 2017).

FINAL REMARKS

The theory of compartmental dynamical systems offers a formal theoretical framework to express and analyze models of the carbon cycle and other biogeochemical elements that meet mass balance requirements. Using a matrix representation of carbon storage in ecosystem pools, it is possible to use the theory of compartmental dynamical systems to study important characteristics of models such as their long-term behavior, the presence of traps that retain carbon indefinitely in a model, and the response of ecosystem compartments to disturbances.

The representation of ecosystem models as compartmental systems is also useful to study system level properties of ecosystems (see Chapter 15). It is a useful mathematical representation that can relate ecosystem concepts to formal mathematical properties of dynamical systems.

SUGGESTED READING

General introductions to compartmental systems can be found in the monograph by Anderson (1983), and the comprehensive review of Jacquez and Simon (1993). More specific results about the application of compartmental systems to model the terrestrial carbon cycle can be found in the reference list and other chapters of this book.

QUIZZES

1. According to the general classification of models with respect to their dynamical properties, what type of compartmental systems have a fixed-point steady-state?
2. Can linear compartmental systems show transitions through tipping points?
3. What is the analogue of a steady state for nonautonomous systems? Why?

CHAPTER EIGHT

Practice 2

MATRIX REPRESENTATION OF CARBON BALANCE EQUATIONS AND CODING

Yuanyuan Huang

Climate Science Centre, CSIRO, Canberra, Australia

CONTENT

Introduction / 65

This practice helps you to learn how to write a matrix equation from carbon balance equations for the CENTURY model. You will also learn how to numerically solve the matrix equation through Python code using the CarboTrain package.

INTRODUCTION

In Chapter 5 we saw that a key step to develop the matrix approach to land carbon cycle modeling is to represent a land carbon cycle model in

a matrix form. This can be achieved by organizing the carbon balance equations of a model into one matrix equation. For many common models, the matrix version is mathematically identical to the original model without any changes in represented processes. Conversely, a matrix equation can be expanded row by row to give carbon balance equations for individual carbon pools. Exercise 1 focuses on how to derive the matrix version from the carbon balance equations of a carbon cycle model with multiple pools. Exercise 2 provides

EXERCISE 1: Deriving the matrix model from carbon balance equations

In this exercise we will develop a matrix form of the CENTURY model from the carbon balance equations. If you performed Exercise 1 in Chapter 4, you have written carbon balance equations of the CENTURY model, whose carbon flow diagram is shown in Figure 4.1.

The carbon cycle as depicted in the carbon flow diagram in Figure 4.1 can be represented by five carbon balance equations as below:

$$\frac{dx_1}{dt} = I * \beta_1 - k_1x_1$$

$$\frac{dx_2}{dt} = I * \beta_2 - k_2x_2$$

$$\begin{aligned} \frac{dx_3}{dt} = & f_{31} * k_1x_1 + f_{32} * k_2x_2 \\ & + f_{34} * k_4x_4 + f_{35} * k_5x_5 - k_3x_3 \end{aligned}$$

$$\frac{dx_4}{dt} = f_{41} * k_1x_1 + f_{43} * k_3x_3 - k_4x_4$$

$$\frac{dx_5}{dt} = f_{53} * k_3x_3 + f_{54} * k_4x_4 - k_5x_5$$

where x_1 , x_2 , x_3 , x_4 , and x_5 correspond to structural carbon, metabolic carbon, active soil carbon, slow soil carbon, and passive soil carbon pools. I is the input rate from plant residue. β_1 and β_2 indicate the proportion of this carbon input that is allocated to structural and metabolic litter. f_{ij} are fractions of carbon transferred from pool j to pool i . k_1 to k_5 are rates of soil organic carbon decomposition.

To develop a matrix model from the carbon balance equations, we need to know the following items. The first item is the general form of the matrix equation. If you do not remember it, please go back to Chapter 5 to find the

general form of the matrix equation. Second, you need to remember what the scalar, vectors, and matrices are in the general matrix equation.

For example, is $\frac{dX}{dt}$ a scalar, vector or matrix?

What does it mean? Please list the scalar, vectors and matrices in the general matrix equation. Third, you need to write all the scalar elements of these vectors and matrices. For example, $X = [x_1 \ x_2 \ x_3 \ x_4 \ x_5]^T$ indicates a vector of pools related to structural carbon, metabolic carbon, active soil carbon, slow soil carbon and passive soil carbon. What are the other vectors and matrices of the matrix model?

EXERCISE 2 Coding and running the matrix model

This exercise uses the package CarboTrain, which enables you to do exercises in this and other units with minimal background training in modeling and programming. Instructions for installing and working with CarboTrain are available in Appendix 3 of this book. The following steps will guide you through the coding and running of the matrix version of the TECO model using the CarboTrain package.

practice in coding and running the matrix model using Python through the CarboTrain package. Through coding, you will find some benefits of working with the matrix model, such as being easy to code and run model simulations.

Step 1: Run the default source code.

In the main window of CarboTrain, select Unit 2 and Exercise 2, specify the path of the output directory in which you wish to store your model result, click on Run Exercise. The notification “Task submitted!” will pop

up. Click on OK. Wait until the notification “Finished” appears, then click on OK. Go to the output directory you specified. There are two files. One is result.png, which plots the change of seven carbon pools through time. A second file is output.csv, which saves the value of each carbon pool in each simulation year.

Step 2: Understand the default source code.

In the CarboTrain main window, click on *Edit source code*. You will see the following source code (black text in your GUI). The code is based on the matrix model of TECO. The first part of the code loads packages and environment, and reads the output path you specified in the previous step.

The second part of the code, shown below, constructs the matrices and specifies our desired simulation length in years. The vector *iv_list* is the initial carbon pool size, *input_fluxes* specifies the input rate, *B* is the allocation vector, *A* is the transfer matrix, and *K* is the turnover rate matrix.

The third part of the code (Figure 8.3) calls the function *GeneralModel* to solve the carbon balance equations and generate the result.

```
# You specify values of Initial carbon pool size (iv_list), Input rate (input_fluxes),
# Allocation (B), Transfer (A), Turnover rate (K), and simulation length,
# GeneralModel solve the carbon balance equations and generate the result
mod = GeneralModel(times, B, A, K, iv_list, input_fluxes)

res = mod.get_x()
```

```

from GeneralModel import GeneralModel
import numpy as np
import matplotlib.pyplot as plt
import sys

if __name__ == '__main__':
    output_folder = sys.argv[1]

```

Figure 8.1. First part of the source code of the matrix version ofTECO.

```

B = np.array([0.45, 0.55, 0, 0, 0, 0, 0]).reshape([7,1]) # allocation

f31 = 0.72; f41 = 0.28; f42 = 1; f53 = 0.45; f54 = 0.275; f64 = 0.275;
f65 = 0.296; f75 = 0.004; f56 = 0.42; f76 = 0.03; f57 = 0.45;

A = np.array([-1, 0, 0, 0, 0, 0, 0,
              0, -1, 0, 0, 0, 0, 0,
              f31, 0, -1, 0, 0, 0, 0,
              f41, f42, 0, -1, 0, 0, 0,
              0, 0, f53, f54, -1, f56, f57,
              0, 0, 0, f64, f65, -1, 0,
              0, 0, 0, 0, f75, f76, -1]).reshape([7,7]) # tranfer

#turnover rate per day of pools: foliage, wood, metabolic litter, structural
#litter, soil microbial,slow soil, passive soil
temp = [0.00176, 0.000100104, 0.021468, 0.000845, 0.008534, 8.976e-005, 0.00000154782]

K = np.zeros(49).reshape([7, 7])

for i in range(0, 7):
    K[i][i] = temp[i]

#Unit of turnover rate from day^-1 to second^-1
#1 day = 86400 seconds
K = np.multiply(K, 1/86400)

# Cinput_const
input_fluxes = 0.00002245 #

nyear = 10000 # number of simulation years

times = np.linspace(0, nyear*365*86400, num = nyear)

iv_list = [0,0,0,0,0,0,0] # Initial carbon pool size

```

Figure 8.2. Second part of the source code ofTECO.

```

# Plot carbon pools and save results into csv file
fig = plt.figure(6*2, figsize=(14, 7.68))
plt.subplots_adjust(left = 0.1, right = 0.95, bottom = 0.10, top = 0.9, wspace =0.3, hspace =0.4)
x = list(range(1,nyear + 1, 1))

pool_names = ["foliage", "wood", "metabolic litter", "structural litter", "soil microbial", "slow soil", "passive soil"]

for i in range(1, 4):
    for j in range(1, 4):
        if ((i-1) * 3 + j) > 7 :
            break
        ax = plt.subplot(3, 3, (i-1) * 3 + j)
        ax.plot(x, res[(i-1) * 3 + j - 1,:])
        plt.xlabel("year", fontsize = 12)
        plt.ylabel(pool_names[(i-1) * 3 + (j-1)] + " pool ($g C m^{2}$)", fontsize = 12)
plt.savefig(output_folder + "/result" + ".png", dpi = 500)
#plt.show()

print(res[:,nyear-1]) # print result of the last year

mod.write_output(output_folder + ".output.csv")

```

Figure 8.3. Call to the functions that solve the carbon balance equations and retrieve the result.

The fourth part of the code (Figure 8.4) generates the output figure (result.png) and saves results into the csv text file (output.csv).

Step 3: Modifying the default source code.

To get a sense of how the code works and what controls the system dynamics, you could change some of the default values in the second part of the code. For example, if

```

from GeneralModel import GeneralModel
import numpy as np
import matplotlib.pyplot as plt
import sys
if __name__ == "__main__":

    output_folder = sys.argv[1]

    B = np.array([0.4, 0.35, 0.25, 0, 0, 0, 0]).reshape([7,1]) # allocation

    f41 = 0.7; f51 = 0.25; f42 = 0.65; f52 = 0.25; f43 = 0.15; f53 = 0.75;
    f64 = 0.3; f65 = 0.05; f75 = 0.04;

    A = np.array([-1, 0, 0, 0, 0, 0, 0,
                  0, -1, 0, 0, 0, 0, 0,
                  0, 0, -1, 0, 0, 0, 0,
                  f41, f42, f43, -1, 0, 0, 0,
                  f51, f52, f53, 0, -1, 0, 0,
                  0, 0, 0, f64, f65, -1, 0,
                  0, 0, 0, 0, f75, 0, -1]).reshape([7,7]) # transfer

    #turnover rate per day of pools:
    #Leaf,root,wood, metabolic litter, structural litter, fast SOM, passive SOM
    temp = [0.0017, 0.002, 0.0001, 0.01, 0.001, 0.0001, 0.000001]

    K = np.zeros(49).reshape([7, 7])

    for i in range(0, 7):
        K[i][i] = temp[i]

    #Unit of turnover rate from day^-1 to second^-1
    #1 day = 86400 seconds
    K = np.multiply(K, 1/86400)

    # Cinput_const, assume to be constant
    input_fluxes = 0.00002245 #

    nyear = 10000

    times = np.linspace(0, nyear*365*86400, num = nyear)

    iv_list = [0,0,0,0,0,0,0]

```

Figure 8.4. Code segment that generates visual output and saves results to a text file.

we change the number of simulation years (nyear) to 100, how does this affect the output from the run? If we make the passive soil carbon turnover faster by changing the default value of 0.00000154782 per day to 10^{-5} (1e-5 in Python) per day, what difference do you see in comparison with the default? You could explore around and understand how different parts of the matrices control the system dynamics. Every time you modify the code through *Edit source code*, please make sure you save the code before you click on *Run Exercise*. It is good practice

to change one place at a time and change the value back to the default value after you finish the practice.

Step 4: Building a new carbon model (optional).

Suppose we have a system with leaf (pool 1), root (pool 2), wood (pool 3), metabolic litter (pool 4), structural litter (pool 5), fast soil organic matter (pool 6), and passive soil organic matter (pool 7). Carbon dynamics of the system are given by the matrix equation below. Based on the default TECO model source code, above, are you able to

code this matrix model and check on carbon dynamics through time?

$$\frac{dX}{dt} = BI + AKX$$

$$B = \begin{bmatrix} \beta_1 \\ \beta_2 \\ 1 - \beta_1 - \beta_2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$A = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ f_{41} & f_{42} & f_{43} & -1 & 0 & 0 & 0 \\ f_{51} & f_{52} & f_{53} & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & f_{64} & f_{65} & -1 & 0 \\ 0 & 0 & 0 & 0 & f_{75} & 0 & -1 \end{bmatrix}$$

$$K = \begin{bmatrix} k_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & k_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & k_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & k_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & k_5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & k_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & k_7 \end{bmatrix}$$

We assume 40% of the net photosynthetic carbon is allocated to leaf, 35% to root and the remaining to wood. $f_{41}=0.7$; $f_{51} = 0.25$; $f_{42} = 0.65$; $f_{52} = 0.25$; $f_{43} = 0.15$; $f_{53} = 0.75$; $f_{64} = 0.3$; $f_{65} = 0.05$; $f_{75} = 0.04$; k_1 to k_7 are $[0.0017, 0.002, 0.0001, 0.01, 0.001, 0.0001, 0.000001]$ per day. For other parameters, if not specified above (e.g., the input rate), we assume they take the same values as the default TECO model. Thus, you do not need to change these values.

Does your new matrix model work? Can you run it? What results do you get?

Click on [Open solutions](#) in CarboTrain to check if the source code of your new matrix model is similar to the code shown in Figure 8.5.

```

mod = GeneralModel(times, B, A, K, iv_list, input_fluxes)

res = mod.get_x()

fig = plt.figure(6*2, figsize=(14, 7.68))
plt.subplots_adjust(left = 0.1, right = 0.95, bottom = 0.10, top = 0.9, wspace =0.2, hspace =0)
x = list(range(1, nyear+1, 1))

pool_names = ["foliage", "root", "wood", "metabolic litter", "structural litter", "fast soil", "passive soil"]

for i in range(1, 4):
    for j in range(1, 4):
        if ((i-1) * 3 + j) > 7 :
            break
        ax = plt.subplot(3, 3, (i-1) * 3 + j)
        ax.plot(x, res[(i-1) * 3 + j - 1,:])
        plt.xlabel("year", fontsize = 12)
        plt.ylabel(pool_names[(i-1) * 3 + (j-1)] + " pool ($g C m^{-2}$)", fontsize = 12)
plt.savefig(output_folder + "/result" + ".png", dpi = 500)
#plt.show()

print(res[:, nyear-1])

#mod.write_output("./output.csv")

```

Figure 8.5. Solution to Step 4.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

UNIT THREE

Carbon Cycle Diagnostics for Uncertainty Analysis



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER NINE

Unified Diagnostic System for Uncertainty Analysis

Yiqi Luo

Cornell University, Ithaca, USA

CONTENTS

Uncertainty in Land Carbon Cycle Modeling / 73
One Formula to Represent Land Carbon Cycle Models / 74
A Three-dimensional (3D) Space to Describe Model Outputs / 75
Five Traceable Components for Traceability Analysis / 78
Suggested Reading / 78
Quizzes / 78

All model intercomparison projects (MIPs) have shown large uncertainties in prediction of carbon sequestration among models and poor model-data matches. Although great efforts have been made, it is still difficult to identify causes of model uncertainty. This chapter offers a unified diagnostic system, which is also called a 1-3-5 scheme of diagnostics, for uncertainty analysis in carbon cycle modeling. The number 1 stands for one formula to unify the land carbon models, the number 3 is for one three-dimensional (3D) space to evaluate all model outputs, and the number 5 is for five traceable components to pinpoint uncertainty sources via traceability analysis.

UNCERTAINTY IN LAND CARBON CYCLE MODELING

Hundreds of land models have been developed to predict the future state of ecosystems in an attempt to inform management practices for climate change mitigation. However, all model intercomparison projects (MIPs) have shown large uncertainties in prediction of carbon sequestration among models and poor model-data matches (Friedlingstein et al. 2006, 2014, Luo et al. 2015). For example, eleven earth system models (ESMs) participating in the Coupled Model Intercomparison Project Phase 5

(CMIP5) perform poorly in predicting the distribution of global land surface soil carbon (Todd-Brown et al. 2013). Similarly, the spread among 11 ESMs participating in the subsequent CMIP6 exercise has not changed significantly from CMIP5 results (Arora et al. 2020). The model predictions are an order of magnitude more uncertain over land than over ocean. Regionally, terrestrial ecosystem models did not accurately characterize a wide range of vegetation functional traits associated with net primary productivity (NPP) in the East Asian monsoon area (Cui et al. 2019).

Great efforts have been made in the past decades to identify causes of model uncertainty. For example, model development teams add more and more processes into land carbon models in the hope of representing ecosystem processes more realistically. This practice yields mixed results. Incorporation of the nitrogen cycle into more models has been suggested to reduce the spread of CMIP6, whereas different treatments of processes in permafrost regions results in divergent model prediction. In general, increasing details in process representation in models hinders our understanding of holistic system behavior (Sierra et al. 2018). Benchmark analysis has been used to evaluate model performance skills, quantify model-data mismatches, and identify processes that need to be improved (See Chapter 19).

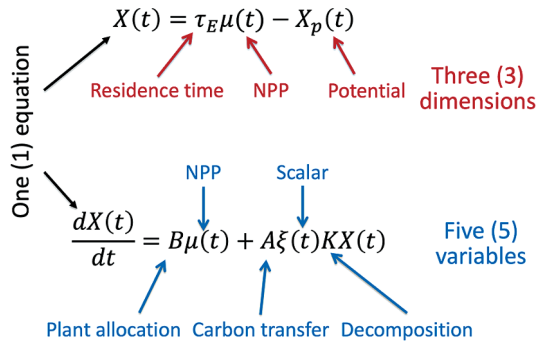


Figure 9.1. The unified diagnostic system with one (1) equation, three dimensions (3D), and five (5) variables.

Data assimilation has been proposed to improve data-model consistency (See Chapter 37). None of the approaches, however, help full understanding of the causes of model uncertainty. Without identifying sources of uncertainty, it is extremely difficult to focus model improvement efforts to realistically predict future carbon dynamics.

This chapter introduces a 1-3-5 scheme as a unified diagnostic system for uncertainty analysis (Figure 9.1). This diagnostic system is based on the matrix approach to representation of land carbon cycle models in one (1) general equation without compromising any details of process representation. Land carbon cycle dynamics are defined by a three-dimensional (3D) space with axes of coordinates being carbon input, residence time, and carbon storage potential. Model uncertainty can be traced to five (5) variables, which are carbon input, plant partitioning, decomposition, carbon transfer, and environmental scalars. Thus, the 1-3-5 scheme provides an analytic framework to understand the structure of complex models, their dynamic behavior, and uncertainty.

ONE FORMULA TO REPRESENT LAND CARBON CYCLE MODELS

The unified diagnostic system offered by the matrix approach is based on one formula to unify land carbon cycle models. In the previous two training units, we have shown you that the matrix equation can unify the land carbon cycle models.

We have converted 18 models to the matrix equations. These models are CLM3.5, CLM4, CLM4.5, CLM5, CABLE, LPJ-GUESS, IBIS, CASA', CENTURY, ORCHIDEE, TEM, TECO, DELAC2, ELM, GECO, FBDC, BEPS, and YASSO. In these models, we reorganize the original carbon balance equations into one matrix equation without changing

any processes. For example, CLM5 uses hundreds of carbon balance equations to simulate carbon transfer among 18 plant pools of 17 vegetation types and 140 soil pools over 20 soil layers. These carbon balance equations are organized into one matrix equation for the 18 vegetation pools and one matrix equation for the 140 soil pools (Lu et al. 2020). The vegetation and soil matrix equations are connected through litterfall and can be integrated into one matrix equation.

Sierra and Müller (2015) review all soil carbon models according to the principles underlying models. The matrix equation we have focused on in this training course satisfies five principles: mass balance, substrate dependence, heterogeneity of decomposition rates, transformation of organic matter, and environmental effects. When decomposition rates or transfer rates are functions of substrate, we still can use a matrix equation to describe carbon dynamics. In this case, the matrix equation becomes a nonlinear model. Indeed, Carlos Sierra's group from Max Planck Institute of Biogeochemistry, Germany, uses the nonlinear matrix models to represent more than ten microbial models. Thus, we can understand models at different levels of complexity under one overarching theory. Please be mindful that nonlinear matrix models will have different properties from the linear ones.

With one general formula to represent all the models, we can seek to understand the general behavior of the land carbon cycle and diagnose model uncertainty on a common ground. For example, the widely used model CLM5 includes 18 plant pools for each of the 17 vegetation types and 140 soil pools (i.e., 7 pools per layer over 20 layers) (Lu et al. 2020). Thus, CLM5 simulates carbon transfer among 158 pools (i.e., 140 for soil + 18 for plant) in one grid-cell if it is occupied by one vegetation type and 194 (i.e., 140 + 3 × 18) pools if

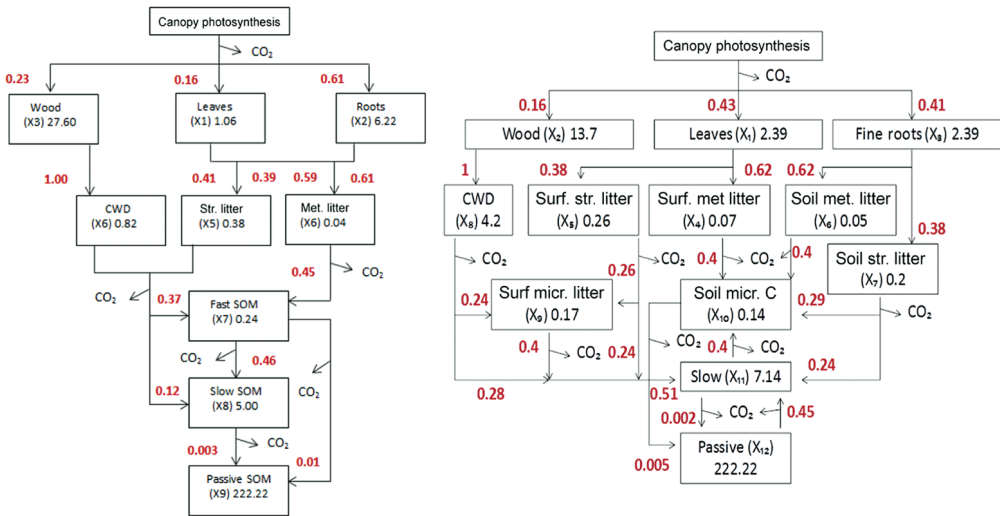


Figure 9.2. Model structure and parameterization of CABLE (a) and CLM3.5 (b). Carbon enters the system through photosynthesis and is partitioned among live plant pools. From live pools, carbon is transferred to litter pools, and then to soil carbon pools. Values in boxes show the pool residence times. Values outside the boxes show the partitioning and transfer coefficients. Abbreviation CWD for coarse woody debris, Str. for structural litter, Met. for metabolic litter, Surf. for surface litter, micr. for microbial biomass, and SOM for soil organic matter (From Rafique et al. 2016).

one grid-cell is occupied by three vegetation types. As a consequence, the matrix equation has at least 158 elements in the pool vector. In contrast, the CABLE model has nine pools so that the pool vector has nine elements (Xia et al. 2012). Despite the different numbers of carbon pools treated by these two models, they share the fundamental structure represented in the matrix equation, a structure that is shared by all land carbon cycle models.

Besides the differences in pool numbers, each of the five components of the matrix equation (i.e., equation 1.6 in Chapter 1) is represented differently either by fixed values, functions, or nested models. The differences in parameterization of the five components also contribute to different simulation results. For example, CABLE allocates 61% of NPP to roots, 23% to wood and 16% to leaves (Figure 9.2a) whereas CLM3.5 allocates 43% of NPP to leaves, 16% to wood and 41% to roots (Figure 9.2b) (Rafique et al. 2016). Similarly, a large difference exists in carbon transfers from live plants to litter and soil. CABLE transfers dead tissues into three litter pools (including coarse woody debris, CWD) after senescence, whereas CLM3.5 transfers dead plant tissues to six litter pools (including CWD) after mortality. CLM3.5 and CABLE also differ in representing decomposition (i.e., K matrix in Figure 9.1). While the two models can be both represented by one matrix formula, CLM3.5 realizes each of

five components differently from CABLE, leading to different model projections of carbon dynamics. Despite these differences in model structure and parameterization, all these models can be represented by the same matrix equation.

A THREE-DIMENSIONAL (3D) SPACE TO DESCRIBE MODEL OUTPUTS

Now let us explore what the 3D space means for evaluating model outputs. Chapter 1 explains that we need three variables: carbon input, residence time, and carbon storage potential to describe the transient dynamics of the land carbon cycle. The three variables become three dimensions to form a 3D space that can place model outputs, no matter how differently models are executed, for evaluation on a common ground. The mathematical foundation for using the 3D space to evaluate the transient dynamics of the carbon cycle is presented in detail in a paper by Luo et al. (2017). Here is a description of each of the three dimensions.

The first dimension of transient dynamics is carbon input through primary production. The primary production, being either gross primary production (GPP) or net primary production (NPP), quantifies the amount of carbon that enters an ecosystem to go through a variety of processes of the land carbon cycle before being released back

to the atmosphere. Of the three variables, carbon input via GPP or NPP has been studied most.

The second dimension of transient dynamics is carbon residence time. Residence time is approximated as a mean value (i.e., mean residence time, MRT) or turnover time by dividing pool by flux. The approximation is valid when the carbon cycle is at equilibrium but yields substantial deviation from residence times for a multiple compartmental system when the carbon cycle is not at equilibrium (Lu et al. 2018b). Residence time or transit time is explained in detail in Chapter 15 and can be estimated from data assimilation for individual ecosystems. Overall, the residence time quantifies the duration of carbon staying in an ecosystem before being released back to the atmosphere.

The two terms, NPP and residence time, together quantify equilibrium carbon storage capacity. Figure 9.3 shows simulated carbon storage capacity by three land models, the Beijing Climate Center (BCC) model, the Canada (CAN) model, and the Community Earth System Model Biogeochemical module (CESM GBC). In response to climate change, all the three models simulate increases in NPP but decreases in residence time. The CESM BGC simulates the smallest NPP and residence time, leading to the lowest carbon storage capacity. The BCC model simulates the largest increase in NPP, leading the large increase in carbon storage capacity. The Canada model simulates the highest residence time and thus estimates the highest carbon storage capacity.

However, when models are used to simulate responses of ecosystems to climate changes, the carbon cycle is no longer at equilibrium but in dynamic disequilibrium. Therefore, we need the third term, the carbon storage potential, to describe disequilibrium of the land carbon cycle. The CESM GBC model has the smallest carbon storage potential whereas the CAN and BCC models have high storage potential (Figure 9.3).

We use one more example to explain the three dimensions (3D) of land carbon cycle dynamics, particularly the third dimension. The third dimension, carbon storage potential, is a relatively new concept and worth more explanation.

A free-air CO₂-enrichment (FACE) experiment was conducted in Duke Forest in North Carolina, USA. The FACE experiment started in the mid-1990s and ended in late 2000s. The CO₂ concentration in the treatment rings was elevated by 200 parts per million above the ambient CO₂ concentration.

To illustrate the concept of 3D space of carbon dynamics in the FACE experiment, let us make a few assumptions. Let us assume that NPP is 1000 g C m⁻² yr⁻¹ and residence time is 40 years at ambient CO₂ concentration. Then, we have the steady-state carbon pool size in the Duke Forest as 40 kg C m⁻² before the FACE treatment. We assume that elevated CO₂ treatments increase NPP by 40% but have no effect on residence time. The steady-state pool size is 56 kg C m⁻² at elevated CO₂ treatments. For the sake of simplicity, we also assume no seasonal and diurnal changes in forcing variables and carbon processes.

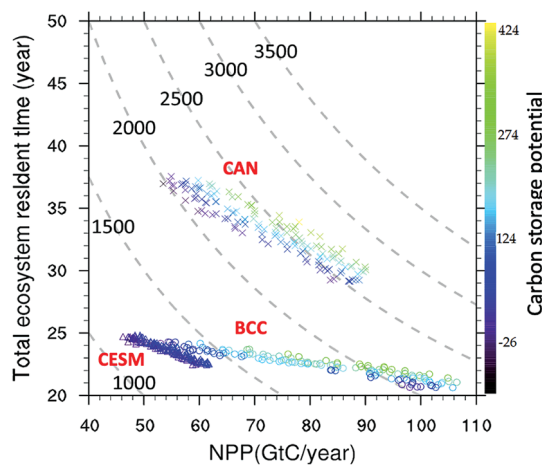


Figure 9.3. The 3D model output space (NPP in x-axis, carbon residence time in y-axis, and carbon storage potential in color), for three models from CMIP5. The points represent the global annual values of carbon storage for the three variables. The contours represent the carbon storage capacity. Colors show the values of the carbon storage potential.

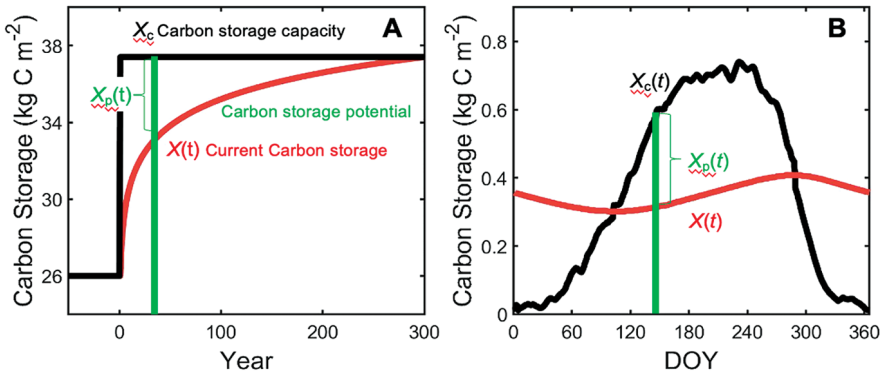


Figure 9.4. Carbon storage dynamics as determined by carbon storage capacity and potential. Panel A presents an ideal case in which carbon storage capacity (X_c) is a constant while carbon storage potential ($X_p(t)$) and carbon storage itself ($X(t)$) vary with time. In this case, the capacity is assumed to abruptly increase by 40%, mainly due to instantaneous increase in carbon input as in an elevated CO_2 experiment (Luo & Reynolds, 1999). Consequently, the potential immediately increases and then gradually declines as $X(t)$ increases toward the equilibrium (i.e., the carbon storage capacity at elevated CO_2 treatment). Panel B illustrates time-dependent $X(t)$, its capacity and potential in a non-autonomous system over day of year (DOY). Seasonal change in the capacity is due to change in carbon input, which is low in winter and high in summer. The capacity is a moving target that $X(t)$ chases. The rate of chasing is proportional to the storage potential.

In this example, we have two equilibrium states of carbon storage (i.e., carbon storage capacity), one at ambient $[\text{CO}_2]$, equaling 40 kg C m^{-2} and the other at the elevated $[\text{CO}_2]$, denoted as X_E , equaling 56 kg C m^{-2} (Figure 9.4A). At the very beginning of the FACE experiment, we need to determine in which direction and how fast the current carbon storage $X(t)$ would change. Generally speaking, we expect that the carbon storage changes toward the equilibrium state at elevated $[\text{CO}_2]$ (X_E) but the rate of the change is fast in the first few years and slow in later years. Let us denote the difference between the equilibrium carbon storage and current storage to be the carbon storage potential $X_p(t)$. The rate of carbon storage change is proportional to the potential $X_p(t)$.

The above example illustrates three key findings about carbon storage dynamics. First, carbon storage is always moving toward the carbon storage capacity. Second, the capacity is the ultimate attractor which current carbon storage chases (or changes toward). Third, the rate of the carbon storage change is proportional to the carbon storage potential.

The Duke FACE example assumes equilibrium carbon storage capacity. However, the carbon storage capacity, which equals NPP times residence time, is changing over time in the real world as both NPP and residence time vary with time. This is the reason why the carbon cycle in the terrestrial ecosystem is mathematically considered a non-autonomous system.

We use simulated seasonal change of fine root biomass in Harvard Forest (Luo et al. 2017) as an example to illustrate the non-autonomous system of carbon cycle dynamics. The carbon storage capacity of the fine root pool is a theoretical quantity, which is NPP times residence time. As NPP is low in winter and high in summer, the root carbon storage capacity as indicated by the black line is low in winter and high in summer (see Figure 9.4B). The red line indicates the current carbon storage in fine roots, which is higher than the black line in the winter and lower than the black line in the summer. The red line is always moving toward the black line. That is a mathematical explanation why fine root amount declines in fall and winter but increases in spring and summer.

For a non-autonomous system in which both NPP and residence time are changing with time, the carbon storage capacity still controls the direction of carbon storage change. Likewise, the carbon storage potential still determines the rate of carbon storage change. Therefore, carbon input, residence time, and carbon storage potential form a 3D space within which all model outputs can be evaluated.

Zhou et al. (2018) have applied the 3D space to evaluate model performance from three model intercomparison projects (MIPs): CMIP5, TRENDY, and MsTMIP. The 3D space is defined by NPP as x-axis, residence time as y-axis, and color for carbon storage potential to place outputs from 25 models in the same space to evaluate their

performance (see Chapter 18 for details). The three variables can also be plotted to indicate current carbon storage in relatively smooth lines, the carbon storage potential in shaded areas, and the carbon storage capacity in the zig-zag lines over a time course. This is the first time we are able to evaluate all model outputs from different MIPs in a simple, 3D space. See Chapter 18 for more applications of the 3D space to the model evaluation.

Enqing Hou has recently led a study to show the uncertainty among eight matrix models can expand or shrink, depending on differences in residence time when carbon input into the eight models is the same (Hou et al. in prep.). As all the terms to calculate residence time, which are A , $\xi(t)$, K , and B , in equation 1.13 in Chapter 1 are standardized, model prediction trajectories can be brought to be one identical one. This study indicates the model uncertainty is all related to carbon input and residence time.

FIVE TRACEABLE COMPONENTS FOR TRACEABILITY ANALYSIS

Now, let us examine the five traceable components of carbon cycle dynamics, which have been effectively used in traceability analysis.

The matrix equation describes land carbon dynamics via carbon input (e.g., NPP), allocation of carbon input to different plant parts, carbon process rates via litterfall or decomposition of organic matter, carbon transfer among pools, and environmental scalar. They consist of five components of the matrix equation. Mathematically, the five components are largely independent from each other in models (Xia et al. 2012) although they may interact in the real world. Moreover, each component can be further traced to its subcomponents as far as individual carbon cycle processes and their parameter values. This mathematical property enables us to trace sources of model uncertainty to individual processes and parameters via traceability analysis (see Chapter 17).

This traceability analysis was first developed by Xia et al. (2013) for the equilibrium carbon storage capacity and expanded by Jiang et al. (2017) and Zhou et al. (2018) to the transient dynamics of the land carbon cycle. The traceability analysis can trace the model uncertainty in simulated carbon storage hierarchically along the traceable pathways down to vegetation traits, climate forcing, and soil attributes. Chapter 17 shows an

authentic traceability analysis that requires matrix models to generate all the traceable components so that it enables uncertainty analysis to be analytically transparent. Chapter 18 shows post-MIP traceability analysis that does not require matrix models but is applied to any model output. The post-MIP traceability analysis can attribute uncertainty among models to variations in NPP, residence time, and carbon storage potential. The variations in NPP, residence time, and carbon storage potential can be further traced to environmental scalars and model parameters.

For example, the Australian CABLE model predicts lower carbon storage capacity than CLM3.5 due to lower NPP. But CABLE has higher residence time than CLM3.5 (Rafique et al. 2016). The higher residence time in CABLE is mainly caused by setting lower decomposition coefficients, leading to higher baseline residence time. The environmental scalars among the two models are similar. The example shows that the traceability framework can help trace sources of model uncertainty down to individual processes or parameter values.

The training in unit 3 shows you how we can add diagnostic variables in matrix models so that we can use the 1-3-5 scheme for uncertainty analysis and traceability analysis. Unit 5 will specifically show you how to do traceability analysis.

SUGGESTED READING

Luo YQ, Shi Z, Lu X et al. (2017) Transient dynamics of terrestrial carbon storage: mathematical foundation and its applications. *Biogeosciences*, 14, 145–161.

QUIZZES

1. What is the 1-3-5 scheme of the diagnostics system?
2. Please briefly describe traceability analysis.
3. What does the term “carbon storage potential” mean?
4. A non-autonomous system is
 - a. a system which does not conserve mass balance.
 - b. a system with its properties changing with time.
 - c. a system with constant pool sizes.
 - d. a system with its pool sizes changing with time.

CHAPTER TEN

Sensitivity Analysis with Matrix Equations

A CASE STUDY WITH ORCHIDEE

Yuanyuan Huang

Climate Science Centre, CSIRO, Canberra, Australia

CONTENTS

What is Sensitivity Analysis? / 79
Sobol Sensitivity Analysis / 80
One-at-a-time Sensitivity Analysis / 82
Spatial Pattern / 85
Suggested Reading / 85
Quizzes / 85

Sensitivity analysis is an important procedure to understand how model parameter values, input or forcing data, and sometimes model structure, influence behaviors of the modeled system. However, sensitivity analysis is rarely conducted for complex land carbon models due to high computational cost. The matrix approach makes it computationally feasible to conduct sensitivity analyses of land carbon models.

WHAT IS SENSITIVITY ANALYSIS?

Sensitivity analysis is the study of how the variability in the output of a mathematical model or system (numerical or otherwise) can be divided and allocated to different sources of variability in its inputs, parameters, and structures (Wikipedia). Sensitivity analysis allows us to identify the parameter or set of parameters that have the greatest influence on the model output, and helps us to understand model dynamics, trace uncertainty sources and calibrate model parameters.

As a common practice, we sometimes change the value of one model parameter, conduct model simulations with and without the parameter

change, and compare the results. Quantifying the change in the result per unit change of the parameter can give us a sense of how sensitive the model is to that parameter. By repeating the analysis for other parameters of the model, we may discover which parameters are most important in determining the predictions of the model. This is the one-at-a-time approach.

Let's say we have several data points of a variable X and corresponding values of a second variable Y . Suppose we believe that Y and X are related in such a way that a change in X induces a response in Y . A preliminary approach we could take is to make a scatter plot or a regression analysis of Y onto X . If there is a steep relationship between Y and X , that means Y is very sensitive to X : a small change in X induces a large change in Y . These examples are sufficient for linear relationships or simple cases in which the sensitivity of Y to change in X does not vary due to interactions with other parameters. In dealing with nonlinear responses or nonadditive cases, where interactions among several parameters affect the sensitivity of a model's output to its input, variance-based sensitivity methods offer an alternative that can be

more informative. The variance-based approach decomposes and attributes the uncertainty of model output to inputs. For example, for a model with two inputs and one output, if 30% of the variance of the output is attributable to parameter P1, 45% is caused by parameter P2 and 25% by the interaction between P1 and P2, we could interpret these percentages as measures of the model sensitivity (Wikipedia).

SOBOL SENSITIVITY ANALYSIS

In this chapter, we focus mainly on one variance-based method, the Sobol method, as an approach to partitioning the sensitivity of the output of a model to its parameters, inputs or structures. Here we focus on parameters. Sobol sensitivity analysis is a global sensitivity method. Here, “global” refers to the whole parameter space, in comparison to some methods that can only give us information around a limited region of the parameter space, i.e., the local sensitivity. The Sobol method relies on Monte Carlo sampling and can separate the single and interaction effects among parameters. Land carbon models typically consist of many equations, parameters, and state variables interacting with each other. Nonlinear responses are common in these models. A method that measures the sensitivity across the whole parameter space and quantifies the interactions is very attractive. The shortcoming of this method is its relatively high computational cost. Process-based land carbon models often track diverse processes with response scales ranging from minutes (e.g., half-hourly time step for photosynthetic carbon uptake) to centuries or millennia (e.g., turnover of recalcitrant soil organic matter pools). Global or regional simulations using these models are normally computationally expensive and rely on the use of supercomputers. A large number of simulations required by the Sobol method and the high computational cost of complex land carbon models, especially over large spatial-temporal scales, make direct applications of the Sobol method to complex land carbon models very challenging. The matrix version of the model, thanks to its efficient semi-analytical spin-up, brings us the capability in fulfilling such a task.

Here I take the ORCHIDEE-MICT model as an example. As introduced in Chapters 3 and 5, the litter and soil carbon component of the model simulates carbon dynamics across 100 pools to capture multiple real-world processes

that govern terrestrial carbon cycling. Vegetation carbon enters litter pools during processes such as leaf senescence, root and wood turnover, and fire disturbance. ORCHIDEE-MICT tracks four classes of litter, namely aboveground metabolic litter, belowground metabolic litter, aboveground structural litter, and belowground structural litter. Aboveground and belowground litter differ in the environmental conditions (temperature and moisture) that modify the rate of litter decomposition. Litter carbon (both aboveground and belowground) is transferred into vertically resolved soil carbon pools as litter decomposes. ORCHIDEE-MICT divides soil carbon into 32 layers, down to a depth of 38 m. The thickness of soil layer increases from shallow to deep soil layers. In each soil layer, ORCHIDEE-MICT tracks three different types of soil carbon pools, that is, the active soil organic carbon (SOC) with a default potential turnover time of 0.145 year and the slow and passive SOC pools with potential turnover times of 5.48 and 241 years, respectively. Once it enters the soil, carbon is transferred among a complex network of soil carbon pools, distinguished by vertical layer and turnover time. Ultimately, soil carbon can enter the atmosphere through respiration, be transformed into more recalcitrant pools with longer turnover times, or be buried into deep soil layers through cryoturbation or bioturbation. Advection is not considered in this model version. The original model is rewritten into the matrix form introduced in Chapter 5.

Each element in these matrices might be linked to several processes or functions as mathematical expressions of mechanisms controlling land carbon dynamics. To evaluate which sets of parameters are most important in determining carbon dynamics in the modeled system, we perform sensitivity analysis on the matrix model.

Based on our understanding of carbon cycling and the original model, we choose 34 parameters that could potentially be important (see Table 10.1). These parameters regulate elements from different matrices. For example, the lignin content affects the transfer of carbon flux from one pool to another. It also modifies the turnover rate. We take advantage of the matrix version of the model for semi-analytical spin-up (see Chapter 14), which saves us a lot of computational time. We randomly sample parameters across the whole parameter space and conduct 2,720,000 ($=34 \times 100 \times 100 \times 8$) model simulations (see more details below). As you can

TABLE 10.1
Parameters affecting litter and SOC dynamics in the ORCHIDEE-MICT model

	Name	Matrix source	Description	Default value	Range	Unit
1	ins	I	Input scalar	1	[0,1]	
2	p4lf	I	Partition (structural vs. metabolic) coefficient for leaf	0.6916	[0,1]	
3	p4sa	I	Aboveground sapwood partition, structural vs. metabolic	0.598	[0,1]	
4	p4sb	I	Belowground sapwood partition, structural vs. metabolic	0.598	[0,1]	
5	p4ha	I	Aboveground heartwood partition, structural vs. metabolic	0.598	[0,1]	
6	p4hb	I	Belowground heartwood partition, structural vs. metabolic	0.598	[0,1]	
7	p4ro	I	Root partition, structural vs. metabolic	0.6916	[0,1]	
8	p4fr	I	Fruit partition, structural vs. metabolic	0.6916	[0,1]	
9	p4ca	I	Carbohydrate reserve partition, structural vs. metabolic	0.6916	[0,1]	
10	fam2a	A	Transfer fraction, aboveground metabolic litter to active SOC	0.45	[0,1]	
11	fbm2a	A	Transfer fraction, belowground metabolic litter to active SOC	0.55	[0,1]	
12	fas2a	A	Transfer fraction, aboveground structural litter to active SOC	0.45	[0,1]	
13	fbs2a	A	Transfer fraction, belowground structural litter to active SOC	0.45	[0,1]	
14	fas2s	A	Transfer fraction, aboveground structural litter to slow SOC	0.7	[0,1]	
15	fbs2s	A	Transfer fraction, belowground structural litter to slow SOC	0.7	[0,1]	
16	fa2p	A	Transfer fraction, active to passive SOC	0.004	[0,0.15]	
17	fs2a	A	Transfer fraction, slow to active SOC	0.42	[0,0.5]	
18	fs2p	A	Transfer fraction, slow to passive SOC	0.03	[0,0.5]	
19	fp2a	A	Transfer fraction, passive to active SOC	0.45	[0,1]	
20	zlit	A	Factor control vertical distribution of litter input to SOC	PFT dependent	[0.2, 1.25]	
21	clay	A, ξ_{Cl}	Clay content	0.2	[0,0.6]	
22	lgc	A, ξ_L	Lignin coefficient on structural litter decomposition	3	[0,10]	
23	lga	A, ξ_L	Aboveground lignin content	0.76	[0,1]	
24	lgb	ξ_L	Belowground lignin content	0.72	[0,1]	
25	temps	ξ_T	Temperature sensitivity	0.69	[0,1]	
26	ms	ξ_W	Moisture scalar	1	[0.8,1.2]	
27	tau4ml	K	Turnover time, metabolic litter	0.066	[0,0.066]	year
28	tau4sl	K	Turnover time, structural litter	0.245	[0,0.245]	year
29	tau4a	K	Turnover time, active SOC	0.149	[0,0.149]	year
30	tau4s	K	Turnover time, slow SOC	5.48	[0,5.48]	year

(Continued)

TABLE 10.1 (CONTINUED)

	Name	Matrix source	Description	Default value	Range	Unit
31	tau4p	K	Turnover time, passive SOC	241	[0,241]	year
32	cryo	V	Cryoturbation rate	0.001	[0,1]	m ² /year
33	bio	V	Bioturbation rate	0.0001	[0,1]	m ² /year
34	alt	V	Maximum active layer depth of the last year	0.2	[0,3]	m

see, this is a large number (nearly three million) of simulations. It is not realistic to conduct so many simulations with the original model due to its high computational requirements, but with the matrix model it is feasible. Based on the simulation results, we calculate the single as well as the total effect of the targeted parameters on the output of the model, using Sobol’s index.

To be specific, we randomly sample parameters within ranges provided in Table 10.1 assuming a uniform distribution for each parameter. In choosing a uniform distribution, we assume each value within the range is equally possible for the parameter. Parameter ranges are chosen based on model information, parameter meaning (e.g., the transfer fraction cannot be bigger than 1), and empirical knowledge. With these randomly chosen parameters, we conduct $34 \times 100 \times 100$ simulations. 34 is the number of parameters. 100 is the number of random values sampled from the potential distribution of parameter i . For each random value of parameter i , we randomly sample 100 sets of parameters excluding parameter i . The first-order Sobol sensitivity index S_i for parameter i (p_i) is given by:

$$S_i = \frac{V_{p_i} (E_{p_{-i}} (Y | p_i))}{V(Y)} \quad (10.1)$$

where $V(Y)$ denotes the total variance; $Y | p_i$ indicates simulation outputs given p_i ; p_{-i} represents the parameter space excluding the i th parameter; and E corresponds to the expectation from multiple simulations. Similarly, the total-order Sobol index is given by:

$$ST_i = \frac{E_{p_{-i}} (V_{p_i} (Y | p_{-i}))}{V(Y)} \quad (10.2)$$

Due to the random sampling, the Sobol index may be dependent on the specific samples

especially when sample size is not big enough, the so-called convergence issue. To avoid the convergence issue, we repeat the previous step eight times ($34 \times 100 \times 100 \times 8$ simulations in total) and calculate our final Sobol index as the average from the eight replicates.

The results of the sensitivity analysis are shown in Figure 10.1, the left two panels show the parameters ranked by sensitivity from high to low based on the effect of the parameter on total soil organic carbon. The first panel is the total effect (single + interaction), quantified by the total-order Sobol index, ST_i (Equation 10.2). The second panel shows the single effect, quantified by the first-order Sobol index, S_i (Equation 10.1). As you can see, the model is most sensitive to the parameter that controls the external carbon input into this system (*ins*), followed by the turnover time of the passive soil organic carbon pool. The right three panels show the sensitivity of active SOC, slow SOC and passive SOC to different parameters at different soil depths. Different carbon pools are sensitive to different parameters at different soil depths. The point here is, with the matrix version of the carbon model, we are able to look into different aspects of the model sensitivity in a great deal of detail.

ONE-AT-A-TIME SENSITIVITY ANALYSIS

To investigate more closely the sensitivity of the model to key parameters, revealed by the Sobol analysis, we can perform one-at-a-time sensitivity analysis. For example, Figure 10.1 tells us that the parameter that regulates external carbon input (*ins*) and that scaling the turnover of passive SOC (*tau4p*) are important in terms of the sensitivity of total SOC, while the active layer depth (*alt*) seems to be important in determining the relative amount of carbon stored in active, slow and passive SOC pools. In Figure 10.2, we change these parameters in even steps denoted by different colors and look

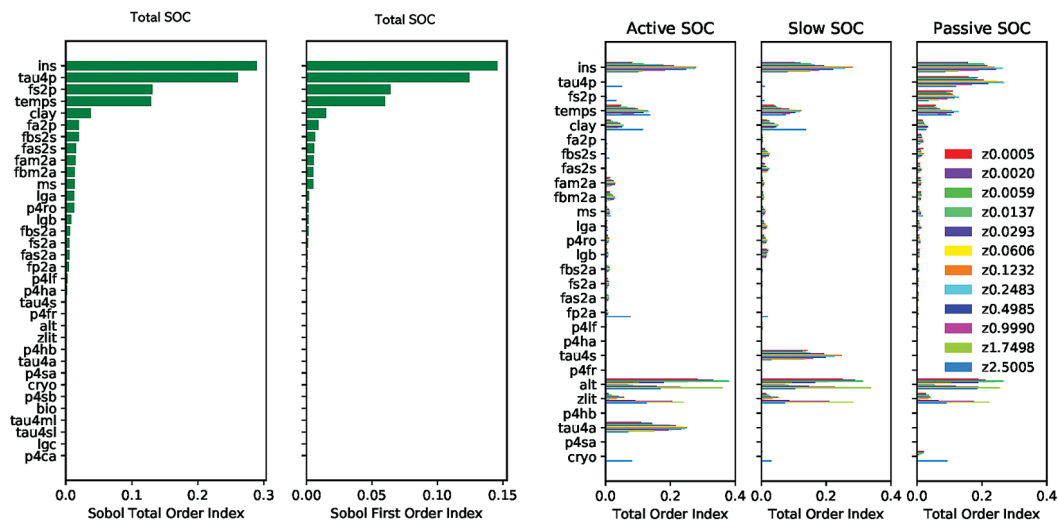


Figure 10.1. Sensitivities of soil organic carbon to 34 relevant parameters through the Sobol's method. The left two panels are for total soil organic carbon. The right three panels are for different soil organic carbon pools in different soil layers. Colors indicate different soil layers. Adapted from Huang et al. (2018a).

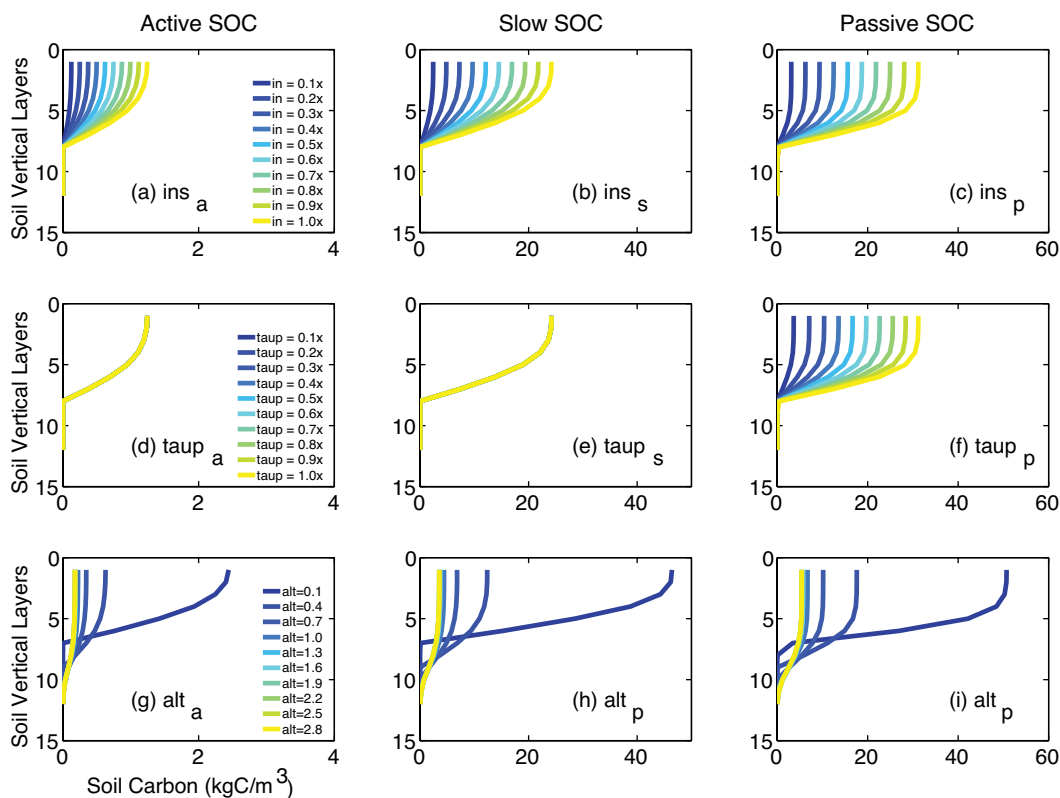


Figure 10.2. Sensitivities of different SOC pools to litter input scalar (top three panels, (a)-(c), ins), passive SOC turnover time scalar (middle three panels, (d)-(f), $tau4p$) and the maximum active layer depth of the last year (bottom three panels, (g)-(i), alt) through changing each parameter one-at-a-time. The x-axis corresponds to soil carbon content and the y-axis corresponds to soil vertical layers (larger numbers mean deeper soil layers). Colors denote different parameter values. Adapted from Huang et al. (2018a).

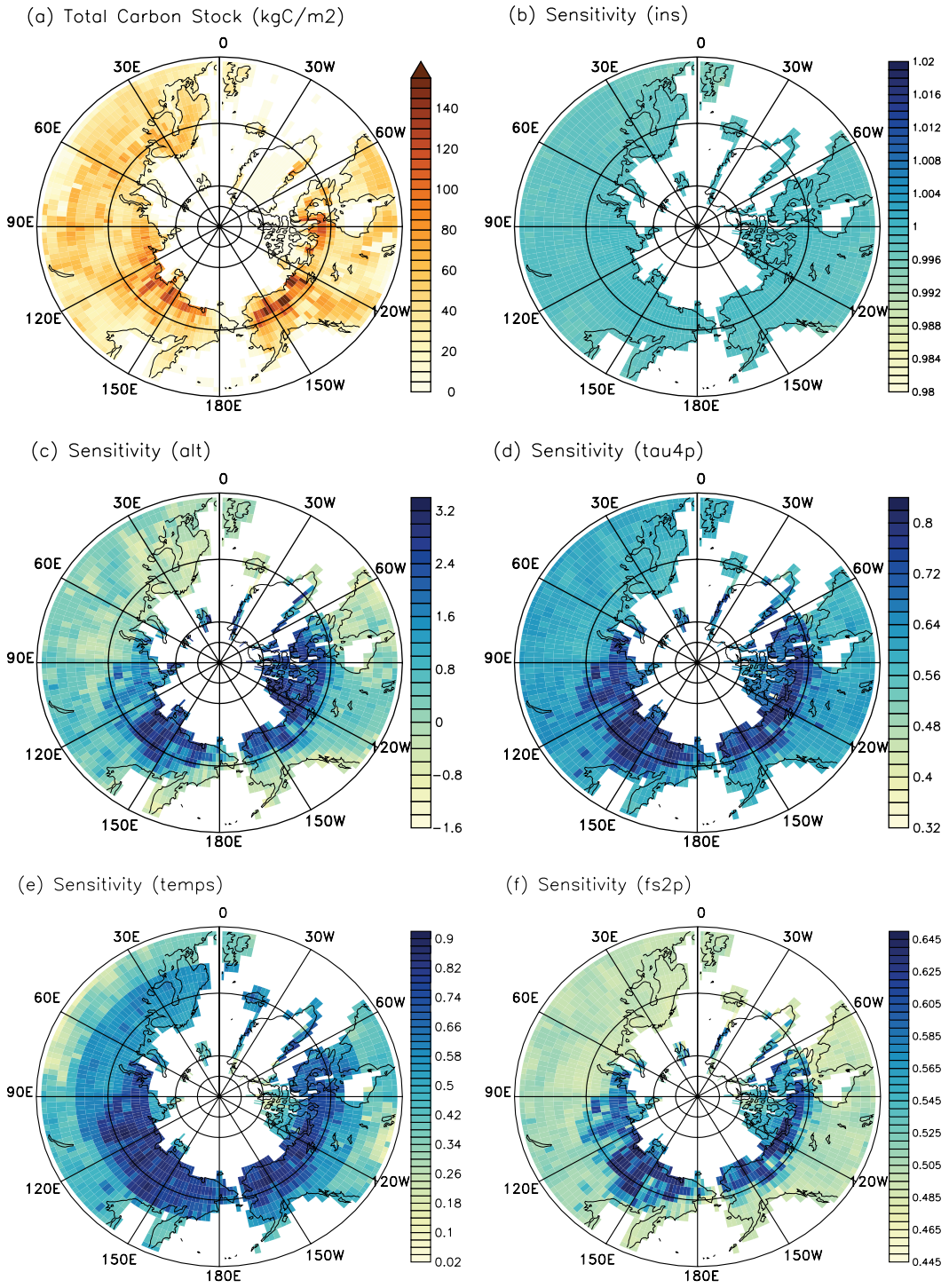


Figure 10.3. Sensitivity of high latitude ($> 50^{\circ}\text{N}$) total SOC stocks to 20% change in carbon input (*ins*), maximum active layer depth (*alt*) and turnover time of passive SOC (*tau4p*). Sensitivity is quantified as the ratio between the change in total SOC and the corresponding parameter. Panel (a) shows total SOC stock. Panels (b), (c), (d), (e), (f) show the sensitivity of total SOC to the indicated parameter. Adapted from Huang et al. (2018a).

into changes of active SOC (left panels), slow SOC (middle panels) and passive SOC (right panels). For the carbon input parameter, the sensitivity is relatively linear in surface soil layers indicated by the even space between different colored lines, while the sensitivity to active soil layer depths is generally nonlinear.

SPATIAL PATTERN

The spatial or geographic pattern of a model's sensitivity can often be insightful. This is because the importance of different parameters can vary depending on the environmental drivers, which vary along geographic gradients, and in different biomes. In Figure 10.3, the top left panel shows total carbon stock simulated by the matrix model over the northern high latitudes. The top right panel is the sensitivity to the external input parameter (*ins*), the middle-left panel is the sensitivity to active layer depth (*alt*), the middle right panel is the sensitivity to the turnover time of passive SOC pool (*tau4p*), the bottom left panel is the sensitivity to temperature (*temp*), and the bottom right shows the sensitivity to the fraction of slow soil carbon transferred to passive carbon pools (*f2p*). The sensitivity to *ins* is spatially homogenous as the response of

the model to this parameter is linear, while sensitivities to *alt* and *tau4p* vary with location.

In summary, the matrix version of the carbon model brings flexibility in comprehensive sensitivity analyses. Hopefully the case study of this chapter will provide you with inspiration to make use of the matrix version of carbon models for your own studies.

SUGGESTED READING

Huang, Y.Y., Zhu D., et al. 2018. Matrix-Based Sensitivity Assessment of Soil Organic Carbon Storage: A Case Study from the ORCHIDEE-MICT Model. *Journal of Advances in Modeling Earth Systems* **10**, 1790–1808, doi:10.1029/2017ms001237.

QUIZZES

1. What is a sensitivity analysis?
2. Why does it take nearly 3 million simulations to understand parameter sensitivities in this study?
3. Why is sensitivity analysis computationally costly?
4. What are the key benefits of the matrix model form for sensitivity analysis?



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER ELEVEN

Matrix Phosphorus Model and Data Assimilation

Enqing Hou

Northern Arizona University, Flagstaff, USA

CONTENTS

Introduction / 87
A Brief Overview of Soil P Dynamics Models / 88
Matrix Approach to Soil P Modeling and Data Assimilation / 88
An Example of Applying a Matrix Model and Data Assimilation to Soil P / 89
Data Selection and Description / 89
Construction of the P Matrix Model / 89
Model Validation and Data Assimilation / 91
New Knowledge Emerging from Data Assimilation with the Matrix Model / 91
Soil P Dynamics Quantified by Data Assimilation / 91
Soil P Dynamics in Relation to Other Ecosystem Properties / 93
Summary / 93
Suggested Reading / 94
Quizzes / 94

Soil phosphorus supply regulates terrestrial carbon dynamics. To improve our understanding of this regulation, we need to improve our understanding of soil phosphorus dynamics first. This chapter first briefly reviews research progress in modeling soil phosphorus dynamics, with a focus on the diversity of representations among models. This chapter then demonstrates how to construct a soil phosphorus model and how to transfer it to a matrix form. Finally, this chapter presents an example study to show how assimilating data into the matrix model can improve our understanding of soil phosphorus dynamics and availability. Overall, this chapter demonstrates that the matrix approach and data assimilation are very useful techniques to study terrestrial nutrient dynamics.

INTRODUCTION

Phosphorus (P) is a key element of macromolecules such as deoxyribonucleic acid (DNA), ribonucleic

acid (RNA), and adenosine triphosphate (ATP), of which the former two are carriers of genetic information and the last is the carrier of energy during biochemical reactions. Given the important roles of P in life, plants and other organisms need P for growth and reproduction. In natural ecosystems, plants mainly obtain P through uptake from the soil. Since soil P supply is not always sufficient to meet plant P demand, P limits plant production and its response to elevated atmospheric carbon dioxide (CO₂) in many terrestrial ecosystems. Moreover, P can regulate ecosystem carbon storage and cycling by affecting soil microbial activity. Therefore, improving our understanding of terrestrial P dynamics and P–C interactions is crucial to realistically simulate the land C cycle as well as its responses to future global change.

The application of data assimilation to quantify terrestrial C dynamics is well established (e.g., Luo et al. 2016). However, data assimilation to inform a soil P model has rarely been tried (Hou et al. 2019), despite suitable soil P measurements being

available in the literature (Hou et al. 2018), The data assimilation approach is potentially complementary to other currently available techniques (e.g., isotope dilution technique) in quantifying soil P dynamics. For example, it can use multiple sources of existing observations (e.g., soil P pool size and plant P uptake), simultaneously quantify the rates of all major soil P processes, and provide information about the uncertainties of the parameters related to soil P. Moreover, it may be particularly useful to quantify the dynamics of slow-cycling soil P pools (e.g., soil occluded P), which can hardly be achieved by any currently available experimental technique.

This chapter will first briefly review research progress in modeling soil P dynamics, with a focus on the diversity of representations among models. The chapter will then propose how a matrix framework and data assimilation can potentially improve our modeling of soil P dynamics. After that, an example study is presented to show how assimilating data into a matrix model can help to improve our understanding of soil P dynamics and availability.

A BRIEF OVERVIEW OF SOIL P DYNAMICS MODELS

Soil P dynamics is a key component of terrestrial P dynamics. To account for soil P supply as a determining factor for plant growth, land models are increasingly being extended to incorporate soil P dynamics. The first well-known land model to include soil P dynamics was CENTURY, published in the 1980s (Parton et al. 1988). In the last decade, because of growing interest in P cycle regulation of the land C cycle, over ten land models have been extended by incorporating soil P dynamics.

Current soil P models share some common features. Most are constructed based on the well-known soil P pools including soil labile P (readily available to plants), organic P (in organic forms), secondary mineral P (associated with secondary minerals such as iron and aluminum oxides), primary mineral P (apatite P), and occluded P (associated with soil clay and minerals and not directly available to plants). The dynamics of P as it accumulates and flows among these pools are usually expressed by a set of equations, based on empirical understanding of soil P dynamics. Despite a generally good understanding of soil P processes, most soil P models are not well parameterized, calibrated, or validated, because of limited

long-term observations of soil P dynamics. In current modeling practice, soil P models are typically not fully spun up to a steady state before formal simulations, even though spin-up to steady state is normally regarded as essential for land modeling. High computational cost associated with the initialization of slow-cycling soil P pools such as soil occluded P, as well as the unidirectional change in soil primary mineral P with time, make a full spin-up impractical for most current models.

Current P models differ both in structure and parameters. Regarding structure, some models include both water soluble P and labile P pools. The former can be directly available to plants and the latter may not be directly available to plants but exchangeable with the water soluble P pool. However, other models do not include a soil water soluble P pool and assume that soil labile P is directly available to plants. Both empirical studies and theory have suggested that soil occluded P pool may deplete in some conditions (e.g., an anaerobic environment) and accumulate in some other conditions (e.g., an aerobic environment). However, most land models that incorporate soil P dynamics assume that soil occluded P accumulates all the time. Moreover, model structure with respect to soil organic P differs largely among models, because soil organic P dynamics are coupled with soil organic C dynamics via soil C:P ratios in models, while models have very different structures for soil organic C dynamics (e.g., different numbers of soil organic C pools and vertically resolved vs. unresolved schemes).

Parameter values scaling soil P dynamics also differ among models. The literature provides many observations of soil P pools but few observations of soil P fluxes, especially fluxes from slow-cycling soil P pools (e.g., secondary mineral P and occluded P). Given these limitations, most soil P dynamics models are poorly parameterized and thus suffer from large uncertainties when they are used to predict future changes. Moreover, regulation by environmental factors (e.g., soil temperature and moisture content) of soil P dynamics and enzyme-mediated soil P mineralization are not well represented in current soil P models. In summary, the modeling of soil P dynamics is still in its infancy.

MATRIX APPROACH TO SOIL P MODELING AND DATA ASSIMILATION

The matrix approach and data assimilation can potentially help soil P modeling in several ways.

First, a matrix representation of the soil P system makes it easier to depict, understand, and compare models than the traditional approach using many difference equations. In the standard approach, change per time step in each soil P pool is represented by one equation with multiple state variables and parameters. Different soil P pools are mathematically linked by cross-terms in multiple equations, but such linkages may not be apparent to experimentalists who may have the data to validate the model but may not have any modeling experience. A matrix representation of soil P dynamics transfers the equations for different soil P pools into a unified form, no matter how many equations or what type of model, as long as a model is structured to describe transfers among multiple soil P pools. The matrix representation of soil P dynamics facilitates our understanding of model structure (e.g., the number of pools and the linkages among pools) and enables a direct comparison of structure among models.

A matrix representation of soil P dynamics also has the advantage of enabling a fast model spin-up and data assimilation that may be difficult or impossible with a traditional P model due to the prohibitive amount of computational power required. Soil P pools besides primary mineral P are generally in equilibrium in natural terrestrial ecosystems. Therefore, equilibrium soil P pool sizes are usually needed before a model simulation or data assimilation procedure. The turnover of soil occluded P in the field is typically slow (hundreds to tens of thousands of years), potentially even slower than the turnover of soil passive organic C (hundreds to thousands of years) which constitutes a bottleneck in spin-up of carbon-only ecosystem models. Therefore, it is a computational challenge to spin up a soil P model in the traditional way (e.g., repeat forcing many times). If soil P dynamics are represented in a unified matrix form, the semi-analytical spin-up (SASU) method introduced in Chapter 14 can be used to perform model spin-up, with the pool size of soil primary mineral P set to be constant. Application of SASU to soil P modeling may accelerate model spin-up by one or more orders of magnitude and may make it computationally feasible to assimilate soil P measurements from the field into models.

Data assimilation enables a quantitative and predictive understanding of soil P dynamics and availability. The understanding is fundamental to an accurate management of soil P availability,

which can further meet the societal needs of efficient use of P fertilizer in croplands, prohibiting eutrophication of water bodies, and developing strategies to alleviate P constraints on terrestrial C sequestration. To this end, an example is given below to show how matrix model and data assimilation approaches can be used in soil P studies to gain insights into soil P dynamics and availability.

AN EXAMPLE OF APPLYING A MATRIX MODEL AND DATA ASSIMILATION TO SOIL P

Data Selection and Description

The example assimilates datasets reported in Guo et al. (2000) into a soil P dynamics model. Guo et al. (2000) reported consistent (i.e., using the same fractionation procedure) and repeated (seven or eight times) measurements of P fractions of eight soils that represent four of the twelve major USDA soil types. The datasets are selected, because most soil P fractions changed substantially during the study period, which potentially offers a constraint on modeled soil P dynamics. Additionally, the soil samples are representative of common soil types, promoting the scope of applicability of the model. A less desirable feature of these datasets is that they are derived from experiments performed in the relatively artificial conditions of a greenhouse. One likely consequence is that the P pools in these experiments might have higher turnover rates than in the field.

In a nutshell, Guo et al. (2000) used crops to remove labile P from the eight soils to trigger changes in P fractions in these soils over a total of 14 cropping periods. After every two croppings, they sampled small amounts of the soils to determine soil P fractions using a modified Hedley P fractionation procedure. They fractionated soil P into several soil P fractions, corresponding to P pools with different turnover times and plant availability. Further details of the experimental design, cropping, sampling and preparation of the soils, and determination of the P fractions and physicochemical properties of the soils, are available in Guo et al. (2000).

Construction of the P Matrix Model

Our example follows the study of Hou et al. 2019. Here we construct a soil P dynamics model that aligns with the datasets acquired by Guo et al.

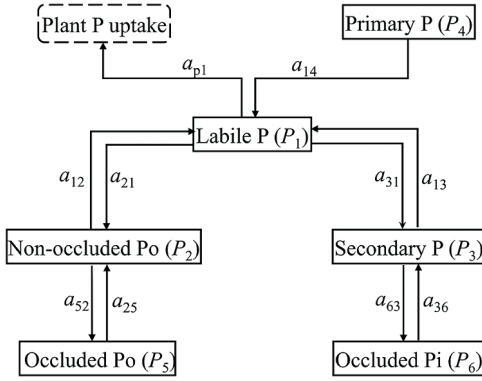


Figure 11.1. Schematic representation of the soil P model. Primary P indicates primary mineral P; secondary P indicates secondary mineral P; Pi indicates inorganic P; Po indicates organic P. An 'a' on an arrow indicates the coefficient of soil P transformation: plant immobilization (a_{p1}), weathering (a_{14}), microbial immobilization (a_{21}), mineralization (a_{12}), sorption/precipitation (a_{31}), desorption/dissolution (a_{13}), and solid-phase transformations (a_{25} , a_{52} , a_{36} , and a_{63}). Derived from Hou et al. (2019).

(2000). Our model groups the measured soil P fractions into six ecologically meaningful soil P pools, whose dynamics we track in the model. These six pools are labile P (P_1 in Figure 11.1), non-occluded Po (P_2 in Figure 11.1), secondary mineral P (P_3 in Figure 11.1), primary mineral P (P_4 in Figure 11.1), occluded Po (P_5 in Figure 11.1), and occluded Pi (P_6 in Figure 11.1). The labile P is inorganic P that is readily available to plants. The non-occluded Po is organic P that is sorbed by soil particles or secondary minerals (e.g., aluminum and iron oxides) and that can be mineralized by enzymes. The secondary mineral P is inorganic P that is sorbed by secondary minerals, which is not readily available to plants but is exchangeable with the labile P. The primary mineral P is P that is associated with primary minerals and exists mainly as apatite P. The occluded Po is organic P that is stabilized by soil minerals and that is presumably not mineralizable by enzymes unless dissolved. The occluded Pi is inorganic P that is occluded by soil minerals or aggregates and turns over very slowly. In Guo et al. (2000), the occluded Po and occluded Pi were not separated but determined together as residual P (P not extracted by the chemical reagents used). A proportion of residual P in organic forms (OP_o) is introduced here to represent the amounts of occluded Po (calculated as residual P \times OP_o) and occluded Pi (calculated as residual P \times $(1-OP_o)$).

We consider all major soil P transformations in our model, as detailed in Figure 11.1. We calculate plant P uptake during a specific period as the sum of the decreases in P_{1-6} during the period. For instance, we calculate plant P uptake after cropping 14 as the difference between the sum of P_{1-6} at cropping 0 and the same sum after cropping 14. We don't consider P leaching in our model, because soil moisture content was maintained near soil available water capacity during the experiment. We don't consider atmospheric P deposition in our model either because the experiment was performed in a greenhouse. Moreover, we don't consider litterfall in our model, because the plants were young (≤ 45 days of growth) and unlikely to produce any litterfall. Matlab code and the eight datasets of soil P fractions are freely accessible via Hou et al. (2019).

Reflecting the structure depicted in Figure 11.1, we represent soil P dynamics in the model by the following set of balance equations.

$$\left\{ \begin{array}{l} \frac{dP_1(t)}{dt} = a_{14}P_4(t) + a_{12}P_2(t) + a_{13}P_3(t) - k_1P_1(t) \\ \frac{dP_2(t)}{dt} = a_{21}P_1(t) + a_{25}P_5(t) - k_2P_2(t) \\ \frac{dP_3(t)}{dt} = a_{31}P_1(t) + a_{36}P_6(t) - k_3P_3(t) \\ \frac{dP_4(t)}{dt} = -k_4P_4(t) \\ \frac{dP_5(t)}{dt} = a_{52}P_2(t) - k_5P_5(t) \\ \frac{dP_6(t)}{dt} = a_{63}P_3(t) - k_6P_6(t) \end{array} \right. \quad (11.1)$$

Note that we don't list plant P uptake in Equation (11.1), because we treat it as a soil P flux. We don't use a_{p1} in Equation (11.1), because we can estimate it directly from a_{21} and a_{31} by the following equation:

$$a_{p1} = 1 - a_{21} - a_{31} \quad (11.2)$$

We can summarize the set of balance Equations (11.1) by the following first-order matrix equation:

$$\frac{dP(t)}{dt} = AKP(t) \quad (11.3)$$

where A , K , and $P(t)$ are the matrices given by

$$A = \begin{pmatrix} -1 & a_{12} & a_{13} & 1 & 0 & 0 \\ a_{21} & -1 & 0 & 0 & 1 & 0 \\ a_{31} & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1-a_{12} & 0 & 0 & -1 & 0 \\ 0 & 0 & 1-a_{13} & 0 & 0 & -1 \end{pmatrix}$$

$$K = \text{diag}(k) = \begin{pmatrix} k_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & k_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & k_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & k_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & k_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & k_6 \end{pmatrix}$$

$$P(t) = \begin{pmatrix} P_1(t) \\ P_2(t) \\ P_3(t) \\ P_4(t) \\ P_5(t) \\ P_6(t) \end{pmatrix}$$

Matrix A gives the transfers of P between the individual P pools, as described by the arrows in Figure 11.1. The elements (a_{ij}) are the P transfer coefficients, representing the fraction of P entering the i th (row) pool from the j th (column) pool. a_{52} is calculated as $1 - a_{12}$; a_{63} is calculated as $1 - a_{13}$; a_{14} , a_{25} , and a_{36} are fixed at 1.0. K is a 6×6 diagonal matrix representing the release rates of six soil P pools (units: $\text{g P g}^{-1} \text{P d}^{-1}$; for convenience, $\text{g g}^{-1} \text{d}^{-1}$ was used in the following), i.e., the amount of P leaving each of the soil P pools per day. $P(t)$ describes the sizes of the soil P pools at time t .

Model Validation and Data Assimilation

We first validate the matrix P model with the eight datasets of soil P pool measurements in Guo et al. (2000). We use the same parameter values for simulating P pools of all the eight soils. In general, the soil P model can simulate temporal changes in soil P pools reasonably well, with a better performance for labile P , secondary mineral P , and occluded P than for non-occluded P_0 and primary mineral P .

We then use a data assimilation approach to estimate the values of parameters describing soil P

dynamics. The approach, known as the Metropolis-Hastings algorithm, is described in more detail in the lectures and practice in unit 6 on terrestrial C dynamics (see Chapter 22). We run data assimilation formally for five replicates and 500,000 times for each soil to examine the convergence of the parameters. We test the convergence of the sampling chains by the Gelman-Rubin (G-R) diagnostic method to ensure that the within-run variation is roughly equal to the between-run variation. The Gelman-Rubin method is described in detail in Chapter 22.

After data assimilation, the soil P model can simulate well temporal changes in P pools of all the eight soils. Results on two soils are shown in Figure 11.2. Relationships between the measured and modeled soil P pools were mostly significant ($P < 0.05$), with R^2 values generally larger for labile P (mean 0.90), secondary mineral P (0.82), and occluded P (0.87) than for non-occluded P_0 (0.43) and primary mineral P (0.50). The relatively poor simulation of non-occluded P_0 was probably due to its relatively large measurement errors and dynamic nature.

NEW KNOWLEDGE EMERGING FROM DATA ASSIMILATION WITH THE MATRIX MODEL

Soil P Dynamics Quantified by Data Assimilation

Maximum likelihood estimates and uncertainty of the turnover rates of all soil P pools were estimated, including those of soil occluded P_i and occluded P_0 , which are rarely quantified in measurements. Estimated parameter values for a slightly weathered soil and a strongly weathered soil are shown in Table 11.1. Turnover rates of soil inorganic P pools generally decreased in the following order: labile $P >$ secondary mineral $P >$ occluded P_i (Table 11.1). The turnover rate of soil non-occluded P_0 was faster than that of soil occluded P_0 (Table 11.1).

The turnover rates of soil P pools (Table 11.1) were generally comparable with the values reported in previous studies using isotopic and spectroscopic measurements in the laboratory. However, the turnover rates of soil P pools here could be higher than those in the field, because the datasets used for data assimilation were derived from an experiment in a greenhouse environment, where plant P uptake and soil P depletion could be faster than in the field.

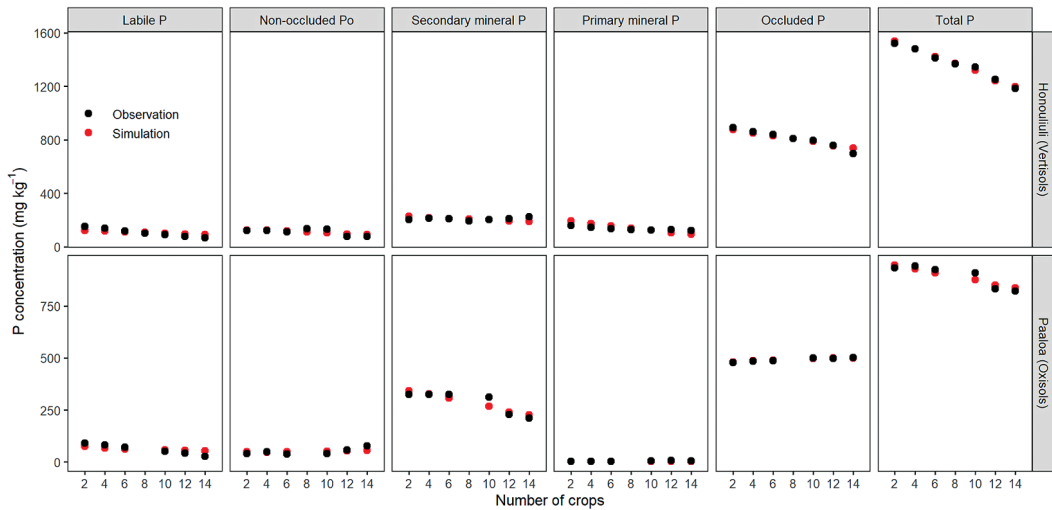


Figure 11.2. **Observed vs. simulated temporal changes in P pools of two soils.** The two soils are Honouliuli and Paaloa, which are typical Vertisols (slightly weathered) and Oxisols (strongly weathered), respectively. Derived from Hou et al. (2019).

Rates of transformations among all major P pools were also estimated (Table 11.1). These estimates can convey deep insights into soil P dynamics and soil P bioavailability. For example, the proportions of labile P flowing to secondary mineral P (mean 0.35) and non-occluded P_0 (0.52) were on

average larger than that flowing to plants (0.13) (Table 11.1). This result suggests that soil secondary minerals and microbes were stronger competitors of soil labile P than plants. Both turnover rate of labile P and transfer coefficients related to labile P differed among soils (Table 11.1), suggesting that

TABLE 11.1
Physicochemical properties and maximum likelihood estimates of model parameters describing P dynamics of two soils

Parameter	Unit	Honouliuli	Paaloa
Soil order		Vertisols	Oxisols
Weathered extent		Slightly	Strongly
Total P concentration at the start of experiment	g kg^{-1}	1840	596
0.5 M NaHCO_3 extractable P concentration	mg kg^{-1}	26.3	1.1
pH in water		7.26	5.05
Organic C concentration	g kg^{-1}	16.6	40
Exchange cation concentration	$\text{cmol}_c \text{ kg}^{-1}$	30.6	5.5
Acid ammonium oxalate extracted Fe concentration	g kg^{-1}	3.4	7.48
Acid ammonium oxalate extracted Al concentration	g kg^{-1}	1.43	2.98
Sand content	g kg^{-1}	56.5	53.8
Silt content	g kg^{-1}	363.6	205.4
Clay content	g kg^{-1}	580	740.8
Turnover rate of labile P (k_1)	$\text{g g}^{-1} \text{ d}^{-1}$	0.04	0.048
Turnover rate of non-occluded P_0 (k_2)	$\text{g g}^{-1} \text{ d}^{-1}$	0.022	0.073
Turnover rate of secondary mineral P (k_3)	$\text{g g}^{-1} \text{ d}^{-1}$	0.044	0.012
Turnover rate of primary mineral P (k_4)	$\text{g g}^{-1} \text{ d}^{-1}$	0.00193	0.00015

(Continued)

TABLE 11.1 (CONTINUED)

Parameter	Unit	Honouliuli	Paaloa
Turnover rate of occluded P ₀ (k_5)	$\text{g g}^{-1} \text{d}^{-1}$	0.0077	0.0067
Turnover rate of occluded P _i (k_6)	$\text{g g}^{-1} \text{d}^{-1}$	0.0062	0.009
Coef. of transfer from labile P to non-occluded P ₀ (a_{21})	Unitless	0.51	0.53
Coef. of transfer from labile P to secondary mineral P (a_{31})	Unitless	0.33	0.38
Coef. of transfer from labile P to plant (a_{p1})	Unitless	0.16	0.09
Coef. of transfer from non-occluded P ₀ to labile P (a_{12})	Unitless	0.97	0.23
Coef. of transfer from non-occluded P ₀ to occluded P ₀ (a_{52})	Unitless	0.03	0.77
Coef. of transfer from secondary mineral P to labile P (a_{13})	Unitless	0.23	0.72
Coef. of transfer from secondary mineral P to occluded P _i (a_{63})	Unitless	0.77	0.28
Proportion of occluded P in organic form (OP_o)	Unitless	0.28	0.02

Derived from Hou et al. (2019)

labile P dynamics varied among soils. This result suggests that not only the amount of but also the dynamics of soil labile P control soil P availability.

The estimated model parameters from data assimilation provide information about the datasets and the model in two other aspects. Firstly, posterior distributions of parameters can tell how well model parameters were constrained by the datasets. Secondly, relationships among model parameters can reflect relationships defined by the model structure, correlations between the soil P pools, errors, or any combination of all three.

Soil P Dynamics in Relation to Other Ecosystem Properties

Soil P dynamics were revealed to differ between the lightly and the strongly weathered soils (Table 11.1). Turnover rate of secondary mineral P was ~ 4 times higher in the lightly weathered soil ($0.044 \text{ g g}^{-1} \text{ d}^{-1}$) than in the highly weathered soils ($0.012 \text{ g g}^{-1} \text{ d}^{-1}$) (Table 11.1). The proportion of labile P flowing to plants was higher in the slightly weathered soil (0.16) than in the strongly weathered soils (0.09); the opposite was true for proportion of labile P flowing to secondary mineral P (slightly: 0.33; strongly: 0.38) (Table 11.1).

The proportion of labile P flowing to plants increased with soil pH; correspondingly, the proportion of labile P flowing to secondary mineral P decreased with soil pH (Figure 11.3c). Relationships of soil organic C concentration with

model parameters were generally opposite to those of soil pH (Figure 11.3). These results suggest the regulation of plant and soil microbial competition for labile P by soil pH and organic C concentration. The proportion of labile P flowing to secondary mineral P decreased with increasing soil pH. This was probably because of decrease in soil P sorption capacity with increasing soil pH. This mechanism also explains the increase in turnover rate of secondary mineral P with increasing soil pH (Figure 11.3a). Increase in the proportion of labile P flowing to secondary mineral P with increasing soil organic C concentration (Figure 11.3d) may be attributable to the sorption of P by organic-metal complexations in soils. This mechanism additionally explains the decrease in turnover rate of secondary mineral P with increasing soil organic C concentration (Figure 11.3b).

SUMMARY

Despite increasing research efforts, there are still some uncertainties in the structure of soil P models and much larger uncertainties in the parameter values of soil P models. These uncertainties can be potentially reduced by the adoption of matrix and data assimilation approaches in soil P modeling. The matrix representation can make it easier to depict, understand, and compare soil P models compared to the traditional representation as a set of balance equations. Matrix approaches can also reduce the computational costs of models by enabling semi-analytical spin-up and make it computationally

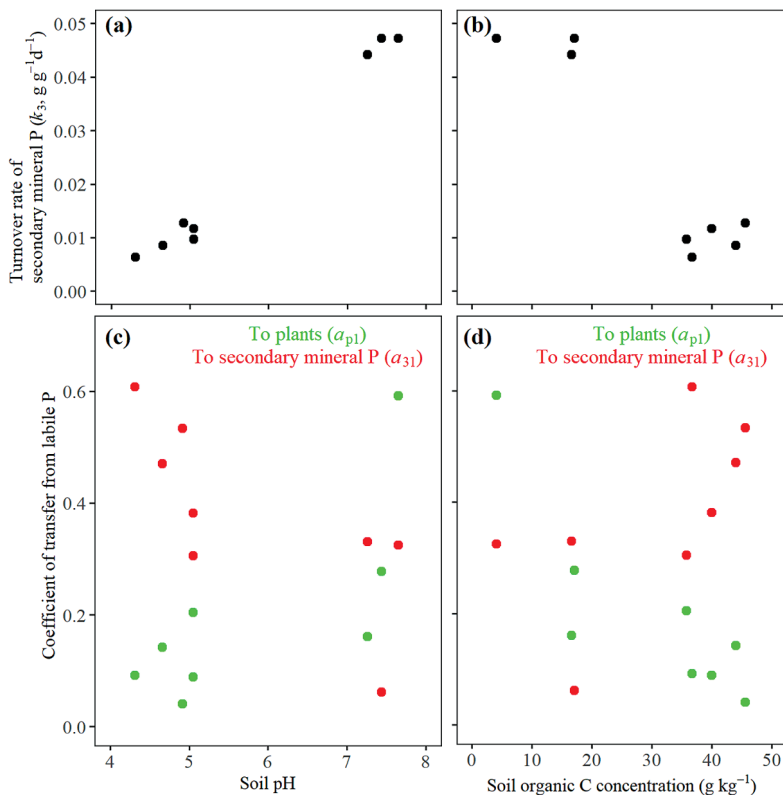


Figure 11.3. **Soil P dynamics in relation to soil pH and organic C concentration.** Turnover rate of secondary mineral P vs. (a) soil pH ($R^2 = 0.99$, $P < 0.001$) and (b) soil organic C concentration ($R^2 = 0.87$, $P = 0.001$). (c) Coefficients of transfer from labile P to plants (green; $R^2 = 0.53$, $P = 0.042$) and secondary mineral P (red; $R^2 = 0.71$, $P = 0.009$) vs. soil pH. (d) Coefficients of transfer from labile P to plants (green; $R^2 = 0.56$, $P = 0.034$) and secondary mineral P (red; $R^2 = 0.38$, $P = 0.106$) v. soil organic C concentration. Derived from Hou et al. (2019).

more feasible to assimilate soil P observations into models. An example study was given to show how to build a soil P model in matrix form, and how assimilating soil P observations into the matrix model can expose insights into soil P dynamics and availability. Overall, this chapter shows that assimilating soil P observations into a matrix model of soil P dynamics can improve our understanding of soil P dynamics and availability.

SUGGESTED READING

Hou, E., X. Lu, L. Jiang, D. Wen and Y. Luo (2019). Quantifying soil phosphorus dynamics: a data

assimilation approach. *Journal of Geophysical Research: Biogeosciences* **124**: 2159–2173.

QUIZZES

1. Why is a unified matrix equation preferred over traditional balance equations in soil P studies?
2. Was turnover rate of soil labile P constant across soils?
3. Which model parameter is soil labile P pool size most sensitive to?
4. What purpose was data assimilation used for in the example study?

CHAPTER TWELVE

Practice 3

DIAGNOSTIC VARIABLES IN MATRIX MODELS

Xingjie Lu

Sun Yat-sen University, Guangzhou, China

CONTENTS

Motivation of the Uncertainty Diagnostics / 95
The Mathematical Foundation of the Diagnostics of Land Carbon Cycle Models / 95
Carbon Storage Capacity and Carbon Storage Potential / 97
Residence Time and Carbon Input / 98
Suggested Reading / 99

This practice helps you understand diagnostic variables in biogeochemistry matrix models. The key diagnostic variables include carbon storage capacity, storage potential, residence time, and input. We use a TECO matrix model to demonstrate how to incorporate diagnostic variables from carbon balance equations. We verify the theory that carbon storage capacity represents an attractor of carbon storage dynamic. Meanwhile, we understand how changes in carbon turnover rate and input allocation fraction affect the carbon residence time and therefore the carbon storage capacity. Carbon residence time and carbon input are two essential diagnostics to characterize the steady state carbon storage in land carbon cycle modeling.

MOTIVATION OF THE UNCERTAINTY DIAGNOSTICS

Land carbon cycle models are frequently used to estimate biosphere-atmosphere feedback. However, different models often provide quite divergent predictions in land carbon uptake and storage as revealed via model intercomparison projects. Great efforts have been made to understand differences among models. Even so, we still have difficulties in identifying causes of model differences.

To improve understanding of why models behave differently, we need effective diagnostic tools. This is one of the motivations for developing the matrix approach. The traceability analysis with the matrix approach offers a framework to understand each traceable component first and then assemble them together to analyze the carbon storage dynamics.

The matrix approach enables decomposition of modeled carbon storage to a few traceable components according to the mathematical properties of the matrix equation. In this way, we can conduct traceability analysis (see unit 5) to understand the causes of model uncertainty.

THE MATHEMATICAL FOUNDATION OF THE DIAGNOSTICS OF LAND CARBON CYCLE MODELS

Previous chapters have demonstrated that the carbon storage dynamics in most land carbon cycle models can be formulated in a matrix form (Luo et al. 2017; Chapter 1):

$$\frac{dX(t)}{dt} = Bu(t) + G(t)X(t) \quad (12.1)$$

where $X(t)$, as a vector, represents carbon storage in multiple pools, $u(t)$ as a scalar represents the amount of carbon input from net primary production (NPP, i.e., photosynthesis minus autotrophic respiration), B , as a vector, is the carbon partitioning from NPP to multiple pools. G , as a matrix, indicates the carbon transfer network among pools.

Equation (12.1) can be reformulated to a diagnostic format:

$$X(t) = X_c(t) + X_p(t) \quad (12.2)$$

where X_c is carbon storage capacity and X_p is carbon storage potential. The carbon storage capacity can be further decomposed into ecosystem residence time (τ_E) and carbon input (u):

$$X_c(t) = \tau_E u(t) \quad (12.3)$$

where ecosystem residence time is defined by:

$$\tau_E = G^{-1}B \quad (12.4)$$

The carbon storage potential is the product of the inverse of matrix G and the carbon storage growth rate:

$$X_p(t) = G^{-1} \frac{dX(t)}{dt} \quad (12.5)$$

In this practice, we are going to focus on the carbon storage capacity (X_c), carbon storage potential (X_p), carbon residence time (τ_E) and carbon input (u), which are the most important diagnostics in the matrix approach.

Equations 12.3–12.5 show that the three terms carbon storage capacity (X_c), carbon storage potential (X_p), and carbon residence time (τ_E) are related to the matrix G . Therefore, one of the most critical steps to calculate the above diagnostics is to find the matrix G , which equals $A(t)\xi(t)K$.

Exercise 1 and Exercise 2 are to identify diagnostic variables from the TECO and CLM5 matrix models.

EXERCISE 1

The matrix equation of the TECO model can be written into one matrix equation:

$$\frac{dX(t)}{dt} = Bu(t) + A(t)\xi(t)KX(t) \quad (12.6)$$

u is the C input scalar ($\text{gC m}^{-2} \text{s}^{-1}$). B represents the C allocation fraction vector (unitless). A is the partitioning coefficient matrix (unitless). K is a diagonal turnover rate matrix (s^{-1}). X is C pool size vector (gC m^{-2}).

Ex 1.1 Find the **matrix G** of Equation 12.6. This matrix was defined in Equation 12.1: $G = \dots$

Ex 1.2 Write down the **diagnostic form** of Equation 12.6: $X(t) = \dots$

Ex 1.3 Identify the residence time, C input, C storage capacity and C storage potential from the diagnostic form.

Residence time: $\tau_E = \dots$

C input: $u = \dots$

C storage capacity: $X_c = \dots$

C storage potential: $X_p = \dots$

EXERCISE 2 (optional)

The matrix equation of the CLM5 vegetation C cycle model can be written in the form of one matrix equation:

$$\begin{aligned} \frac{dX(t)}{dt} = & Bu(t) + A_{ph}(t)K_{ph}(t)X(t) \\ & + A_{gm}K_{gm}X(t) + A_{fi}K_{fi}(t)X(t) \end{aligned} \quad (12.7)$$

u is the C input scalar, ($\text{gC m}^{-2} \text{s}^{-1}$). B represents the C allocation fraction vector (unitless). A is the partitioning coefficient matrix (unitless). K is a diagonal turnover rate matrix (s^{-1}). X is C pool size vector (gC m^{-2}). The subscripts, **ph**, **gm**, and **fi** represent the partitioning coefficient and turnover rate related to phenology, gap mortality, and fire processes.

Ex 2.1 Find the **matrix G** of Equation 12.7: $G = \dots$

Ex 2.2 Write down the **diagnostic form** of Equation 12.7: $\mathbf{X}(t) = \dots$

Ex 2.3 Identify residence time, C input, C storage capacity, and C storage potential from the diagnostic form.

Residence time: $\tau_E = \dots$

C input: $\mathbf{u} = \dots$

C storage capacity: $X_c = \dots$

C storage potential: $X_p = \dots$

CARBON STORAGE CAPACITY AND CARBON STORAGE POTENTIAL

The carbon storage capacity and carbon storage potential are two important diagnostics. The carbon storage capacity represents the maximal amount of carbon that a land ecosystem can store. The carbon storage potential represents the difference between carbon storage capacity and current

carbon storage. The two diagnostics are the first tier variables to define land carbon dynamics for any modeling study. The carbon storage capacity is the attractor and the sign of the carbon storage potential represents the direction. That is, the carbon storage will ultimately reach carbon storage capacity regardless of initial values, carbon inputs and other model parameters. **Exercise 3** will allow us to verify this notion.

EXERCISE 3

Run the TECO matrix model (Ex 3.1), then change the initial value (Ex 3.2) and input (Ex 3.3). Learn whether carbon storage changes to approach the carbon storage capacity. Learn how carbon storage potential changes.

Ex 3.1. Follow instructions to run the TECO matrix model in CarboTrain:

- Select **Unit 3**
- Select **Exercise 3**
- Select **Default**
- Select **Set Output Folder**
- Select **Open Source Code**
- Read lines 10–44, get familiar with parameters, input, and initial value. The file `test_p3.py` can be edited at this step.
- Run Exercise**
- Check results in Output Folder. Time-dependent variation in carbon input, each pool size, residence time, total ecosystem carbon storage, total ecosystem carbon storage capacity, and carbon storage potential appear in the output file `output.xls`. Figures are in `results.png`.

Questions: Does carbon storage change towards carbon storage capacity? How does carbon storage potential change?

Ex 3.2. Try different initial values

- Repeat Ex 3.1, but select “**Change initial pool size I**” instead of “**Default**”.

Open source code and change the initial value to (1000.0 20000.0 50.0 3000.0 200.0 15000.0 40000.0) at line 44 of `test_p3.py`. **Run Exercise** and check the total ecosystem carbon storage and total ecosystem carbon storage capacity in `results.png`.

- Repeat Ex 3.1, but select “**Change initial pool size II**” instead of “**Default**”. **Open source code** and change the initial value to (280.0 5000.0 15.0 800.0 50.0 4000.0 10000.0) at line 44 of `test_p3.py`. **Run Exercise** and check the total ecosystem carbon storage and total ecosystem carbon storage capacity in `results.png`.

Questions: If the initial pool size changes, does carbon storage change towards carbon storage capacity? Does the carbon storage capacity depend on the initial value? How does carbon storage potential change?

Ex 3.3. Try different C inputs

- Repeat Ex 3.1, but select “**Change carbon input I**” instead of “**Default**”. **Open source code** and change the C input to 0.00001123 at line 38 of `test_p3.py`. **Run Exercise** and check the total ecosystem carbon storage and total ecosystem carbon storage capacity in `results.png`.
- Repeat Ex 3.1, but select “**Change carbon input II**” instead of “**Default**”.

Open source code and change the C input to 0.00004490 at line 38 of `test_p3.py`. **Run Exercise** and check the total ecosystem carbon storage and total ecosystem carbon storage capacity in `results.png`.

Questions: If the carbon storage capacity changes, does carbon storage still change towards carbon storage capacity? Does the carbon storage capacity depend on the C input? How does carbon storage potential change? Would the carbon storage always equal the carbon storage capacity at steady-state? What does zero carbon storage potential stand for?

These questions are critical to the conceptual foundation of the carbon storage capacity and carbon storage potential. Results from Exercise 3 guide you to understand why carbon storage capacity and carbon storage potential are important diagnostics for carbon cycle modeling and how the definitions of the two variables reflect the inherent property of land carbon cycle.

RESIDENCE TIME AND CARBON INPUT

Now let us further explore the carbon storage capacity, which is described by two additional traceable components: residence time and carbon input. Climate change causes changes in

carbon cycling, usually via changes in residence time and/or carbon input. For example, rising atmospheric CO₂ concentration usually enhances carbon input (u). Rising air temperature usually increases soil carbon decomposition and thus turnover rate of soil pools (K). When a forest is converted to cropland under land use change, the carbon allocation to woody tissue is usually set to zero, leading to changes in the carbon allocation fraction (B). Thus, changes in ecosystem carbon storage in response to various climate change factors can be explained by diagnostics related to residence time and carbon input in the matrix approach. Exercise 4 will let us explore how changes in parameter values related to residence

EXERCISE 4

Run the TECO matrix model, and make following changes in parameters or carbon input (Ex 4.1–4.3). Observe how parameters influence the carbon storage capacity.

Ex 4.1 Try a different turnover rate.

- Repeat Ex 3.1, but Select “**Exercise 4**”, and “**Low foliage turnover**”. **Open source code** and change the foliage turnover rate to 8.8×10^{-4} , which is the first element of “temp” at line 26. **Run Exercise** and check the total ecosystem carbon storage capacity, residence time and carbon input at steady state in `output.xls` and `results.png`.
- Repeat Ex 3.1, but Select “**Exercise 4**”, and “**Low passive soil turnover**”. **Open source code** and change the passive soil turnover rate to 7.739×10^{-7} , which is the last element of “temp” at line 26.

Run Exercise and check the total ecosystem carbon storage capacity, residence time and carbon input at steady state in `output.xls` and `results.png`.

Questions: Is the carbon storage capacity in Ex 4.1a or Ex 4.1b lower or higher than that in the default model (Ex 3.1)? Can the lower or higher carbon storage capacity be attributed to residence time or carbon input? How do changes in turnover of foliage or passive soil carbon impact residence time or carbon input? Which impact is stronger? Is it the impact of lower foliage turnover rate or the impact of lower passive soil turnover rate? Why?

Ex 4.2 Try different allocation fractions.

- Repeat Ex 3.1, but Select “**Exercise 4**”, and “**High allocation to foliage**”. **Open source code** and change the allocation fraction to (0.55, 0.45...), which are the first two elements of “B” at line 10. **Run**

Exercise and check the total ecosystem carbon storage capacity, residence time and carbon input at steady state in `output.xls` and `results.png`.

- b. Repeat Ex 3.1, but Select “**Exercise 4**”, and “**High allocation to wood**”. **Open source code** and change the allocation fraction to (0.2, 0.8...), which are the first two elements of “B” at line 10. **Run Exercise** and check the total ecosystem carbon storage capacity, residence time and carbon input at steady state in `output.xls` and `results.png`.

Questions: Is the carbon storage capacity in Ex 4.2a or Ex 4.2b lower or higher than that in the default model (Ex 3.1)? Can the lower or higher carbon storage capacity be attributed to residence time or carbon input? How do changes in allocation fraction impact residence time or carbon input? Why?

Ex 4.3 Try multiple changes of parameters in different C input and turnover rate.

- a. Repeat Ex 3.1, but Select “**Exercise 4**”, and “**Multiple Changes I**”. **Open source code** and change the C input to 1.123×10^{-5} at line 26 and 38, and change the slow and passive soil turnover rate to 2.99×10^{-5} and 5.159×10^{-7} , which

are the last two elements of “temp” at line 25. **Run Exercise** and check the total ecosystem carbon storage capacity, residence time and carbon input at steady state in `output.xls` and `results.png`.

- b. Repeat Ex 3.1, but Select “**Exercise 4**”, and “**Multiple Changes II**”. **Open source code** and change the C input to 4.49×10^{-5} at line 26 and 38, and change the slow and passive soil turnover rate to 2.69×10^{-4} and 4.643×10^{-6} , which are the last two elements of “temp” at line 25. **Run Exercise** and check the total ecosystem carbon storage capacity, residence time and carbon input at steady state in `output.xls` and `results.png`.

Questions: What are the common features or differences in carbon storage dynamics and carbon storage capacity between Ex 4.3a and Ex 4.3b? Can you see how residence time and carbon input cause differences in carbon storage capacity? What causes these differences in residence time or carbon input?

Understanding how carbon storage capacity changes with carbon residence time and carbon input changes is essential in diagnostics for carbon cycle modeling. This will be further explored in unit 5 on traceability analysis.

time or carbon input result in changes in the carbon storage capacity.

Diagnostic capability is one of the most important benefits offered by the matrix approach. This practice only provides you with simple cases to understand the terrestrial ecosystem carbon cycle. To understand the land carbon cycle at the earth system scale, we need to analyze wider ranges of spatial and temporal variations from multiple models in response to rising atmospheric CO₂ concentration, and climate warming (Lu et al. 2018).

SUGGESTED READING

- Luo, Y. Q., Shi, Z., Lu, X. J., Xia, J. Y., Liang, J. Y., Jiang, J., et al. (2017). Transient dynamics of terrestrial carbon storage: mathematical foundation and its applications. *Biogeosciences*, **14**(1), 145–161
- Lu, X., Du, Z., Huang, Y., Lawrence, D., Kluzek, E., Collier, N., Lombardozzi, D., Sobhani, N., Schuur, E., Luo, Y., (2020). Full implementation of matrix approach to biogeochemistry module of Community Land Model version 5 (CLM5). *Journal of Advances in Modeling Earth Systems*, **12**: e2020MS002105.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

UNIT FOUR

Semi-Analytic Spin-Up (SASU)



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER THIRTEEN

Nonautonomous ODE System Solver and Stability Analysis

Ying Wang

University of Oklahoma, Flagstaff, USA

CONTENTS

Introduction / 103
Analytical Solution / 104
First Order Non-homogeneous Scalar Equation / 104
One-Pool Model / 104
Homogeneous Nonautonomous ODEs System / 105
Non-homogeneous Nonautonomous ODEs System / 105
N-Pool Model / 106
Mathematica Calculation for the Analytical Solution of a 3-pool Model / 107
Stability / 108
Instantaneous Steady State / 108
Instantaneous Steady State for a 3-pool Model / 109
Global Attractor / 109
The Global Attractor of the N-Pool Model / 111
General Stability Statements / 111
Suggested Reading / 112
Quizzes / 112

The matrix equation model for the terrestrial carbon dynamics is a system of nonautonomous ODEs, which naturally inherits the mathematical difficulties in the solution process and stability studies of its equilibria. This chapter introduces the mathematical properties of this matrix equation model. In particular, we will study: (1) the analytical solution of the matrix equation model, providing a concrete 3-pool terrestrial carbon dynamics to demonstrate the solution process; and (2) the stability analysis of the matrix equation model.

INTRODUCTION

We consider the following system of nonautonomous ordinary differential equations (ODEs):

$$X' = \xi(t)ACX + Bu(t) \quad (13.1)$$

where $X(t)$ is a vector of carbon pool sizes, X_0 is a vector of initial values of the carbon pools; and $\xi(t)$ is an environmental scalar representing effects of temperature and moisture on the carbon transfer among pools, A and C are carbon transfer coefficients between plant, litter, and soil pools; $u(t)$ is the photosynthetically fixed carbon and usually estimated by canopy photosynthetic models, B is a vector of partitioning coefficients of the photosynthetically fixed carbon to plant pools. For background to this equation, see chapter 1.

Our goal is to develop an analytical solution of (13.1), and to understand and predict how the equilibrium state stability is impacted by various environmental scalar functions $\xi(t)$, transfer matrices A and C , and photosynthetic input $u(t)$.

ANALYTICAL SOLUTION

In mathematical language, governing equation (13.1) is a system of non-homogeneous nonautonomous ODEs, and the derivation of its solution is complicated. Therefore, instead of deriving the solution of (13.1) directly, we will start with studying the analytical solution of a scalar non-homogeneous nonautonomous ODE, and then extend it to the system situation.

First Order Non-homogeneous Scalar Equation

A first-order linear non-homogeneous scalar equation has the following general form:

$$x' + p(t)x = q(t) \quad (13.2)$$

A standard way to solve (13.2) is the integrating factor method, which proceeds as follows. Let:

$$P(t) = \int_{t_0}^t p(s) ds,$$

then the integrating factor is given by:

$$h(t) = e^{P(t)}. \quad (13.3)$$

After multiplying the integrating factor (13.3) to both sides of (13.2), we get:

$$\begin{aligned} e^{P(t)}(x' + p(t)x) &= e^{P(t)}q(t) \\ e^{P(t)}x' + e^{P(t)}p(t)x &= e^{P(t)}q(t) \end{aligned} \quad (13.4)$$

Integration by parts allows us to combine the left-hand side of (13.4) into one derivative:

$$\left(e^{P(t)}x \right)' = e^{P(t)}q(t).$$

Therefore, integrating the entire equation gives us the analytical solution formula for the first-order linear non-homogeneous scalar equation (13.2):

$$\begin{aligned} e^{P(t)}x &= \int_{t_0}^t e^{P(s)}q(s) ds + C \\ x(t) &= e^{-P(t)} \left(\int_{t_0}^t e^{P(s)}q(s) ds + C \right) \end{aligned} \quad (13.5)$$

where:

$$C = x(t_0)$$

Let's take a look at the following straightforward example to get ourselves familiar with this solution procedure:

$$x' - x = e^{2t}. \quad (13.6)$$

The integrating factor in this case is:

$$h(t) = e^{P(t)} = e^{\int_{t_0}^t -1 ds} = e^{-t+t_0}. \quad (13.7)$$

Now, we multiply the integrating factor (13.7) to the ODE (13.6):

$$\left(e^{-t}x \right)' = e^t,$$

then integrate this equation, we get:

$$e^{-t}x = e^t + C,$$

therefore:

$$x = e^{2t} + Ce^t$$

where:

$$C = x(0) - 1$$

One-Pool Model

Now, we consider a one-pool model, that is, the vector X in (13.1) becomes a scalar, then (13.1) becomes a first-order linear non-homogeneous scalar equation (13.2) with

$$p(t) = -\xi(t)AC, \quad q(t) = Bu(t).$$

where $X(t)$, A , C and B are all scalars, instead of matrices. We can apply the solution formulae (13.5) to solve the one-pool terrestrial carbon cycle system model (13.1), we have that:

$$X(t) = e^{-P(t)} \left(\int_{t_0}^t e^{P(s)} Bu(s) ds + C \right)$$

where:

$$P(t) = \int_{t_0}^t -\xi(s)ACds, C = X(t_0)$$

Homogeneous Nonautonomous ODEs System

Before moving on to study the analytical solution for the general n -pool model (13.1), we first prepare ourselves with the solution to an initial value problem of a homogeneous nonautonomous ODEs system in the following form:

$$\begin{cases} X' = A(t)X \\ X(t_0) = e_i \end{cases} \quad (13.8)$$

where:

$$e_i = \left(\underbrace{0, \dots, 0}_{i-1 \text{ copies of } 0}, 1, \underbrace{0, \dots, 0}_{n-i \text{ copies of } 0} \right)^T$$

is an n -vector with only the i th entry equals to 1, and all the other entries equal to 0. The term “homogeneous” means that every term in equation (13.8) includes X , and the term “nonautonomous” means that X' depends on the independent variable t explicitly, i.e., the right-hand side of equation (13.8) contains the term $A(t)$ which is a function of t . Note that system (13.8) is one system with n different initial conditions. If we solve (13.8) for $i = 1, 2, \dots, n$, we will get solutions $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ corresponding to the n different initial conditions e_i 's ($i = 1, 2, \dots, n$), and this set of solutions forms a fundamental matrix:

$$\Phi(t) = \begin{bmatrix} X^{(1)} & X^{(2)} & \dots & X^{(n)} \end{bmatrix} \quad (13.9)$$

of the homogeneous nonautonomous ODEs System:

$$X' = A(t)X. \quad (13.10)$$

Non-homogeneous Nonautonomous ODEs System

Now we consider the following nonhomogeneous nonautonomous ODEs system:

$$X' = A(t)X + g(t). \quad (13.11)$$

The difference between equation (13.11) and equation (13.10) is that equation (13.11) includes a non-homogeneous term $g(t)$, i.e., a term which doesn't include X , that's why equation (13.11) is called a non-homogeneous nonautonomous ODEs system. To look for the solution of equation (13.11), we will need to employ the fundamental matrix $\Phi(t)$ given in equation (13.9) of the corresponding homogeneous system (13.10). Assume that the solution of equation (13.11) has the following form

$$X = \Phi(t)Q(t) \quad (13.12)$$

where:

$$Q(t) = \begin{pmatrix} q_1(t) \\ \vdots \\ q_n(t) \end{pmatrix}$$

is a n -vector, i.e., the solution is a linear combination of the basis of the solution space of the corresponding homogeneous system (13.10). Then we plug the solution formula (13.12) into equation (13.11), we will get:

$$\Phi'(t)Q + \Phi(t)Q' = A\Phi Q + g. \quad (13.13)$$

Since $\Phi(t)$ is a fundamental matrix of equation (13.10), that means that every column of $\Phi(t)$ is a solution of equation (13.10), therefore we have that the fundamental matrix $\Phi(t)$ also satisfies equation (13.10), i.e.:

$$\Phi'(t) = A\Phi(t). \quad (13.14)$$

Equation (13.14) allows us to cancel the $\Phi'(t)Q$ and $A\Phi(t)Q$ terms on the left and right hand sides of equation (13.13), then equation (13.13) becomes:

$$\Phi(t)Q' = g,$$

Equivalently:

$$Q' = \Phi^{-1}(t)g(t)$$

because matrix Φ as a fundamental matrix, is clearly invertible. To solve for Q , we integrate this equation, and get:

$$Q = \int_{t_0}^t \Phi^{-1}(s)g(s)ds + C$$

Therefore, the solution to the non-homogeneous nonautonomous ODEs system (13.11) is:

$$X(t) = \Phi(t)Q + \Phi(t)C + \Phi(t) \int_{t_0}^t \Phi^{-1}(s)g(s)ds \quad (13.15)$$

where:

$$C = \Phi^{-1}(t_0)X(t_0)$$

N-Pool Model

The n-pool model:

$$\begin{aligned} X' &= \xi(t)ACX + Bu(t) \\ X(0) &= X_0 \end{aligned} \quad (13.16)$$

is in principle a non-homogeneous nonautonomous ODEs system (13.11) with:

$$\begin{aligned} A(t) &= \xi(t)AC \\ g(t) &= Bu(t). \end{aligned}$$

Therefore, we can apply the solution formulae (13.15) to solve (13.16):

$$X(t) = \Phi(t)C + \Phi(t) \int_{t_0}^t \Phi^{-1}(s)g(s)ds = \underbrace{e^{\int_0^t \xi(\sigma)AC d\sigma}}_{(a)} X_0 + \underbrace{\left(\int_0^t e^{\int_s^t \xi(\sigma)AC d\sigma} u(s)ds \right) B}_{(d)} \quad (13.17)$$

Notice that, in this case, the fundamental matrix of the corresponding homogeneous system:

$$X' = \xi(t)ACX$$

is:

$$\Phi(t) = e^{\int_0^t \xi(\sigma)AC d\sigma} \quad (13.18)$$

The calculation of the matrix exponential $e^{\int_0^t \xi(\sigma)AC d\sigma}$ is rather complicated. In principle, a matrix exponential e^A is defined in the same way as the scalar exponential:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

to be:

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!}.$$

If the matrix:

$$A = \text{diag}(a_i)_{i=1}^n$$

is a diagonal matrix with $\{a_i\}_{i=1}^n$ as the diagonal entries, then,

$$e^A = \text{diag}(e^{a_i})_{i=1}^n.$$

If the matrix A is diagonalizable by an invertible matrix P (for example, the eigenvector matrix of A), i.e.,

$$A = PAP^{-1}$$

where Λ is a diagonal matrix, then the matrix exponential becomes

$$e^A = Pe^{\Lambda}P^{-1}.$$

From this formulation, we see that the essential step to calculate a matrix exponential is to diagonalize that matrix. Now, we apply this essential step to derive the details in the solution formulae (13.17). In order to calculate terms (a) and (d) in (13.17), we first introduce an auxiliary function:

$$\phi(t) = e^{ACt}. \quad (13.19)$$

To calculate this matrix exponential, we first find the eigenvalues of the matrix product AC :

$$\lambda_i, i = 1, \dots, n$$

and a complete set of orthonormal eigenvectors of AC :

$$v_i, i = 1, \dots, n$$

and define the eigenvector matrix:

$$V = (v_1, \dots, v_n)$$

so that the matrix product AC can be diagonalized by the eigenvector matrix V , i.e.:

$$AC = V \text{diag}(\lambda_i)_{i=1}^n V^{-1},$$

which also gives that:

$$ACt = V \text{diag}(\lambda_i t)_{i=1}^n V^{-1},$$

then $\phi(t)$ given in (13.19) can be calculated by:

$$\phi(t) = e^{ACt} = V \text{diag}(e^{\lambda_i t})_{i=1}^n V^{-1} \quad (13.20)$$

Since the fundamental matrix $\Phi(t)$ given in (13.18) can be written as:

$$\Phi(t) = \phi(f(t)), \text{ where } f(t) = \int_0^t \xi(\sigma) d\sigma$$

this matrix exponential is therefore given as:

$$\begin{aligned} \Phi(t) &= e^{\int_0^t AC \xi(\sigma) d\sigma} = \phi(f(t)) \\ &= V \text{diag}(e^{\lambda_i f(t)})_{i=1}^n V^{-1} \\ &= V \text{diag} \left(e^{\lambda_i \int_0^t \xi(\sigma) d\sigma} \right)_{i=1}^n V^{-1}. \end{aligned}$$

hence, terms (a) and (d) in (13.17) are calculated as following:

$$A = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \quad C = \begin{pmatrix} c_1 & 0 & 0 \\ 0 & c_2 & 0 \\ 0 & 0 & c_3 \end{pmatrix} \quad B = \begin{pmatrix} b \\ 0 \\ 0 \end{pmatrix} \quad (13.23)$$

$$\xi(t) = k_1 e^{b_1 t}$$

$$u(t) = k_2 e^{b_2 t},$$

$$(a) = e^{\int_0^t AC \int_s^t \xi(\sigma) d\sigma} X_0 = V \text{diag}(e^{\lambda_i f(t)})_{i=1}^n V^{-1} X_0 \quad (13.21)$$

And:

$$\begin{aligned} (d) &= \left(\int_0^t e^{\int_s^t AC \int_s^t \xi(\sigma) d\sigma} u(s) ds \right) B \\ &= \int_0^t V \left(\text{diag} \left(e^{\lambda_i \int_s^t \xi(\sigma) d\sigma} \right)_{i=1}^n \right) V^{-1} u(s) ds B \\ &= V \int_0^t \left(\text{diag} \left(e^{\lambda_i \int_s^t \xi(\sigma) d\sigma} \right)_{i=1}^n \right) u(s) ds V^{-1} B \\ &= V \text{diag} \left(\int_0^t e^{\int_s^t \lambda_i \xi(\sigma) d\sigma} u(s) ds \right)_{i=1}^n V^{-1} B \end{aligned} \quad (13.22)$$

This finishes the calculation of the solution (13.17) to the n-pool model (13.16). However, we want readers to note that the calculation of the eigenvalues λ_i 's and eigenvectors v_i 's, which ultimately form the eigenvector matrix V of the carbon transfer coefficients matrix product AC , is far from trivial in most of the practical cases.

Mathematica Calculation For the Analytical Solution of a 3-pool Model

We consider a 3-pool model with:

and we use the mathematical software Mathematica to carry out the following calculations. The matrix product AC is:

$$AC = \begin{pmatrix} -c_1 & 0 & 0 \\ c_1 & -c_2 & 0 \\ 0 & c_2 & -c_3 \end{pmatrix}$$

The eigenvalues of AC are:

$$\lambda_1 = -c_1, \lambda_2 = -c_2, \lambda_3 = -c_3$$

and the eigenvector matrix of AC is:

$$V = \begin{pmatrix} \frac{(c_1 - c_2)(c_1 - c_3)}{c_1 c_2} & 0 & 0 \\ \frac{-c_1 + c_3}{c_2} & -1 + \frac{c_3}{c_2} & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

Let the initial value X_0 be:

$$X_0 = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

then term (a) given in (13.21) has three components:

$$d_3 = \frac{bc_1 c_2^2 \left[\frac{\left(-1 + \frac{c_3}{c_2} \right) \left(-1 + e^{-\frac{c_1(-1+e^{bt})k_1}}{b_1}} \right)}{c_1} - \frac{(c_1 - c_3) \left(-1 + e^{-\frac{c_2(-1+e^{bt})k_1}}{b_1}} \right)}{c_2^2} - \frac{\left(1 - \frac{c_1}{c_2} \right) \left(-1 + e^{-\frac{c_3(-1+e^{bt})k_1}}{b_1}} \right)}{c_3} \right]}{(c_1 - c_2)(c_1 - c_3)(c_2 - c_3)k_1} k_2$$

Even for a seemingly simple 3-pool model, the computation of the analytical solution of this model is already very complicated. For models with more pools, the calculation could be much more challenging for many cases.

$$a_1 = e^{-\frac{c_1(-1+e^{bt})k_1}{b_1}} x_1$$

$$a_2 = \frac{e^{-\frac{(c_1+c_2)(-1+e^{bt})k_1}{b_1}} \left(-c_1 e^{-\frac{c_2(-1+e^{bt})k_1}{b_1}} x_1 + e^{-\frac{c_1(-1+e^{bt})k_1}{b_1}} (-c_2 x_2 + c_1(x_1 + x_2)) \right)}{c_1 - c_2}$$

and a_3 has a much longer expression, and is omitted here. To facilitate the calculation of term (d) given in (13.22), we assume that $b_2 = b_1 = b$, then term (d) given in (13.22) has three components:

$$d_1 = \frac{b \left(1 - e^{-\frac{c_1(-1+e^{bt})k_1}{b_1}} \right) k_2}{c_1 k_1}$$

$$d_2 = \frac{b \left(c_1 - c_1 e^{-\frac{c_2(-1+e^{bt})k_1}{b_1}} + c_2 \left(-1 + e^{-\frac{c_1(-1+e^{bt})k_1}{b_1}} \right) \right) k_2}{(c_1 - c_2) c_2 k_1}$$

STABILITY

Instantaneous Steady State

The instantaneous steady-state of (13.1):

$$X' = \xi(t) ACX + Bu(t) \quad (13.24)$$

is defined as:

$$X_{ss}(t) = -\xi^{-1}(t)u(t)C^{-1}A^{-1}B(t) \quad (13.25)$$

The instantaneous steady state provides the $X(t)$ with zero rate of change at time t .

Instantaneous Steady State for a 3-pool Model

Let us re-visit the 3-pool model (13.23) given in the previous section. A simple calculation shows that the inverse of the carbon transfer coefficients matrices AC is:

$$\text{Inverse}(AC) = \begin{pmatrix} -1/c_1 & 0 & 0 \\ -1/c_2 & -1/c_2 & 0 \\ -1/c_3 & -1/c_3 & -1/c_3 \end{pmatrix}$$

and therefore the instantaneous steady state solution is:

$$X_{ss}(t) = -\xi^{-1}(t)u(t)\text{Inverse}(AC)B(t) = \begin{pmatrix} \frac{bk_2}{c_1k_1} \\ \frac{bk_2}{c_2k_1} \\ \frac{bk_2}{c_3k_1} \end{pmatrix}$$

The Mathematica computational results show that all the eigenvalues $(\lambda_i)_{i=1}^n$ of the matrix AC are strictly negative, therefore the instantaneous steady state solution (13.25) is asymptotically stable. We numerically investigate an example to verify that the solution $X(t)$ does converge to the instantaneous steady-state solution $X_{ss}(t)$ as $t \rightarrow \infty$. To measure the convergence, we need to study the following function:

$$\begin{aligned} \text{distance}(t) &= X(t) - X_{ss}(t) \\ &= (a) + (d) - X_{ss}(t) \end{aligned} \quad (13.26)$$

where (a) , (d) and $X_{ss}(t)$ are given in (13.21, 13.22, 13.25).

In this numerical investigation, we choose $c_1 = 1/10$, $c_2 = 3/10$, $c_3 = 2/10$, $b = 1$, $k_1 = 1$, $k_2 = 1$, $b_1 = 1$, and the initial value $X_0 = (5/10, 2/10, 3/10)$ in the 3-pool model (13.23), Figure 13.1 shows that the three components of the solution $X(t)$, the instantaneous steady-state solution $X_{ss}(t)$ and the distance given in (13.26).

An interesting question to ask is how the parameters affect the rate of convergence to the instantaneous steady state solution. This question will be left to the readers to explore.

Global Attractor

As the name suggests, a global attractor attracts all the solutions regardless of the initial profiles. We will introduce the definition of the global attractor, how to numerically estimate the global attractor, and a mathematical study of the convergence of the solutions to the global attractor as $t \rightarrow \infty$. In the end of this section, we will also give the condition when the global attractor and the instantaneous steady-state solution given in (13.25) converge to each other.

Consider a non-homogeneous nonautonomous system in the general form:

$$X'(t) = A(t)X(t) + b(t) \quad (13.27)$$

where $A(t)$ has the block matrix form:

$$A(t) = \begin{pmatrix} A_{11}(t) & 0 \\ A_{21}(t) & A_{22}(t) \end{pmatrix}$$

$$\begin{aligned} A_{11}(t) &\in \mathbb{R}^{d_1 \times d_1}, 0 \in \mathbb{R}^{d_1 \times d_2}, A_{21}(t) \in \mathbb{R}^{d_2 \times d_1}, \\ A_{22}(t) &\in \mathbb{R}^{d_2 \times d_2}, \text{ and } d_1 + d_2 = d. \end{aligned}$$

Assume that:

- 1) $A_{11}(t)$ is a lower triangular matrix
- 2) $\|b(t)\| \leq B$ for some $B > 0$ (i.e., $b(t)$ is bounded)
- 3) $\exists \delta < 0$, such that
 - $a_{ii}(t) \leq \delta$, for $i = 1, \dots, d_1$
 - $a_{ii}(t) < 0$, for $i = d_1 + 1, \dots, d$ (i.e., $A_{22}(t)$ has negative diagonal entries)
 - $|a_{ii}(t) + \delta| \geq \sum_{j \neq i, j > d_1} |a_{ij}(t)|$, for $i = d_1 + 1, \dots, d$, (i.e., $A_{22}(t)$ is diagonally dominant)

Then:

$$\|\Phi(t,s)\| \leq Ke^{\delta(t-s)} \quad (13.28)$$

for some $K > 0$ and $t > s$, where $\Phi(t,s)$ is the fundamental matrix of:

$$X'(t) = A(t)X(t)$$

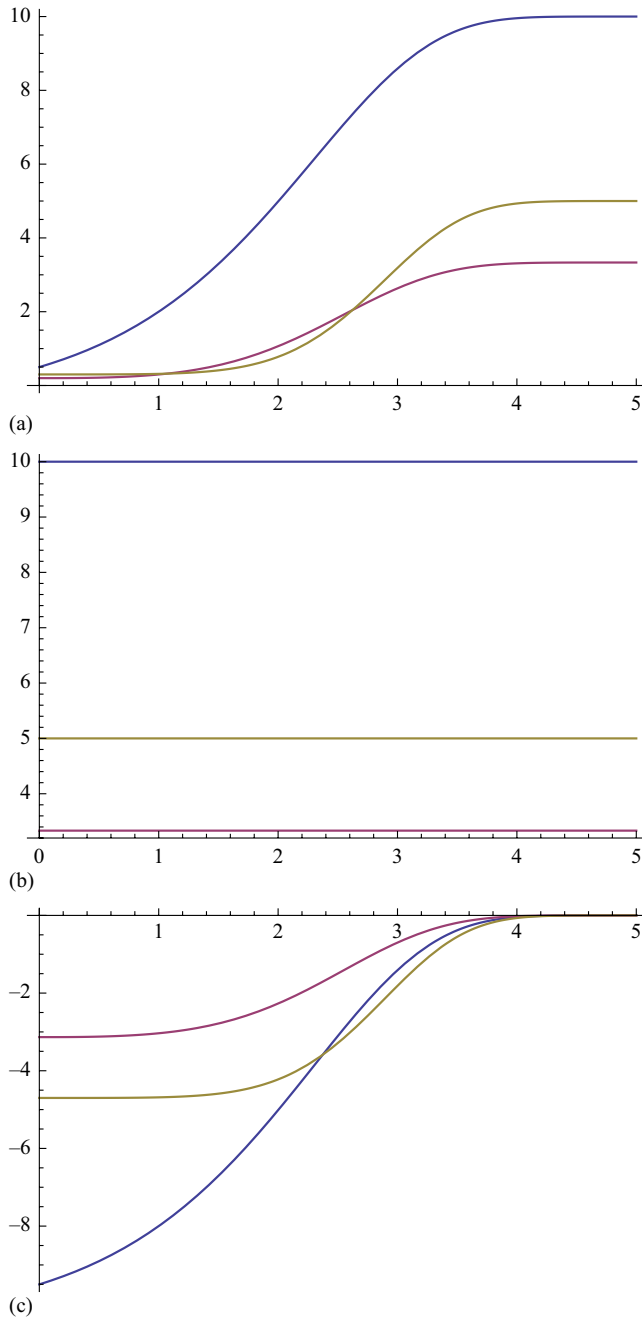


Figure 13.1. The three panels show (a) the solution $X(t)$, (b) the equilibrium solution $X_e(t)$ and (c) the distance versus time, for $0 \leq t \leq 5$. The blue, red, green curves represent the first, second and third components respectively.

i.e., the solution of the general nonautonomous equation (13.27) can be written as: Let:

$$X(t) = \Phi(t, s)X(s) + \int_s^t \Phi(t, \tau)b(\tau) d\tau \quad \mu(t) = \int_{-\infty}^t \Phi(t, s)b(s) ds \quad (13.29)$$

then $\mu(t)$ is the unique global attractor of the non-autonomous system (13.27).

Notice that based on (13.28, 13.29) and the assumption on boundedness of $b(t)$,

$$\begin{aligned}
\|\mu(t)\| &\leq \int_{t-T}^t \|\Phi(t,s)b(s)\| ds \\
&\quad + \int_{-\infty}^{t-T} \|\Phi(t,s)b(s)\| ds \\
&\leq \int_0^T \|\Phi(t,t-\sigma)b(t-\sigma)\| d\sigma \\
&\quad + \int_{-\infty}^{t-T} Ke^{\delta(t-s)} \|b(s)\| ds \\
&\leq \int_0^T \|\Phi(t,t-\sigma)b(t-\sigma)\| d\sigma \\
&\quad + Ke^{\delta T} / |\delta| \\
&\xrightarrow{T \rightarrow \infty} \int_0^T \|\Phi(t,t-\sigma)b(t-\sigma)\| d\sigma
\end{aligned} \tag{13.30}$$

Equation (13.30) gives a way to numerically estimate the global attractor $\mu(t)$.

All the solutions (regardless of the initial conditions) will converge to $\mu(t)$ as $t \rightarrow \infty$. This is because:

$$X(t) - \mu(t) = \Phi(t,s)(X(s) - \mu(s))$$

Therefore:

$$\begin{aligned}
\|X(t) - \mu(t)\| &\leq Ke^{\delta(t-s)} \|X(s) - \mu(s)\| \\
&\xrightarrow{t \rightarrow \infty} 0, \text{ since } \delta < 0
\end{aligned}$$

The Global Attractor of the N-Pool Model

Consider:

$$X' = \xi(t)ACX + Bu(t), \tag{13.31}$$

its global attractor is:

$$\begin{aligned}
\mu(t) &= \int_{-\infty}^t \Phi(t,s)Bu(s) ds \\
&= \int_{-\infty}^t e^{\int_s^t AC \int_s^\sigma \xi(\sigma) d\sigma} (-1) \xi(s) ACX_{ss}(s) ds \\
&= \left[e^{\int_s^t AC \int_s^\sigma \xi(\sigma) d\sigma} X_{ss}(s) \right]_{s=-\infty}^{s=t} - \int_{-\infty}^t e^{\int_s^t AC \int_s^\sigma \xi(\sigma) d\sigma} X'_{ss}(s) ds \\
&= X_{ss}(t) - \int_{-\infty}^t e^{\int_s^t AC \int_s^\sigma \xi(\sigma) d\sigma} X'_{ss}(s) ds
\end{aligned}$$

Therefore:

$$\begin{aligned}
\mu(t) - X_{ss}(t) &= - \int_{-\infty}^t e^{\int_s^t AC \int_s^\sigma \xi(\sigma) d\sigma} X'_{ss}(s) ds \\
&= - \int_{-\infty}^T e^{\int_s^t AC \int_s^\sigma \xi(\sigma) d\sigma} X'_{ss}(s) ds \\
&\quad - \int_{t-T}^t e^{\int_s^t AC \int_s^\sigma \xi(\sigma) d\sigma} X'_{ss}(s) ds
\end{aligned}$$

General Stability Statements

CASE 1

Consider a homogeneous system of ODEs with constant coefficient matrix:

$$X' = AX.$$

Let $\lambda_j, j = 1, \dots, n$ counting multiplicity be the eigenvalues of matrix A , and let \tilde{X} be the equilibrium state.

- if $\text{Re}\lambda_j < 0$ for all $j = 1, \dots, n$, then \tilde{X} is asymptotically stable.
- if $\text{Re}\lambda_j \leq 0$ for all $j = 1, \dots, n$, and the eigenvalues with real part equals to zero are simple, then \tilde{X} is stable.
- if there exists λ_j , such that $\text{Re}\lambda_j = 0$, but is not simple, then \tilde{X} is unstable.
- if there exists λ_j , such that $\text{Re}\lambda_j > 0$, then \tilde{X} is unstable.

CASE 2

Consider a homogeneous system of nonautonomous ODEs:

$$X' = AX + B(t)X$$

where $B(t)$ is a continuous function. Let λ_j , $j = 1, \dots, n$ counting multiplicity be the eigenvalues of matrix A , and let \tilde{X} be the equilibrium state.

- if $\operatorname{Re}\lambda_j < 0$ for all $j = 1, \dots, n$, and $\lim_{t \rightarrow \infty} \|B(t)\| = 0$, then \tilde{X} is asymptotically stable.
- if $\operatorname{Re}\lambda_j \leq 0$ for all $j = 1, \dots, n$, and the eigenvalues with real part equals to zero are simple, and $\int_{t_0}^{\infty} \|B(t)\| < \infty$ then \tilde{X} is stable.
- if there exists λ_j , such that $\operatorname{Re}\lambda_j > 0$, and $\lim_{t \rightarrow \infty} \|B(t)\| = 0$, then \tilde{X} is unstable.

CASE 3

Consider the following general system:

$$X' = AX + B(t)X + f(t, X)$$

where $B(t)$ is a continuous function. Let X_0 be an equilibrium solution. If:

$$\lim_{t \rightarrow \infty} \|B(t)\| = 0$$

- 1) $f(t, X)$ is continuous in t , and has continuous partial derivative in x :

$$\lim_{x \rightarrow 0} \frac{\|f(t, x)\|}{x} = 0 \quad \text{uniformly in } t$$

then:

- If $\operatorname{Re}\lambda_j < 0$, where λ_j are the eigenvalues of A , then $X = X_0$ is asymptotically stable.
- If there exists an eigenvalue λ_j , such that $\operatorname{Re}\lambda_j > 0$, then $X = X_0$ is unstable.

SUGGESTED READING

C. H. Edwards, D. E. Penney, and D. T. Calvis. 2014. *Differential Equations and Boundary Value Problems (5/e)*, Pearson, ISBN 9780321796981.

QUIZZES

Find the general solution of the following ordinary differential equations and system of ODEs.

$$y' + y = 2, y(0) = 0$$

$$xy' + 2y = 3x, y(1) = 5$$

$$2xy' - 3y = 9x^3$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 0 & 8 \\ -2 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 10 & -7 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} -3 & 2 \\ -3 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 3 \\ -2 \end{pmatrix}$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 9 & 1 \\ -8 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2 \\ t \end{pmatrix} e^t$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1 & -5 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2 \sin t \\ -3 \cos t \end{pmatrix}$$

Find the critical point(s) of the following systems, and determine whether it is asymptotically stable, stable, or unstable.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 5 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -8 \\ 10 \end{pmatrix}$$

$$\begin{cases} x' = x - y \\ y' = x^2 - y \end{cases}$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 4 & -5 \\ 5 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 3 \\ 6 \end{pmatrix}$$

$$\begin{cases} x' = xy - 2 \\ y' = x - 2y \end{cases}$$



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER FOURTEEN

Semi-Analytic Spin-Up (SASU) of Coupled Carbon-Nitrogen Cycle Models

Xingjie Lu

SunYat-sen University, Guangzhou, China

Jianyang Xia

East China Normal University, Shanghai, China

CONTENTS

What Is Spin-Up? / 115
Development of Spin-Up Approaches / 116
Semi-Analytic Spin-Up / 117
The Procedure of Semi-Analytic Spin-Up in CABLE / 119
Computational Efficiency / 120
Suggested Reading / 121
Quizzes / 121

This chapter introduces the concept of spin-up in land biogeochemical models. Spin-up is a standard initialization procedure, which is commonly proposed in protocols of model intercomparison projects. Conventional spin-up, repeatedly running the model with recursive environmental forcing to achieve a steady state, is very time consuming. Semi-analytic spin-up (SASU) enabled by the matrix representation significantly improves the computational efficiency of spin-up for land biogeochemical models. The implementation of SASU in the Community Atmosphere and Biosphere Land Exchange (CABLE) model is demonstrated. The SASU method has been proved to save about 92.4% and 86.6% of the computational time for spin-up of the global carbon-only and coupled carbon-nitrogen models.

WHAT IS SPIN-UP?

Spin-up is a standard initialization procedure used by land surface models, atmosphere models,

ocean models, or earth system models. Generally speaking, spin-up provides initial values of state variables, which are very critical to the definite solution. For instance, you might have heard of the so-called butterfly effect in weather/climate forecasting: slightly different initial values might cause different or even completely opposite results.

The initialization procedure is an essential step in simulation of land biogeochemical models. Model intercomparison projects are often carried out in order to elaborate understanding of carbon cycle dynamics and characterize intermodel uncertainty. These exercises require a standardized protocol in order to ensure that results from different participating models can be compared on an equal basis. Model intercomparison projects have for instance been undertaken to assess the anthropogenic impact on land biogeochemical cycles. The industrial revolution is commonly believed to have triggered much of the current anthropogenic impact such as increasing fossil fuel emissions, land use change, and increasing nitrogen deposition.

It represents the starting point of a period when humans have played a more important role than ever before in impacting the earth's geology and ecosystems. Land biogeochemical cycles are experiencing an important transition period unique in the history of the earth. So, in the protocol of model intercomparison projects, the simulation of the transition period in biogeochemical cycles from industrial revolution to present-day is often called the historical simulation. One of the primary goals for the historical simulation is to hindcast the ecosystem changes in response to anthropogenic activity since the onset of industrialization around the 19th century. Initial values matter to the growing trend of ecosystem carbon storage. However, determining these initial values can be a tricky issue when we have little information on the initial state. Direct observations for data such as soil temperature, moisture, carbon, and nitrogen storage are not generally available until very recent times. Thus, in these protocols, *steady state* is commonly used to determine initial values for the historical simulation.

Steady state is achievable by most land biogeochemical models as long as we run the model long enough. Chapter 1 provided the theoretical foundation; that is, for given forcing, ecosystem carbon storage tends to grow towards the same steady state regardless of the initial carbon storage once a model structure is defined and parameter values are given. The unique steady state does not rely on the initial value as the land carbon cycle converges to a steady state that is determined by carbon input and residence time. Thus, the procedure of generating steady state can be standardized across different models.

Conventionally, the steady state of carbon storage can be achieved by repeatedly running the model with recursive environmental forcing. Most model intercomparison projects include a protocol to achieve the steady state. For examples, Multiscale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP) protocol requires that the spin-up uses a random climate driver data package, constant 1801 land cover, constant pre-industrial atmospheric CO₂ concentration, and constant nitrogen deposition to force the model. Under a constant atmospheric CO₂ concentration, constant nitrogen inputs and random climate forcing, ecosystem carbon storage will ultimately reach a quasi-steady state. As a rule, to detect whether steady state has been achieved,

MsTMIP specifies that the 100-yr mean interannual change in total ecosystem carbon stocks for consecutive years must be below 1 gC m⁻² yr⁻¹ for 95% of grid cells. The carbon stocks accumulated at steady state are used as the initial values in the transient historical simulation from 1801 to 2010. This approach, which takes advantage of a model's native dynamics to drive the system towards the steady state, is called native dynamics spin-up (ND), and is the most widely used by current models.

DEVELOPMENT OF SPIN-UP APPROACHES

Several other spin-up approaches have also been developed. The motivation to develop alternative spin-up approaches to ND in land biogeochemical model is to improve the computational efficiency of the spin-up. For most models, the ND approach usually takes hundreds to thousands of simulation years to reach the steady state, which is computationally very expensive. Even this poor level of efficiency is becoming much lower as ongoing development makes models more and more complex. For example, vertically-resolved soil carbon modules are developed to represent carbon vertical mixing and the soil freeze-thaw impact on soil decomposition in permafrost regions. While accounting for these additional processes better represents the nature of the heterogeneity, the extremely slow turnover rates for deep soil carbon in permafrost regions substantially slow down the spin-up. A more efficient spin-up approach is urgently needed.

Spin-up approaches that modify ND but allow the state variables to naturally accumulate to the steady state without any arbitrary adjustments to the state variables themselves are called *ad hoc* approaches. A punctuated nitrogen addition approach (PN) manipulates the rate of new mineral nitrogen addition (Thornton et al., 2002; Thornton and Rosenbloom, 2005). PN is proposed as a solution to reduce spin-up times based on the observation that the time taken by a coupled carbon-nitrogen biogeochemical model to reach a steady state under ND is mainly controlled by the new nitrogen addition. PN is faster than ND but the higher nitrogen inputs also result in larger carbon pools at steady state. Therefore, an additional ND needs to be carried out after PN, so that PN results conform to ND results. The accelerated decomposition approach

(AD) was firstly implemented into the Biome-BGC model (Thornton and Rosenbloom, 2005). AD arbitrarily increases the decomposition rate during spin-up to allow the system to approach a steady state faster. AD is informed by the experience that scaling of litter and soil carbon decomposition produce a new steady state, which is linearly related to the scaling constant (Thornton and Rosenbloom, 2005). AD is much faster than ND, but it also generates lower steady state carbon and nitrogen pools, because interactions with the soil passive pool become much stronger. To achieve the final steady state, additional ND after AD is required, which is called the post-AD process. The post-AD process is sometime much longer than AD itself. AD has been implemented into CLM version 4 (Koven et al., 2013) and the following versions (CLM4.5 and CLM5) with vertically-resolved soil carbon.

In contrast to *ad hoc* approaches, spin-up approaches allowing arbitrary change in state variables are called generalized optimization approaches. The semi-analytic spin-up approach (SASU) updates the state variables with steady state estimates from an analytic solution of the model. The SASU approach was firstly implemented in the Community Atmosphere-Biosphere-Land Exchange (CABLE) model for carbon and nitrogen cycle spin-up (Xia et al., 2012). For carbon-cycle-only simulations, SASU saves well over 90% of the computational cost for CABLE site-level and global simulations. For carbon-nitrogen coupled simulation, SASU saves over 80% of the computational cost. The SASU approach has the additional advantage that it enables parameter sensitivity analysis of the biogeochemical parameters (Huang et al., 2018b), as well as data assimilation for the estimation of these parameters (Tao et al., 2020), which was prohibitive due to large computational cost before SASU was proposed.

Matrix representations have been derived for many land biogeochemical models, in part to enable them to use the SASU approach for spin-up. In the last decade matrix representations have been published for TECO (Jiang et al., 2017; Luo et al., 2017), CABLE (Xia et al., 2013), LPJ-GUESS (Ahlström et al., 2015), ORCHIDEE (Huang et al., 2018b), CLM3.5 (Hararuk et al., 2015; Hararuk et al., 2014; Rafique et al., 2016), CLM4 (Rafique et al., 2017), CLM4.5 (Huang et al., 2018a), and CLM5 (Lu et al., 2020).

The accelerated spin-up (ASU) approach is implemented in the Terrestrial Ecosystem Model (TEM). ASU is similar to SASU and also belongs to the generalized optimization approach, but is applied to the traditional as opposed to the matrix form of the model. ASU numerically solves the model's steady state considering the seasonal cycle of carbon and nitrogen storage. Qu et al. (2018) found that ASU saved 90% and 85% of computational costs for TEM site-level and North American regional simulations, respectively.

SEMI-ANALYTIC SPIN-UP

The SASU approach is built upon the mathematical foundation of biogeochemical cycles in terrestrial ecosystems, introduced in Chapter 1. The biogeochemical cycling of carbon in an ecosystem is usually initiated with plant photosynthesis, which transfers CO_2 from the atmosphere into an ecosystem. The carbon assimilated through photosynthesis is partitioned into compartments of plant biomass, such as leaf, root, and woody biomass. Plant biomass lost through phenological turnover, mortality or damage by herbivores etc. becomes litter, entering metabolic, structural, and coarse woody debris (CWD) litter pools. The litter carbon is partially released to the atmosphere as CO_2 respired by decomposing microbes, and partially converted to soil organic matter (SOM) in fast, slow, and passive pools (Figure 14.1).

The mean carbon residence time varies greatly among different pools, from several months in leaves and roots to hundreds or thousands of years in some woody tissues and SOM (Torn et al., 1997). These carbon processes in terrestrial ecosystems can be mathematically expressed by the following carbon balance equation in a matrix form (Luo et al., 2017; Luo et al., 2003):

$$\frac{dX(t)}{dt} = BU(t) + A\xi(t)KX(t) \quad (14.1)$$

Taking the CABLE model (Wang et al., 2011) as one example, $X(t) = (X_1(t), X_2(t), \dots, X_9(t))^T$ is a 9×1 vector describing pool sizes for the nine carbon pools leaf, wood, root, metabolic litter, structural litter, CWD, fast SOM, slow SOM, and passive SOM, respectively. A and K are then 9×9 matrices given by:

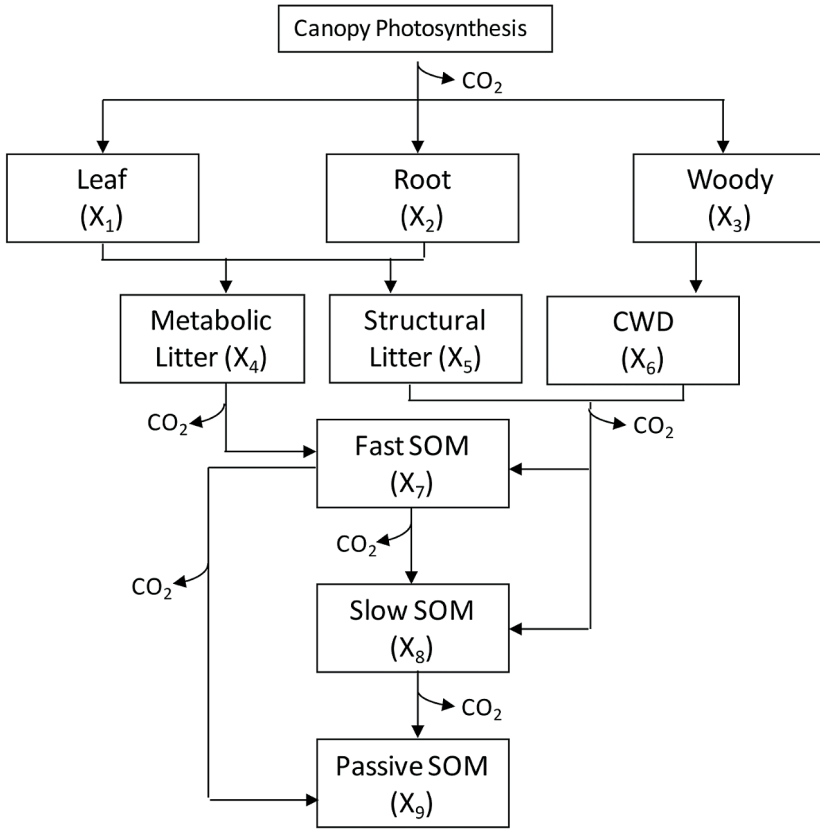


Figure 14.1. Diagram of the carbon processes of CABLE model on which model Equations 14.1–14.3 are based. SOM stands for soil organic matter.

$$A = \begin{bmatrix}
 -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 a_{41} & a_{42} & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 a_{51} & a_{52} & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & a_{74} & a_{75} & a_{76} & -1 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & a_{85} & a_{86} & a_{87} & -1 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & a_{97} & a_{98} & -1 & 0
 \end{bmatrix} \quad (14.2)$$

$$K = \text{diag}(k) \quad (14.3)$$

where A denotes the carbon transfer matrix, in which a_{ij} represents the fraction of carbon transfer from pool j to i . The $\text{diag}(k)$ is a 9×9 diagonal matrix with diagonal entries given by vector $k = (k_1, k_2, \dots, k_9)^T$, components k_j ($j = 1, 2, \dots, 9$) quantify the fraction of carbon left from pool X_j ($j = 1, 2, \dots, 9$) after each time step. $\xi(t)$ is an

environmental scalar accounting for effects of soil type, temperature, and moisture on carbon decomposition. $B = (b_1, b_2, b_3, 0, \dots, 0)^T$ represents the partitioning coefficients of the photosynthetically fixed carbon into different pools. $U(t)$ is the input of carbon via plant photosynthesis. Other models may have a slightly different set of biomass, litter and soil carbon pools, but in general, Equation 14.1 can adequately summarize C cycle processes in most land models (Luo et al., 2015; Luo et al., 2017).

Equation 14.1 cannot be directly solved to obtain the steady-state values of carbon pools because the environmental scalar $\xi(t)$, ecosystem carbon influx $U(t)$, possibly matrix A , and vector B vary with time and driving variables. Since carbon influx involves fast processes, its steady-state value U_{ss} can be obtained from a short ND spin-up, which we call initial spin-up. The initial spin-up uses recursive meteorological forcing to drive modeled carbon and nitrogen dynamics. Thus, it is

possible to calculate averaged values of the environmental scalar ($\bar{\xi}$), the carbon transfer (\bar{A}), and partitioning (\bar{B}) coefficients within one cycling period of the meteorological variables. With carbon input at steady state U_{ss} and the mean values of the time-varying variables ($\bar{\xi}$, \bar{A} , and \bar{B}), we can analytically calculate the steady-state carbon pool sizes $X_{ss,C}$ by letting the right-hand side of Equation 14.1 equal zero as:

$$X_{ss,C} = -(\bar{\xi} \bar{A} \bar{K})^{-1} \bar{B} U_{ss} \quad (14.4)$$

We can divide $X_{ss,C}$ by C/N ratio in each pool to obtain steady-state nitrogen pool sizes $X_{ss,N}$. Equation 14.4 stands for the steady state in the absence of soil nitrogen feedback to NPP. Final steady state will be achieved by the full procedure, which will be described in the next section.

THE PROCEDURE OF SEMI-ANALYTIC SPIN-UP IN CABLE

The procedure of SASU implementation in a land surface model is demonstrated by the case of CABLE. In general, the procedure consists of four preparation and three execution steps: (1) development of flow diagram; (2) organization of matrix equation; (3) identification of the time-varying elements; (4) coding the analytic solution; and three modeling steps: (5) initial spin-up; (6) analytic solution of steady state; and (7) final ND spin-up. In more detail, these seven steps entail (Figure 14.2):

1. Developing a flow diagram reflecting the linkages between ecosystem carbon pools in the target model structure, as shown in Figure 14.1 (also see lectures and practice in unit 1).
2. Organizing CABLE into a matrix equation by encoding the linkage between carbon pools and fluxes in the carbon transfer matrices A and K , and plant carbon partitioning coefficients in vector B . Values for non-zero elements in vector B , and matrices A and K were assigned based on parameter values in the original CABLE model.
3. Figuring out how the time-varying elements of A , B and ξ in Equation 14.1 are determined in the model.
4. Coding the analytic solution of the biogeochemical steady state in CABLE, which includes two steps: (a) creating new variables to store the mean values of the time-varying parameters; and (b) creating equations to calculate the analytical solutions of each pool according to the structures of matrix A , K and vector B .
5. Performing the initial spin-up by running the model using repeated meteorological forcing until NPP (or all plant pools) reach steady state (U_{ss}). Run the model until the mean change in NPP over each loop is smaller than 0.01% per year. Meanwhile, the values of all the time-varying parameters in Equation 14.3 are updated by the mean values from initial spin-up.

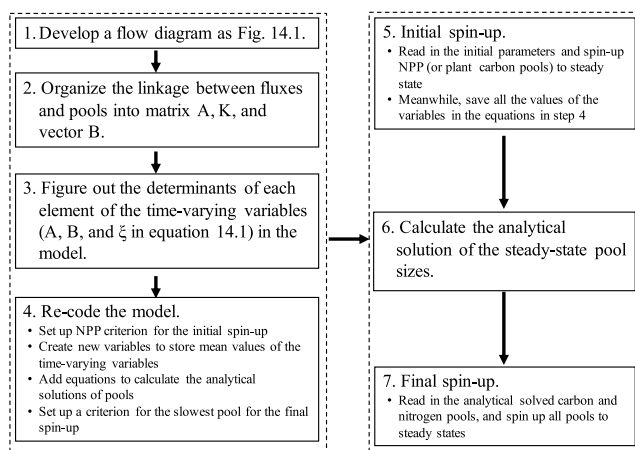


Figure 14.2. Procedure of the semi-analytical spin-up (SASU). (Reproduced from Xia et al. (2012)).

6. Calculating the analytical solution of the steady states of carbon and nitrogen pools. The steady-state carbon pools are solved by setting carbon influx equal to efflux for each pool (Equation 14.4). Nitrogen pools are obtained by dividing the steady-state carbon pools by their C/N ratios at the end of the initial spin-up.
7. Performing the final spin-up by using the analytically solved carbon and nitrogen pools as initial values until the steady-state criterion for the soil carbon pools is met. Here our steady-state criterion will be that the change in the passive soil carbon pool (ΔC_{pass}) within each simulation year is smaller than $0.5 \text{ gC m}^{-2} \text{ yr}^{-1}$. This criterion was chosen following a recommendation of Thornton and Rosenbloom (2005). According to the difference in turnover rate, a slower pool needs a longer time to reach steady state during the spin-up. When the criterion for recognizing that steady state has been achieved is small enough the final spin-up is determined by the dynamic of the slowest carbon pool.

COMPUTATIONAL EFFICIENCY

In our example using CABLE, the traditional spin-up method spends 2780 and 5099 simulation years for carbon-only and coupled carbon-nitrogen simulation, respectively, before the change in the slowest carbon pool meets the steady-state criterion ($\Delta C_{\text{pass}} < 0.5 \text{ gC m}^{-2} \text{ yr}^{-1}$; Figure 14.3). After implementing SASU, the initial spin-up takes 200 simulation years to achieve steady states of plant carbon pools for both the carbon-only model and the coupled carbon-nitrogen model. This allows the sizes of all SOM pools to be calculated analytically after this initial spin-up (step 6 above). With the SASU method, all carbon pools in the carbon-only model reach steady states after analytical calculation without a need for any final spin-up (step 7 above; black arrow in Figure 14.3). In the coupled carbon-nitrogen model, SASU needs another 483 simulation years to obtain the steady states of all pools (gray arrow in Figure 14.3) after the analytical calculation. This means that the SASU method saved about 92.4% and 86.6% of the computational time for spin-up of the global carbon-only and coupled carbon-nitrogen models to

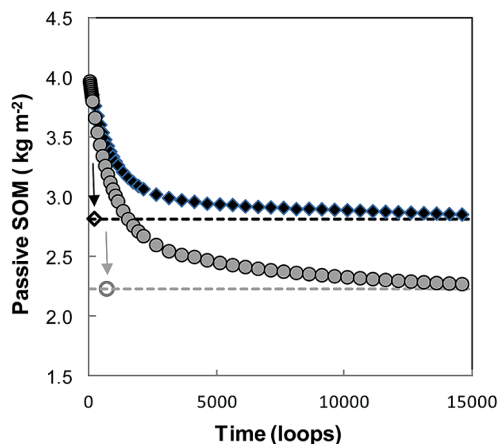


Figure 14.3. Dynamics of global mean passive soil carbon pool in carbon-only (filled black diamond) and coupled carbon-nitrogen (filled gray circle) simulations with traditional method and the SASU framework (dotted lines) using the CABLE model. The arrows and open symbols show the times and values estimated with the SASU method for carbon-only (black) and coupled carbon-nitrogen coupled (gray) simulations. (Adapted from Xia et al. (2012)).

steady states, respectively, in the case of the CABLE model (Figure 14.3).

With the traditional spin-up method, the passive SOM pool continued to decrease after it reached the steady-state criterion (Figure 14.3). Because of the slow turnover rate of this pool, thousands of additional simulation years are needed for the traditional method to reach a steady state of the passive SOM pool. In contrast, the analytical solution obtained by the SASU method allowed this steady state to be accurately predicted after only 200 simulation years (Figure 14.3).

The results of this exercise using CABLE show that SASU can reduce the bottle-neck of biogeochemical model spin-up by around 90%. That means spin-up with the SASU method can be around ten times as fast as the traditional method. The computational efficiency with the SASU method is higher than the accelerated decomposition method identified by Thornton and Rosenbloom (2005) for site-level spin-up in an evergreen needle-leaf forest. A similar analytic spin-up method developed by Lardy et al. (2011) accelerates spin-up with a pasture model (PaSim) by up to 20 times as well.

The SASU method can be easily implemented for biogeochemical models at site, regional, and global scales once a biogeochemical model is converted to a matrix model to enable analytical

calculation of steady states of carbon and nutrient pools together with initial and final spin-ups.

SUGGESTED READING

Xia J, Luo Y, Wang Y-P, Weng E, Hararuk O. (2012). A semi-analytical solution to accelerate spin-up of a coupled carbon and nitrogen land model to steady state. *Geoscientific Model Development*, **5**:1259–1271.

QUIZZES

Select one option from the given answers

1. Which statement for the spin-up is **NOT** true?
 - a. Spin-up provides an initial state for model simulation.
 - b. Spin-up uses atmosphere CO₂ concentration at pre-industrial level.
 - c. Spin-up can speed up the historical simulation.
 - d. Spin-up in land biogeochemical models aims for the steady state.
2. The motivation to develop different spin-up approaches in land biogeochemical model is to
 - a. improve the simulation results.
 - b. improve the computational efficiency.
 - c. improve model structure.
 - d. improve model diagnostic capacity.
3. Semi-analytic spin-up (SASU) in CABLE calculates the steady state of carbon storage by setting
 - a. carbon storage to zero.
 - b. carbon influx to zero.
 - c. carbon influx equal to efflux.
 - d. carbon efflux to zero.
4. Which statement about CABLE spin-up efficiency is **NOT** correct.
 - a. The CABLE spin-up efficiency for the carbon-only model is higher than for the carbon-nitrogen coupled model.
 - b. SASU can be up to ten times as fast as traditional spin-up.
 - c. In the coupled carbon-nitrogen model, SASU does not need the final spin-up after the analytical calculation.
 - d. With the traditional spin-up method, the passive SOM pool continues to decrease after it reaches the steady-state criterion.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER FIFTEEN

Time Characteristics of Compartmental Systems

Carlos A. Sierra

Max Planck Institute for Biogeochemistry, Jena, Germany

CONTENTS

Introduction /	123
Age and Transit Time Distributions for Autonomous Systems in Equilibrium /	124
Age and Transit Time Distributions for Nonautonomous Systems /	125
Age Distributions /	126
Transit Time Distributions /	126
Final Remarks /	126
Suggested Reading /	127
Quizzes /	127

To understand carbon flows through terrestrial ecosystems, it is very important to use metrics to quantify the time carbon spends in the entire system and in particular compartments. In this chapter, we introduce the concepts of age and transit time as two fundamental metrics that characterize the speed at which carbon flows through ecosystems. Age is defined as the time carbon atoms spend in an ecosystem, from when they enter through photosynthesis until they are observed in a particular compartment. Transit time is defined as the time carbon atoms require to pass through the entire ecosystem, from the time they enter through photosynthesis until they are lost in gas, liquid, or solid form. We review here mathematical formulas for computing age and transit time in compartmental systems, distinguishing between formulas for autonomous systems in equilibrium and nonautonomous systems moving along a known trajectory.

INTRODUCTION

One of the advantages of representing models in the compact form of compartmental systems is that we can derive system-level diagnostics that help to

better understand system dynamics. Differences in process representations, parameterizations, or size of compartments required to represent a system, can be compared using simple aggregated metrics at the level of the entire system.

Two important system-level diagnostics for describing compartmental systems are the concepts of system age and transit (residence) time (Bolin and Rodhe 1973; Sierra et al. 2017). We define *system age* as the age of all atoms or particles inside the system, from the time they entered t_e until the time of observation t . *Transit time* is defined as the average time required for atoms or particles to traverse the system from their arrival time until they leave in the output flux. In other words, system age characterizes the age structure of all the atoms or particles in the system, while transit time characterizes the age structure of all atoms or particles in the output flux (Figure 15.1).

It is also possible to characterize the age structure of the atoms or particles inside each pool or compartment. We define *pool age* as the time elapsed since the atoms or particles entered the system until the time of observation t inside a pool i (Figure 15.1). Therefore, the system age is the aggregated result of the pool ages for all pools.

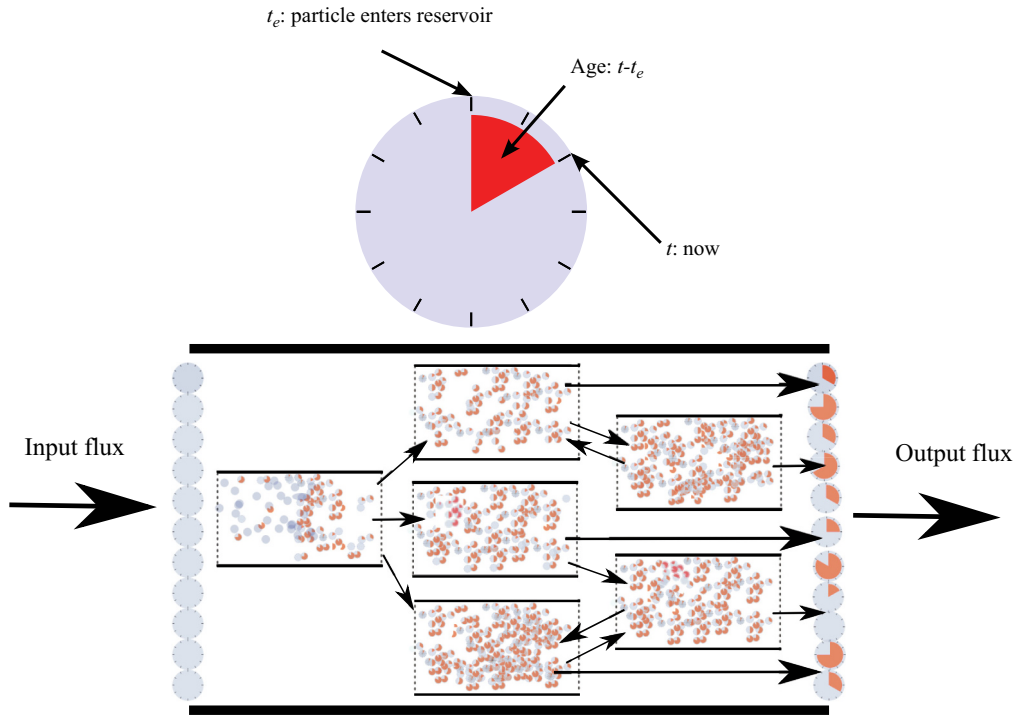


Figure 15.1. Graphical representation of the concepts of system age, transit time, and pool age. Mass entering a compartmental system can be conceptualized as being composed of small particles or atoms, each of them with a 'clock' that measures the time they have been in the system since they entered. All particles in the input fluxes have an age of zero. If we collect all particles inside the system at any given time and organize this information as a distribution of ages, we obtain the system age distribution of particles inside the system. If we collect the particles inside a specific pool and organize particles according to their age, we obtain the pool age distribution. Collecting particles in the output flux and organizing this information as a distribution of ages provides the transit time distribution. Figure extracted from Sierra et al. (2017).

Since the age and transit time concepts are defined for all individual atoms or particles inside a system, we can also think about them in terms of distributions that quantify the proportional distribution of the mass in age classes. Therefore, these distributions can be characterized by statistics such as the mean, standard deviation, and quantiles such as the median.

In the following sections, we will introduce mathematical formulas to quantify age and transit time distributions for two separate cases, (1) autonomous systems in equilibrium and (2) non-autonomous systems. Note that we will not be making a distinction between linear and nonlinear, because for case (1), linear and nonlinear systems in equilibrium can be treated similarly since the vector of states does not change once the equilibrium is reached and the system behaves similarly as a linear system. For case (2), the formulas rely on a linearization of the specific trajectory of a nonlinear system, therefore we will first introduce

the linearization strategy and then provide the formulas for the linear nonautonomous case.

AGE AND TRANSIT TIME DISTRIBUTIONS FOR AUTONOMOUS SYSTEMS IN EQUILIBRIUM

The derivation of the formulas for age and transit time distribution of linear autonomous systems in equilibrium was originally introduced in Metzler and Sierra (2018). For their derivation, we were able to show that linear compartmental systems are analogous to absorbing continuous-time Markov chains. This means that linear compartmental systems can also be interpreted in a stochastic sense, with the deterministic system of differential equations representing the macroscopic behavior of entire masses, and the absorbing Markov chains representing the stochastic behavior of individual atoms of particles with respect to their age. For details about the stochastic process and derivation

of formulas, interested readers can refer to Metzler and Sierra (2018) for additional details.

Let's consider linear autonomous systems introduced in Chapter 7, of the form of equation 7.2, with an equilibrium point given by equation 7.4. Let's also consider the 1-norm of a vector, defined as $\|\mathbf{v}\|_1 = \|\mathbf{v}\| := \sum_{i=1}^n |v_i|$, which is simply the sum of

all the elements in the vector. We say that the random variable age a that measures age of atoms or particles in the system is distributed according to a Phase-Type (PH) distribution of the form:

$$f(a) = \mathbf{z}^\top e^{a\mathbf{B}} \boldsymbol{\eta} = \mathbf{z}^\top e^{a\mathbf{B}} \frac{\mathbf{x}^*}{\|\mathbf{x}^*\|}, \quad a \geq 0.$$

Note that this density distribution is composed of three terms: the vector of fractional release coefficients, the matrix exponential of the compartmental matrix evaluated at each value of age, and the proportional distribution of mass at steady state. Since the fractional release coefficients can be computed directly from \mathbf{B} , we can say that the system age distribution follows a Phase Type distribution with two parameters: the probability vector of mass at steady state, and the transition rate matrix generated by the compartmental matrix. This can be abbreviated as $a \sim \text{PH}(\boldsymbol{\eta}, \mathbf{B})$.

The mean or expected value \mathbb{E} of the system age distribution can be obtained as:

$$\mathbb{E}[a] = -\mathbf{1}^\top \mathbf{B}^{-1} \boldsymbol{\eta} = \frac{\|\mathbf{B}^{-1} \mathbf{x}^*\|}{\|\mathbf{x}^*\|},$$

where $\mathbf{1}$ is a vector containing ones, and \top is the transpose operator.

To obtain the pool-age density distribution, we define first a diagonal matrix with the steady-state values for each compartment as $\mathbf{X}^* := \text{diag}(x_1^*, \dots, x_n^*)$. The vector-valued function that returns the age distribution for each pool is then given by:

$$\mathbf{f}(a) = (\mathbf{X}^*)^{-1} e^{a\mathbf{B}} \mathbf{u}, \quad a \geq 0,$$

and the mean age for each pool:

$$\mathbb{E}[\mathbf{a}] = -(\mathbf{X}^*)^{-1} \mathbf{B}^{-1} \mathbf{x}^*.$$

The density distribution of the random variable transit time τ is also Phase-Type distributed, with the probability vector given by the proportional distribution of the input flux $\boldsymbol{\beta}$, and the compartmental matrix as the transition rate matrix; i.e. $\tau \sim \text{PH}(\boldsymbol{\beta}, \mathbf{B})$. It can be obtained as:

$$f(\tau) = \mathbf{z}^\top e^{\tau \mathbf{B}} \boldsymbol{\beta} = \mathbf{z}^\top e^{\tau \mathbf{B}} \frac{\mathbf{u}}{\|\mathbf{u}\|}, \quad \tau \geq 0,$$

with mean transit time given by:

$$\mathbb{E}[\tau] = \|\mathbf{B}^{-1} \boldsymbol{\beta}\| = \frac{\|\mathbf{x}^*\|}{\|\mathbf{u}\|}.$$

Notice that the mean transit time is given by the ratio between the total mass at steady state and the total input flux.

AGE AND TRANSIT TIME DISTRIBUTIONS FOR NONAUTONOMOUS SYSTEMS

We consider now the nonlinear nonautonomous compartmental system introduced in Chapter 7 of the form of equation 7.9, for which we can always find a unique numerical solution of the form $\mathbf{x}(t, t_0, \mathbf{x}_0)$. To obtain time-dependent age and transit time distributions for this system, we will use the known solution to construct an equivalent linear nonautonomous system with the exact same solution. Details about the approach are presented in Metzler, Müller, and Sierra (2018).

Plugging in the known solution $\mathbf{x}(t) = \mathbf{x}(t, t_0, \mathbf{x}_0)$ into a new linear version of the system, we can define a new vector of inputs as $\tilde{\mathbf{u}}(t) := \mathbf{u}(\mathbf{x}(t), t)$, and a new compartmental matrix as $\tilde{\mathbf{B}}(t) := \mathbf{B}(\mathbf{x}(t), t)$. Then, we obtain a linear nonautonomous compartmental system of the form:

$$\dot{\mathbf{y}}(t) = \tilde{\mathbf{u}}(t) + \tilde{\mathbf{B}}(t) \cdot \mathbf{y}(t), \quad t > t_0, \quad \mathbf{y}(t_0) = \mathbf{x}_0,$$

which has a unique solution $\mathbf{y}(t, t_0, \mathbf{y}_0)$. Since we assume that the original nonlinear system of equation 7.9 also has a unique solution, both solutions must be identical; i.e. $\mathbf{y}(t, t_0, \mathbf{y}_0) = \mathbf{x}(t, t_0, \mathbf{x}_0)$. We can then use the solution of a nonlinear nonautonomous system to construct an equivalent linear nonautonomous compartmental system of the general form of equation 7.7, which has a general solution given by equation 7.8.

Age Distributions

We assume now that the initial content \mathbf{x}_0 has an initial age distribution $\mathbf{f}_0(a)$ such that $\mathbf{x}_0 = \int_0^{\infty} \mathbf{f}_0(a) da$.

This initial age distribution is then perturbed by the time-dependent mass inputs and process rates of the system, generating a time-dependent age distribution of the form

$$\mathbf{f}(a,t) = \mathbf{g}(a,t) + \mathbf{h}(a,t),$$

where the term $\mathbf{g}(a,t)$ is the time evolution of the age distribution of the initial mass in the system, and $\mathbf{h}(a,t)$ is the time evolution of the age distribution of mass that enters the system after t_0 .

The nonautonomous age distribution of the initial mass is given by:

$$\mathbf{g}(a,t) = I_{[t-t_0,\infty)}(a) \cdot \Phi(t,t_0) \cdot \mathbf{f}_0(a - (t - t_0))$$

where the indicator function $I_S(a)$ of a set S equals 1 if $a \in S$, or zero otherwise. The state transition operator $\Phi(t,t_0)$ is defined as in Chapter 7.

The nonautonomous age distribution of the mass that enters the system after t_0 is given by:

$$\mathbf{h}(a,t) = I_{[0,t-t_0)}(a) \cdot \Phi(t,t-a) \cdot \mathbf{u}(t-a)$$

To obtain the age distribution of the entire system, we simply sum the densities over all pools as:

$$\mathbf{f}(a,t) = \|\mathbf{f}(a,t)\|.$$

Transit Time Distributions

To obtain transit time distributions in the nonautonomous case, it is necessary to distinguish between the concepts of backward versus forward transit times. The backward transit time is defined as the age of particles in the output flux at the time of release from the system t_r . Using the fractional release coefficients, it is possible to obtain the vector of outflow rates at time t_r as:

$$z_j(t_r) = - \sum_{i=1}^n B_{ij}(t_r), \quad j = 1, 2, \dots, n.$$

The backward transit time distribution can be obtained as:

$$f_{\text{BTT}}(a, t_r) = \mathbf{z}^T(t_r) \cdot \mathbf{f}(a, t_r) \quad t_r \geq t_0.$$

Now, the forward transit time is defined as the age of an atom or particle that enters the system at an entering time $t_e > t_0$ and exits at time $t_r = t_e + a$. The forward transit time distribution can be obtained as:

$$f_{\text{FTT}}(a, t_e) = \mathbf{z}^T(t_e + a) \cdot \mathbf{f}(a, t_e + a).$$

Both distributions are tightly connected, with the forward transit time distribution of particles entering at time t_e being equal to the backward transit time distribution of the particles being released from the system at time t_r , i.e.:

$$f_{\text{FTT}}(a, t_e) = f_{\text{BTT}}(a, t_r).$$

FINAL REMARKS

The compartmental system representation also unveils analogies between deterministic systems that conserve mass with stochastic systems that conserve probabilities. This stochastic representation can be used to obtain formulas for the age of particles or atoms in the compartmental systems. With this approach, we derived formulas for the age of mass inside a compartmental system (system age), and the age of mass in the output flux (transit time). The concept of age can be very valuable to assess how old carbon and biogeochemical elements can be in an ecosystem. The concept of transit time can be very useful to understand how fast biogeochemical elements are processed inside an ecosystem, integrating all transfers and transformations that may take place.

There are other opportunities to further explore carbon cycle models in a stochastic setting. This could be particularly useful for studying, for example, the macroscopic properties at larger scales where patterns emerge by the action of microorganisms acting at microscopic scales. Also, the compartmental system representation may help to integrate concepts from other disciplines such as graph theory or control theory to address a number of questions not being explored yet in carbon cycle science.

SUGGESTED READING

A general introduction to the concepts of ages and transit times can be found in Bolin and Rodhe (1973). More specific results for the derivation of formulas and the computation of ages and transit times can be found in Rasmussen et al. (2016) for the mean of their distributions in nonautonomous systems, and for complete distributions in autonomous systems in Metzler and Sierra (2018), and for complete distributions in nonautonomous systems in Metzler, Müller, and Sierra (2018).

QUIZZES

1. Give examples of systems where the mean transit time is higher than the mean system age.
2. Give examples where the mean system age is higher than the mean transit time.
3. In what type of systems are the mean system age, the mean transit time, and the turnover time equal?



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER SIXTEEN

Practice 4

EFFICIENCY AND CONVERGENCE OF SEMI-ANALYTIC SPIN-UP (SASU) IN TECO

Xingjie Lu

Sun Yat-sen University, Guangzhou, China

CONTENTS

SASU to Improve Computational Efficiency of Spin-up of Biogeochemical Models / 129
Spin-up in the Simplified TECO Model / 129
Spin-up with Different Model Parameters / 131
Spin-up in a Weak Nonlinear System / 132

This practice aims to help the reader understand the convergence and efficiency of semi-analytic spin-up (SASU) for different carbon cycle representations in a model. By conducting the TECO spin-up using native dynamics (ND) and the SASU methods, we explore the convergence of different spin-up approaches, and the spin-up efficiency under different carbon input and soil turnover rates. Under weak nonlinear parameterization that assumes carbon input and soil turnover rates are functions of carbon pool sizes, the convergence of two approaches is verified and the spin-up efficiencies are compared.

SASU TO IMPROVE COMPUTATIONAL EFFICIENCY OF SPIN-UP OF BIOGEOCHEMICAL MODELS

Spin-up in biogeochemical models is an essential initialization procedure, which sets up the initial carbon and nitrogen pool sizes at a steady state for a model (see Chapter 14). When running a complex biogeochemical model, the spin-up is usually the most time-consuming procedure. So, the efficiency of spin-up becomes an important topic in biogeochemical model development.

Semi-analytic spin-up (SASU) is a recently developed technique to improve spin-up efficiency. The SASU approach builds on the fact that a matrix model of the carbon cycle can be semi-analytically solved to obtain steady-state values of pool sizes (Xia et al. 2012). Once biogeochemical models are presented in a matrix form, SASU can be easily implemented.

So far, SASU has been incorporated into the TECO, CABLE, ORCHIDEE, and CLM5 models. By implementing SASU, the spin-up computational efficiency is improved by 70% to 93% for different models under various configurations, including the most complicated, CLM5. CABLE was the first model to implement SASU for its carbon (C)-only and carbon-nitrogen (C-N) coupled versions. Results showed that SASU saves 92.4% of computational cost for C-only CABLE and 86.6% for the C-N version, when run at the global scale. Site simulations showed higher improvement in the spin-up efficiency.

SPIN-UP IN THE SIMPLIFIED TECO MODEL

We will apply native dynamics spin-up (ND) and semi-analytic spin-up (SASU) on a simplified

version of the TECO model in Exercises 1–3 in CarboTrain. The simplified TECO model was used in the Unit 3 practice, Chapter 12. It includes 7

pools, constant C input fluxes and turnover rate parameters. In Exercise 1, we are going to run ND and SASU for TECO.

EXERCISE 1

Run ND (default) spin-up for TECO matrix model (Ex 1.1), then change the spin-up approach to SASU (Ex 1.2). Learn whether carbon storage driven by the two alternative spin-up approaches converges to the same steady state. How fast is the SASU spin-up at reaching steady state compared with ND?

Ex 1.1. Follow the steps below to run TECO spin-up using ND approach in CarboTrain:

- Select **Unit 4**
- Select **Exercise 1**
- Select **Default ND**
- Select **Output Folder**
- Press **Run Exercise**
- Check results in Output Folder. Time-dependent variation in carbon input, each pool size, residence time, total ecosystem carbon storage, total ecosystem carbon storage capacity, and carbon storage potential in output.xls. Figures are in results.png. Take note of **years for steady state**, and **Total_C_10000** (the total carbon at year

10000) and complete the first row of Table 16.1, below.

Ex 1.2. Follow the steps below to run TECO spin-up using the SASU approach in CarboTrain:

- Select **Unit 4**
- Select **Exercise 1**
- Select **Default SASU**
- Select **Output Folder**
- Press **Run Exercise**
- Check results in Output Folder. Time-dependent variation in carbon input, each pool size, residence time, total ecosystem carbon storage, total ecosystem carbon storage capacity, and carbon storage potential in output.xls. Figures are in results.png. Take note of **years for steady state**, and **Total_C_10000** and complete the second row of Table 16.1, below.

QUESTIONS:

Does carbon storage converge to the same equilibrium in ND versus SASU? Which spin-up approach is faster at achieving equilibrium?

TABLE 16.1
Comparison of results from Exercise 1

Name	Description	C_{input}	k_{passsoil}	C_{init}	C_{10000}	T_{eq}
Ex 1.1	Default ND	0.00002245	1.5478×10^{-6}	0		
Ex 1.2	Default SASU	0.00002245	1.5478×10^{-6}	0		
Ex 1.3	High C_{input} ND	0.00004490	1.5478×10^{-6}	0		
Ex 1.4	High C_{input} SASU	0.00004490	1.5478×10^{-6}	0		
Ex 1.5	High k_{passsoil} ND	0.00002245	3.0956×10^{-6}	0		
Ex 1.6	High k_{passsoil} SASU	0.00002245	3.0956×10^{-6}	0		

Note. C_{input} = carbon input fluxes ($\text{gC m}^{-2} \text{s}^{-1}$); k_{passsoil} = turnover rate of soil carbon (day^{-1}); C_{init} = initial total carbon pool size (gC m^{-2}); C_{10000} = total carbon storage at year 10000 (gC m^{-2}); T_{eq} = spin-up convergence time (year).

SPIN-UP WITH DIFFERENT MODEL PARAMETERS

Previous studies on CABLE spin-up have shown that the C-only and C-N coupled model versions differed in efficiency and equilibrium for both ND and SASU spin-up. We may speculate that the differences in parameters and model structures may be the causes. In *Exercises 1.3–1.6*, we consider idealized cases, and use the simplified TECO model to

study whether or how the parameters will influence the spin-up efficiency and equilibrium.

Run ND (default) spin-up for the TECO matrix model with increasing C input and passive pool turnover rate (Ex 1.3 and 1.5), then change the spin-up approach to SASU (Ex 1.4 and 1.6). Learn whether carbon storage driven by the two spin-up approaches converges to the same steady state. How is spin-up efficiency affected by different parameters or carbon input? Complete Ex 1.3–1.6 and complete the remaining rows of Table 16.1.

Ex 1.3. Tune the carbon input and run TECO spin-up using the ND approach in CarboTrain:

- Select **Unit 4**
- Select **Exercise 1**
- Select **High carbon input ND**
- Select **Output Folder**
- Press **Open source code**
- Change the `input_fluxes` to 0.0000449 at line 46 and save the code
- Press **Run Exercise**
- Check results in Output Folder. Time-dependent variation in carbon input, each pool size, residence time, total ecosystem carbon storage, total ecosystem carbon storage capacity, and carbon storage potential in `output.xls`. Figures are in `results.png`. Take note of **years for steady state**, and **Total_C_10000** and enter results in Table 16.1, below.

Ex 1.4. Tune the carbon input and run TECO spin-up using the SASU approach in CarboTrain:

- Select **Unit 4**
- Select **Exercise 1**
- Select **High carbon input SASU**
- Select **Output Folder**
- Press **Open source code**
- Change the `input_fluxes` to 0.0000449 at line 46 and save the code
- Press **Run Exercise**
- Check results in Output Folder. Time-dependent variation in carbon input, each pool size, residence time, total ecosystem

carbon storage, total ecosystem carbon storage capacity, and carbon storage potential in `output.xls`. Figures are in `results.png`. Take note of **years for steady state**, and **Total_C_10000** and enter results in **Table 16.1**, below.

QUESTIONS:

Compared to the default cases (Ex 1.1 and 1.2), do carbon input increases change the spin-up time for steady state? Do carbon storages using the two spin-up approaches still converge when carbon input increases? Are the steady states the same when carbon inputs are higher? Why/why not?

Ex 1.5. Tune the passive soil turnover rate and run TECO spin-up using the ND approach in CarboTrain:

- Select **Unit 4**
- Select **Exercise 1**
- Select **High K passive ND**
- Select **Output Folder**
- Press **Open source code**
- Change the last column in variable `temp` to 0.00000309564 at line 34 and save the code
- Press **Run Exercise**
- Check results in Output Folder. Time-dependent variation in carbon input, each pool size, residence time, total ecosystem carbon storage, total ecosystem carbon storage capacity, and carbon storage potential in `output.xls`. Figures are in `results.png`. Take note of **years for steady state**, and **Total_C_10000** and enter results in **Table 16.1**, below.

Ex 1.6. Tune passive soil turnover rate and run TECO spin-up using the SASU approach in CarboTrain:

- Select **Unit 4**
- Select **Exercise 1**
- Select **High K passive SASU**
- Select **Output Folder**
- Press **Open source code**
- Change the last column in variable temp to 0.00000309564 at line 34 and save the code
- Press **Run Exercise**
- Check results in Output Folder. Time-dependent variation in carbon input, each pool size, residence time, total ecosystem carbon storage, total ecosystem

carbon storage capacity, and carbon storage potential in `output.xls`. Figures are in `result.png`. Take note of **years for steady state**, and **Total_C_10000** and enter results in **Table 16.1**, below.

QUESTIONS:

Compared to the default cases (Ex 1.1 and 1.2), do increases in passive soil turnover rate change the spin-up time for steady state? By comparing the **delta_C** in `result.png`, are results from SASU more stable? Do carbon storages using the two spin-up approaches still converge when passive soil turnover rate increases? Are the steady states the same when soil carbon turnover rate is higher? Why/why not?

SPIN-UP IN A WEAK NONLINEAR SYSTEM

Nonlinearity exists in terrestrial biogeochemical models. For example, the canopy photosynthesis (i.e., carbon input) is usually dependent on leaf

area, which is a function of leaf carbon pool size. In some models, the soil carbon turnover rate is pool size dependent. The following exercise studies how nonlinearity impacts the spin-up.

EXERCISE 2

Run ND (default) spin-up for the TECO matrix model with foliage nonlinearity and soil turnover rate nonlinearity (Ex 2.1 and 2.3), then change the spin-up approach to SASU (Ex 2.2 and 2.4). Learn whether carbon storage using the two alternative spin-up approaches converges to the same steady state. How long is the spin-up time for a model with nonlinearity in foliage or soil? Complete Ex 2.1–2.4 and enter results to complete Table 16.2.

Ex 2.1. Enable foliage nonlinearity in TECO and then run TECO using the ND spin-up in CarboTrain:

- Select **Unit 4**
- Select **Exercise 2**
- Select **Nonlinear foliage C with ND**
- Select **Output Folder**
- Press **Open source code**
- Add following lines at line 53 (remember to copy the indentation as well):

```
def input_fluxes(t, y):  
    tmp = 0.00000449 + 0.00002245 / 2 * (1 - np.exp(-0.5 * y[0] * 0.008)) / 0.5  
    return tmp
```

TABLE 16.2
Comparison of results from Exercise 1 and 2

Name	Description	C_{input}	$k_{passsoil}$	C_{init}	C_{10000}	T_{eq}
Ex 1.1	Default ND	0.00002245	1.5478×10^{-6}	0		
Ex 1.2	Default SASU	0.00002245	1.5478×10^{-6}	0		
Ex 2.1	Nonlinear foliage C with ND	$f(C_{foliage})$	1.5478×10^{-6}	0		
Ex 2.2	Nonlinear foliage C with SASU	$f(C_{foliage})$	1.5478×10^{-6}	0		
Ex 2.3	Nonlinear soil C with ND	0.00002245	$f(C_{passsoil})$	0		
Ex 2.4	Nonlinear foliage C with ND	0.00002245	$f(C_{passsoil})$	0		

Note. C_{input} = carbon input fluxes ($\text{gC m}^{-2} \text{s}^{-1}$); $k_{passsoil}$ = turnover rate of soil carbon (day^{-1}); C_{init} = initial total carbon pool size (gC m^{-2}); C_{10000} = total carbon storage at year 10000 (gC m^{-2}); T_{eq} = spin-up convergence time (year). $f(C_{foliage})$ = nonlinear relation between carbon input flux and foliage carbon (unitless) (see Ex 2.1 or Ex 2.2 for details). $f(C_{passsoil})$ = nonlinear relation between passive soil turnover rate and passive soil carbon storage (unitless) (see Ex 2.3 or Ex 2.4 for details).

The lines added here describe the nonlinear relation between carbon input fluxes and foliage carbon (see Figure 16.1), which is formulated by:

$$C_{input} = f(C_{foliage}) = C_{input, min} + C_{input0} \cdot \frac{1 - e^{-k_d \cdot C_{foliage} \cdot SLA}}{k_d \cdot LAI_0} \quad (16.1)$$

where $C_{input, min} = 0.00000449 \text{ gC m}^{-2} \text{s}^{-1}$, $C_{input0} = 0.00002245 \text{ gC m}^{-2} \text{s}^{-1}$, $k_d = 0.5$, $SLA = 0.008 \text{ m}^2(\text{gC})^{-1}$, $LAI_0 = 2 \text{ m}^2 \text{m}^{-2}$, $C_{foliage}$ is the foliage carbon.

- g. Press **Run Exercise**
- h. Check results in Output Folder. Time-dependent variation in carbon input, each pool size, residence time, total ecosystem carbon storage, total ecosystem carbon storage capacity, and carbon storage potential in output.xls. Figures are in results.png. Take note of **years for steady state**, and **Total_C_10000** and complete the third row of Table 16.2, below.

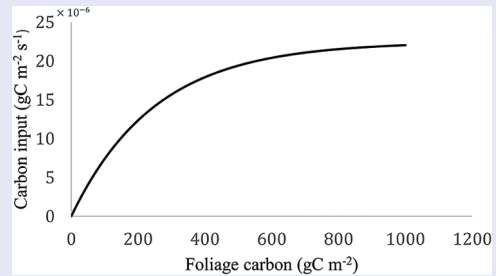


Figure 16.1. The nonlinear relation between carbon input and foliage carbon.

Ex 2.2. Enable foliage nonlinearity in TECO and then run TECO spin-up using SASU approach in CarboTrain:

- a. Select **Unit 4**
- b. Select **Exercise 2**
- c. Select **Nonlinear foliage C with SASU**
- d. Select **Output Folder**
- e. Press **Open source code**

Add following lines at line 53 (remember to copy the indentation as well):

```
def input_fluxes(t, y):
    tmp = 0.00000449 + 0.00002245 / 2 * (1 - np.exp(-0.5 * y[0] * 0.008)) / 0.5
    return tmp
```

The lines added here were explained in Step f of Ex 2.1.

f. Press **Run Exercise**

g. Check results in Output Folder. Time-dependent variation in carbon input, each pool size, residence time, total ecosystem carbon storage, total ecosystem carbon storage capacity, and carbon storage potential in output.xls. Figures are in results.png. Take note of **years for steady state**, and **Total_C_10000** and enter results in Table 16.2, below.

QUESTIONS:

Compared to the default cases (Ex 1.1 and 1.2), does foliage nonlinearity increase the spin-up time for steady state? Does carbon storage using the two spin-up approaches still converge when foliage nonlinearity is enabled? Are the steady states the same between the two spin-up approaches?

Ex 2.3. Enable nonlinearity in passive soil turnover rate and run TECO using the ND spin-up approach in CarboTrain:

- Select **Unit 4**
- Select **Exercise 2**
- Select **Nonlinear soil C with ND**
- Select **Output Folder**
- Press **Open source code**
- Replace line 53 with the following lines (remember to copy the indentation as well):

def fun_K(t, y):

```
K[6][6] = temp[6] / 86400 * (y[6] / 10000.0 + 0.1)
```

```
return K
```

```
mod = GeneralModel(times, B, A, fun_K, iv_list, input_fluxes)
```

The lines revised here describe the relation between passive soil turnover rate and passive soil carbon (see Figure 16.2), which is formulated by:

$$k_{\text{passsoil}} = f(C_{\text{passsoil}}) = k_0 \cdot \left(\frac{C_{\text{passsoil}}}{10000.0} + 0.1 \right) \quad (16.2)$$

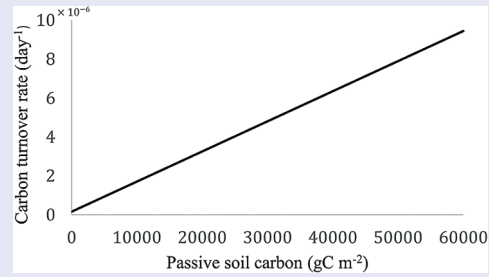


Figure 16.2. The relation between passive soil turnover rate and passive soil carbon

where $k_0 = 0.00000154782 \text{ day}^{-1}$, C_{passsoil} is the passive soil carbon storage.

g. Press **Run Exercise**

h. Check results in Output Folder. Time-dependent variation in carbon input, each pool size, residence time, total ecosystem carbon storage, total ecosystem carbon storage capacity, and carbon storage potential in output.xls. Figures are in results.png. Take note of **years for steady state**, and **Total_C_10000** and enter results in Table 16.2, below.

Ex 2.4. Enable nonlinearity in passive soil turnover rate and run TECO using the SASU spin-up in CarboTrain:

- Select **Unit 4**
- Select **Exercise 2**
- Select **Nonlinear soil C with SASU**
- Select **Output Folder**
- Press **Open source code**
- Replace line 53 with the following lines (remember to copy the indentation as well):

def fun_K(t, y):

```
K[6][6] = temp[6] / 86400 * (y[6] / 10000.0 + 0.1)
```

```
return K
```

```
mod = GeneralModel(times, B, A, fun_K, iv_list, input_fluxes)
```

The lines revised here were explained in Step f of Ex 2.3.

- g. Press **Run Exercise**
- h. Check results in Output Folder. Time-dependent variation in carbon input, each pool size, residence time, total ecosystem carbon storage, total ecosystem carbon storage capacity, and carbon storage potential in `output.xls`. Figures are in `results.png`. Take note of **years for steady state**, and **Total_C_10000** and enter results in Table 16.2, above.

QUESTIONS:

Compared to the default cases (Ex 1.1 and 1.2), does nonlinearity in soil turnover rate increase the spin-up time? Does carbon storage using the two spin-up approaches still converge when nonlinearity in soil turnover rate is enabled? Are the steady states the same between the two spin-up approaches?



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

UNIT FIVE

Traceability and Benchmark Analysis



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER SEVENTEEN

Overview of Traceability Analysis

Jianyang Xia

East China Normal University, Shanghai, China

CONTENTS

A Key Challenge for Earth System Models: Identification of Uncertainty Sources /	139
Traceability Framework: Design and Key Components /	140
Benefits of Traceability Analysis for Identifying Model Uncertainty Sources /	141
Summary /	146
Suggested Reading /	146
Quizzes /	146

This chapter provides an overview of the traceability framework as a method for identifying the key sources of uncertainty in land carbon cycle modeling. The analytical framework is built upon derivations of the matrix equation introduced in Chapter 1. This framework provides a systematic way to decompose the uncertainty of land carbon cycle projections into traceable components. Thus, traceability analysis can facilitate the understanding of carbon cycling, its drivers and controlling processes within a model and across models with different carbon-cycle structures, external climate forcings, and scientific assumptions. This chapter also introduces several examples to demonstrate the application of traceability analysis to model evaluation.

A KEY CHALLENGE FOR EARTH SYSTEM MODELS: IDENTIFICATION OF UNCERTAINTY SOURCES

Carbon cycle schemes incorporated in Earth system models (ESMs) and offline land surface models are widely used to simulate terrestrial biogeochemistry and its feedback to climate change. The processes underpinning the terrestrial carbon (C) cycle constitute one of the most uncertain components, and this uncertainty has become a bottleneck in Earth system modeling. The model-to-model variation in future projections of the

global land C sink remained large from the third assessment report of the Intergovernmental Panel on Climate Change (IPCC), published in 2001, to the fifth report, published in 2017, despite an intervening period of over 15 years of research and model development. The new ESM projections and analysis in Phase 6 of the Coupled Model Intercomparison Project (CMIP6) show that uncertainty on the recent past and potential future evolution of the land carbon cycle, relevant to the assessment and understanding of global climate change, is still a prominent feature in the sixth IPCC report, published in 2021. Thus, one key challenge for Earth system science is how to reduce the large disagreement in carbon cycle predictions among models of the terrestrial biosphere incorporated in ESMs.

To tackle this challenge, we need to answer one question: why are terrestrial carbon cycle predictions different among models? In the past three decades, numerous model intercomparison projects (MIPs) have been established, in part to shed light on this question. Although those MIPs have made important contributions to model evaluation and synthesis, they have been limited by design in their ability to explore the sources of model uncertainty. Thus, our question can be refined into how we can trace model uncertainty back to its sources, such as model structures, parameters and external forcings.

TRACEABILITY FRAMEWORK: DESIGN AND KEY COMPONENTS

In Chapter 1 we noted that the terrestrial carbon cycle has four fundamental properties: (1) photosynthesis as the primary carbon influx pathway; (2) the transfer of assimilated carbon among different C pools (e.g., plant litter and soil); (3) the rate of transfer of this carbon controlled by the donor pool; and (4) the process of carbon transfer from the donor pool to the recipient pool, which can be described by first-order kinetics. These four fundamental properties, with minor variations, have emerged in most models using a pool-and-flux structure. Thus, the modeled terrestrial carbon cycle can be tracked by the equation:

$$\frac{dX(t)}{dt} = Bu(t) + A\xi(t)KX(t) \quad (17.1)$$

where the $\frac{dX(t)}{dt}$ characterizes the dynamic of terrestrial carbon storage, u is the carbon inputs from photosynthesis, and B is the partitioning coefficients to live carbon pools (e.g., leaf, wood, and root). Thus, the term $Bu(t)$ represents the partitioning of the assimilated carbon from photosynthesis among different plant carbon pools. The second term, $A\xi(t)KX(t)$, describes the movement and exit rates of carbon atoms along their transferring paths (Xia et al., 2013; Luo et al., 2017). A is a transfer coefficients matrix to represent the movements of carbon atoms among multiple carbon pools. K represents the exit rates of different carbon pools. $\xi(t)$ modifies the decay rate of different pools by adding influences from environmental factors (temperature, moisture, nutrients, soil texture, and so on). Different models may have different combinations of plant and soil pools and different values in vector B or matrices A and C , but they can be unified by this matrix equation.

At steady state, terrestrial carbon reservoirs reach their maximum storage capacity and there are no further net carbon exchanges. Therefore, the steady-state land carbon storage (X_{ss}) can be solved by letting Equation 17.1 equal 0 (i.e., $d(X)/dt = 0$):

$$X_{ss} = (-A\xi K)^{-1}BU_{ss} \quad (17.2)$$

where X_{ss} is a vector containing the steady-state pool sizes of all carbon pools, and U_{ss} is the carbon

influx at steady state. The term $(-A\xi K)^{-1}B$ measures the ecosystem carbon residence time (τ_E), which is an important ecosystem property involving multiple processes, including carbon allocation (the B matrix), carbon transfer network (the A matrix), decomposition processes (the K matrix) and modifications from environmental factors (the ξ matrix). Here, net primary productivity (NPP) is treated as carbon inputs, and the X_{ss} can then be decomposed into NPP and τ_E :

$$X_{ss} = NPP \times \tau_E \quad (17.3)$$

As mentioned above, τ_E is determined by four items: $(-A\xi K)^{-1}B$. Here, A , K and B are the intrinsic model properties, while ξ represents external influences from environmental factors. We thus rearrange these items and express τ_E as:

$$\tau_E = (-A^{-1}K^{-1}B)\xi^{-1} \quad (17.4)$$

We further merge $(-A^{-1}K^{-1}B)$ and define it as the baseline carbon residence time (τ'_E). Then, τ_E can be decomposed into τ'_E and ξ^{-1} :

$$\tau_E = \tau'_E \xi^{-1} \quad (17.5)$$

where τ'_E is determined by model intrinsic properties, associated with plant trait, soil attributes, number of C pools and how C is transferred among these pools at different cycling rates. As defined above, ξ represents the modifying effects of environmental factors as a fractional value applied multiplicatively to the baseline C residence time. If, as an example, we consider temperature and water availability as the main environmental factors scaling process rates in our system, we may divide ξ into a temperature scalar (ξ_T) and a water scalar (ξ_W):

$$\xi = \xi_T \xi_W \quad (17.6)$$

Through the above mathematical rearrangements of Equation 17.2, the modeled land carbon storage at steady state is decomposed into its determinative components as illustrated in Figure 17.1. X_{ss} is first decomposed into NPP and τ_E . Then, τ_E can be further traced into τ'_E and ξ , while ξ is divided into ξ_T and ξ_W . This framework offers a systematic way to decompose the steady-state land carbon cycle in a form that fits the structure of many common land carbon models. Most recently, advances in terrestrial carbon cycle theory (Luo et al., 2017)

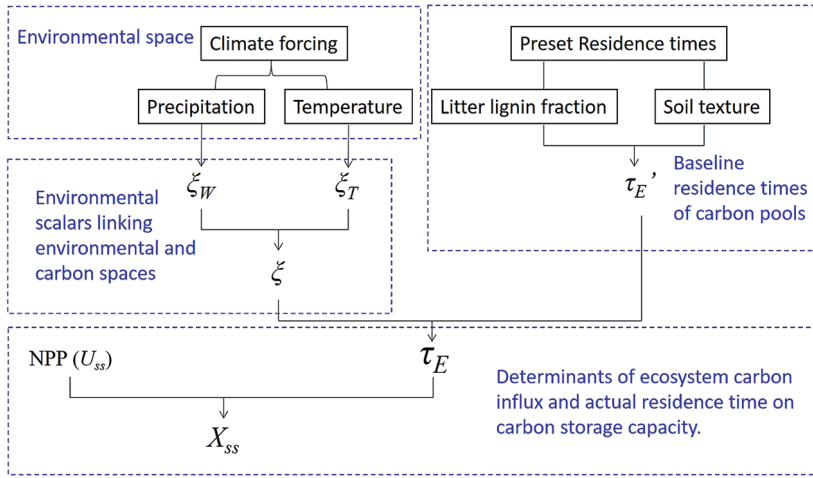


Figure 17.1. The traceability framework for decomposing steady-state terrestrial carbon storage into its traceable components.

have extended the capability of this framework to trace transient-state land carbon storage. This is explored further in Chapter 18.

BENEFITS OF TRACEABILITY ANALYSIS FOR IDENTIFYING MODEL UNCERTAINTY SOURCES

Land carbon cycle schemes in most Earth system models can be mathematically represented in the form of Equation 17.1. Thus, the traceability framework can facilitate the evaluation of land

carbon cycle models and the ESM frameworks to which they are coupled by identifying the sources of simulation difference within and between models. It should be noted that the uncertainty of land carbon cycling in ESMs not only stems from model structure and parameters, but also is affected by climate outputs from the atmospheric components of ESMs (Ahlström et al., 2017). Below are some cases which have shown the application of traceability analyses for model evaluation. More application cases are summarized in Chapter 18.

CASE 1: Understanding Terrestrial Carbon Cycle Variations Within a Model

Modern global land-surface models account for a vast array of different processes, making it difficult to trace features of the results of simulations to the model's constituent processes and interactions to aid understanding and evaluation. For example, all global land carbon models simulate the spatial variability of ecosystem C storage, but it is often unclear how the geographic distribution of ecosystem C storage across a global map is determined. Xia et al. (2013) introduced the framework of traceability analysis and applied it to the Australian Community Atmosphere Biosphere Land Exchange (CABLE) model to help understand differences in modeled carbon processes among biomes and as influenced by nitrogen processes.

They first estimated ecosystem carbon capacity among different biomes in CABLE and traced it to the ecosystem residence time (τ_E) and carbon influx (i.e., NPP or U_{ss} in Equation 17.1). Results indicated that different biomes showed different patterns. For example, evergreen broadleaf forests have a very high NPP but a short carbon residence time. The tundra biome has a very low NPP but a long carbon residence time. Some barren biomes, such as deserts, have low carbon storage capacity because of low NPP and short carbon residence time (Figure 17.2). Using Equation 17.5 they then further decomposed the ecosystem carbon residence time into baseline carbon residence time and environmental scalars. They found that tundra and evergreen broadleaf forests have similar baseline carbon residence times, but the environmental scalars are much lower

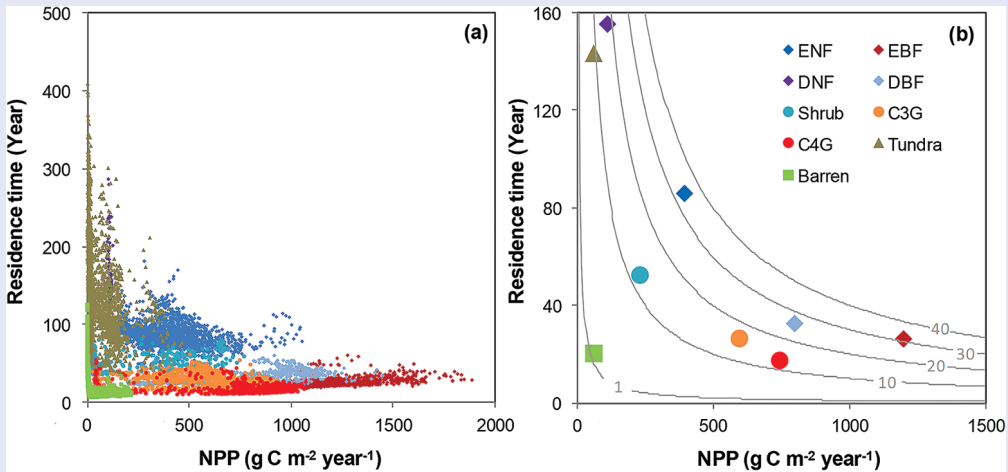


Figure 17.2. Determinants of ecosystem C storage capacity by NPP and residence time. Values of all grid cells are plotted in panel (a). In panel (b), the hyperbolic curves represent constant values of ecosystem carbon storage capacity. ENF – Evergreen needle leaf forest, EBF – Evergreen broadleaf forest, DNF – Deciduous needle leaf forest, DBF – Deciduous broadleaf forest, Shrub – Shrub land, C3G – C3 grassland, C4G – C4 grassland, and Barren –barren/sparse vegetation.

in tundra than evergreen broadleaf forests. The lower value of the environmental scalar signifies stronger environmental limitations on decomposition rates of organic carbon, which implies that tundra has much longer actual carbon residence time than evergreen broadleaf forests, even though the baseline residence time

is similar. The environmental space of temperature and water scalars among biomes in the CABLE model was then considered, which led to the insight that the spatial difference in environmental control of carbon residence time is mainly driven by the temperature scalar rather than the water scalar in CABLE.

Because the CABLE model can be applied with or without carbon-nitrogen coupling enabled, the traceability analysis can be used to evaluate how the incorporation of nitrogen cycle processes can affect carbon cycling within the model. The analysis showed that incorporating the nitrogen cycle into

CABLE reduces ecosystem carbon storage capacity in all biomes in comparison with the carbon-only model. Specifically, the CABLE model simulates lower NPP in woody biomes but shorter carbon residence time for non-woody biomes when the nitrogen cycle is switched on (Xia et al., 2013).

CASE 2: Intermodel Comparisons of Terrestrial Carbon Cycle Simulations

Although model intercomparison projects have shown that the ensemble means of multiple models fit data well, the intermodel difference in carbon cycle pools and fluxes is usually large. To better understand the sources of variations in modeled carbon storage capacity among models, Rafique et al. (2016) used the traceability framework to compare two land carbon cycle models, CLM-CASA' and CABLE. As shown in Figure 9.2, CABLE

and CLM-CASA' both showed a distinctive structure of the ecosystem carbon cycle, like the number of carbon pools, NPP partitioning coefficients, decomposition rates and carbon transfer coefficients among different pools, which resulted in different baseline carbon residence times. Due to more NPP partitioning into roots and wood, the baseline carbon residence time in CABLE was longer than in CLM-CASA'. Despite difference in model structure, the traceability analysis showed that CABLE and CLM-CASA' simulate similar

global mean soil carbon storage capacity. This is because CABLE has lower NPP but longer carbon residence time compared to CLM-CASA'. The longer carbon residence time in CABLE mainly results from the baseline carbon residence time rather than the environmental scalars.

Overall, this case indicated that the major factors contributing to the differences between

the two models were primarily due to parameter settings related to photosynthesis, carbon input, baseline residence times, and environmental conditions. The application of traceability analysis to intermodel comparisons is useful for developing a model with different versions or interpreting the different predictions of land carbon cycle by different models under the same climate forcings.

CASE 3: Evaluating Impacts of External Forcings on Terrestrial Carbon Cycle Simulations

One important uncertainty source in land carbon cycle studies stems from the differences in baseline and projected climate fields generated by different climate and Earth system models.

When a land carbon model is forced by output from different climate models, the resultant predictions of the land carbon cycle also differ – a source of uncertainty propagating through the carbon cycle model from the forcing climate model. By applying traceability analysis on the LPJ-GUESS model when forced with

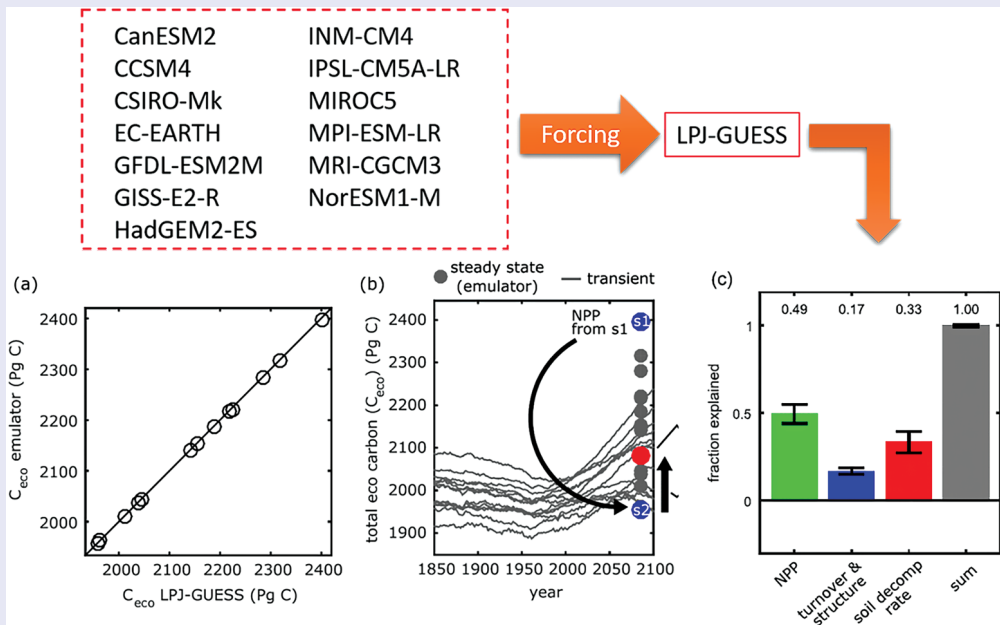


Figure 17.3. Emulator performance and example of experimental design. (a) Comparison between global ecosystem carbon stock (C_{eco}) at steady state as simulated by LPJ-GUESS and emulator solution of global C_{eco} at steady state for the 13 simulations and emulator solutions. The small deviations from the 1:1 relationship (black line) are mainly due to the internal variability in LPJ-GUESS where a true steady state is never reached, in contrast to the emulator, which solves the steady-state conditions analytically. (b) Illustration of experimental design. Gray curves are time trajectories of C_{eco} from transient simulations with LPJ-GUESS; circles denote corresponding emulator-computed steady-state values of C_{eco} . (c) Global partitioning of steady-state C_{eco} uncertainties. From this panel, we can see that NPP is the largest contributor to the difference in modeled land carbon uptake, accounting for almost 50% of variability among simulations with different forcings. Figure reproduced from Ahlström et al. (2015).

13 different climate datasets from CMIP5 general circulation models, Ahlström et al. (2015) studied the impact of climate model uncertainty on the land carbon cycle. LPJ-GUESS is a global dynamic vegetation-ecosystem model that is based on detailed representation of vegetation structure, demography, and resource competition. To understand how ecosystems around the world respond to future projections of atmospheric CO₂ concentrations and climate, transient and steady-state simulations were performed forced by output fields from different climate models under the RCP8.5 radiative forcing scenario. The authors then quantified the relative contributions of the three groups of processes – NPP, vegetation dynamics and turnover, and soil decomposition – to future carbon uptake uncertainties. They achieved this

by fitting the traceability framework as an emulator to each of the 13 LPJ-GUESS simulations. Because the emulator has a common structure, it was possible to ‘exchange’ the key carbon cycle processes among the 13 simulations, allowing the importance of each in explaining the variability among simulations to be derived (Figure 17.3). Further details about the model and simulations can be found in Ahlström et al. (2015). Since there is only one carbon model involved in this study, we can be sure that all the differences in the carbon cycle projections stem from the external climate forcing. The results showed that NPP, vegetation turnover, and soil decomposition rate respectively explain 49%, 17%, and 33% of the uncertainty in carbon uptake by terrestrial ecosystems globally under the RCP8.5 future scenario (Figure 17.3c).

The traceability framework could be applied in a similar way to analyze the carbon cycle impacts of other types of external forcings, such as different CO₂ scenarios, disturbance regimes, and land use/cover changes. As shown in Ahlström et al. (2015), this approach can help to diagnose which

processes are most important in determining the model uncertainty under given external forcings. This in turn can identify processes for priority attention in evaluating, revising or improving the carbon cycle model.

CASE 4: Assessment of New Processes in Land Carbon Models

Although the integration of more process detail into a model can increase its utility, it also tends to increase the number of interacting processes and feedbacks influencing the model output, and the relationship of the model output to the model forcing data. As a result, it can become more difficult to understand or evaluate how the new incorporated processes influence the behavior and performance of the model. For example, recognizing the important role that nitrogen (N) availability plays for the dynamics of the world’s ecosystems, many current models that originally included only a carbon cycle have been enhanced to incorporate a nitrogen cycle that interacts with the model’s carbon processes and state. The availability of nitrogen can strongly affect both ecosystem carbon input and mean carbon residence time. Nevertheless, the detailed structure of the C-N

coupling scheme varies greatly among different models. How these diverse representations of C-N interactions affect carbon cycle modeling remains unclear. Thus, Du et al. (2018) incorporated three different C-N coupling schemes, derived from the TECO-CN, CLM4.5, and O-CN models, into the carbon-only version of the Terrestrial ECOSystem (TECO) model and then used the traceability framework to evaluate their impacts on the carbon cycle. The three C-N coupled frameworks are different in many aspects, including C:N stoichiometry in plants and the soil, plant N uptake strategies, down-regulation of photosynthesis under N deficit, and the pathways of N acquisition.

As shown in Figure 17.4a, each N process is simulated with different assumptions by different models. The results showed that each of the integrated C-N coupling schemes reduced the carbon storage capacity compared with the carbon-only version of TECO. However, the magnitude of the reduction varies among the

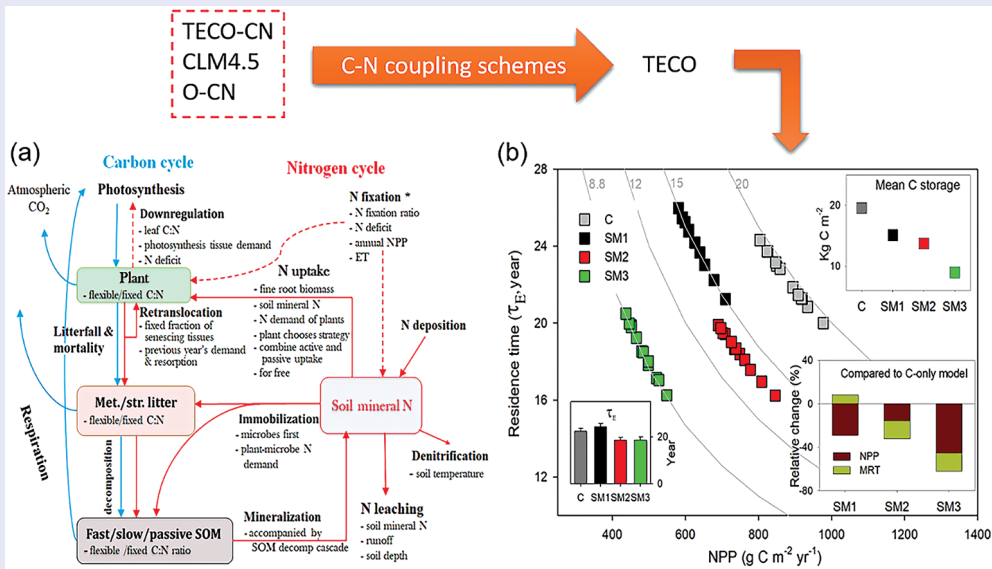


Figure 17.4. Schematic diagram of the terrestrial ecosystem carbon and nitrogen coupling model and its traceable components under different C-N coupling schemes. (a) The major carbon and nitrogen pool-and-flux structure in a terrestrial ecosystem, with alternative assumptions of the N processes represent in SM1 (TECO-CN), SM2 (CLM4.5), and SM3 (O-CN) C-N coupling schemes. Light blue arrows indicate C-cycle processes and red arrows show N-cycle processes. Met./Str. litter – metabolic and/or structural litter; SOM – soil organic matter. *Set N fixation as an option when the plant N uptake is not enough for growth in terms of C investment in SM1, but go directly to soil mineral N pool in SM2 and SM3. (b) Simulation of annual ecosystem carbon storage capacity for 1996 to 2006 at Duke Forest by carbon in flux (NPP, x-axis) and ecosystem residence time (τ_E , y-axis) in the TECO model framework with three C-N coupling schemes (SM1, SM2, and SM3) and in the TECO C-only model (C). The inserted panel in the left bottom corner shows the τ_E in SM1, SM2, SM3, and the C-only model; The top right panel shows the mean ecosystem carbon storage simulated among SM1, SM2, SM3, and the C-only model; The right bottom panel shows the relative change in the simulated NPP and τ_E among the three schemes compared with the C-only model.

three schemes, i.e., -23%, -30%, and -54% for TECO-CN, CLM4.5 and O-CN, respectively. The reduced carbon storage capacity was driven by reduced NPP (-29%, -15%, and -45%) and mean C residence time (9%, -17%, and -17%) (Figure 17.4b). The differences in these results for different N cycle implementations in

TECO indicate that adding interactive nitrogen dynamics to a carbon cycle model could generate new uncertainty sources for carbon cycle-climate feedbacks. This example illustrates that the traceability analysis can improve our understanding of the impacts of newly incorporated processes on an existing carbon cycle model.

CASE 5: Accelerating the Pace of Model Evaluation Via Online Tools

Increasing model complexity not only leads to a large divergence in simulations of the land carbon cycle by different models, but also tends to increase the computational consumption of model runs, which can become a bottleneck for model evaluations. As illustrated above, the traceability framework allows a carbon cycle

model to be simplified and generalized into several traceable components, which can be further decomposed to quantify the structural sources of the uncertainty built into the full version of the model. Thus, an online traceability analysis system for model evaluation (TraceME) was built to accelerate the pace of model evaluation of the land carbon cycle, with Earth system models in mind.

The TraceME system is a cloud-based platform (Zhou et al., 2021). It provides user-friendly interfaces and scientific workflow to automatically perform the model evaluation. On the website (<http://www.traceme.org.cn/>, last access: October 2020), the user can select the data of interest, and submit the task through the browser to complete the traceability analysis. The collaborative framework of TraceME provides a convenient data-sharing platform for all users to filter data of interest from the entire system for traceability analysis. Once a task is requested through a web browser, the scientific workflow of TraceME is triggered and it will execute the corresponding processes, such as data preprocessing, traceability

analysis, and evaluation. The submitted data will be systematically decomposed into traceable components for quantifying the variance contributions of these components to land carbon dynamics. After the submitted task is completed, TraceME provides a visual interface to show and download the results in the forms of figures and Network Common Data Form-format (NetCDF) files. These files can be used to perform further analysis. TraceME is a convenient tool to evaluate models based on the traceability analysis framework. It has, for example, been applied to evaluate land carbon dynamics in CMIP6 Earth system models (Zhou et al., 2021).

SUMMARY

The traceability analysis provides a relatively new approach to the evaluation of model uncertainties impacting simulations of the terrestrial carbon cycle. The framework builds on the fundamental properties of the terrestrial carbon cycle, which are reflected broadly by different models despite differences in structure and process detail. Equation 17.1 provides the theoretical basis for the traceability analysis. The application cases described illustrate how traceability analysis can benefit the understanding of variations in terrestrial carbon cycling within a model, the intercomparison of terrestrial carbon cycling among models, evaluation of the contributions of external forcings to carbon-cycle uncertainty, the assessment of newly incorporated processes into carbon cycle models, and the development of online tools for quick and consistent model evaluation. The application cases presented in this chapter mainly focus on the steady-state ecosystem C storage. In Chapter 18 we will explore how the traceability analysis can be adapted to apply to the transient dynamics of the land carbon cycle in models.

The traceability framework is developed for the terrestrial carbon cycle, but it can be extended to include nutrient and water processes. Further applications include the integration of benchmark analysis (see Chapter 19) with traceability analysis, the connection between structurally traceable components and model parameters, and the application of traceability analysis to understand

climate-carbon cycle feedbacks in coupled land-atmosphere simulations with Earth system models.

SUGGESTED READING

Xia, J., Luo, Y., Wang, Y.-P., & Hararuk, O. (2013). Traceable components of terrestrial carbon storage capacity in biogeochemical models. *Global Change Biology*, 19(7), 2104–2116

QUIZZES

1. In the steady-state traceability analysis, which variable is first decomposed into NPP and carbon residence time?
 - a. Soil C stock
 - b. Vegetation biomass
 - c. Ecosystem total C stock
 - d. Ecosystem C storage capacity
2. When the nitrogen cycle is incorporated into a carbon cycle model, which components in the terrestrial carbon cycle will be changed? What happened in the CABLE model?
3. Which model has the longer ecosystem carbon residence time, CABLE or CLM-CASA? Why?
4. Describe the external forcings which can contribute to the large model uncertainty of the terrestrial carbon cycle.
5. How do TECO-CN, CLM4.5, and O-CN models differ in their approach to simulating the coupling between the terrestrial carbon and nitrogen cycles? How do these differences impact carbon cycle dynamics in the TECO model?

CHAPTER EIGHTEEN

Applications of the Transient Traceability Framework

Lifen Jiang

Northern Arizona University, Flagstaff, USA

CONTENTS

Introduction /	147
A Traceability Framework for Transient Land Carbon Storage Dynamics /	148
Transient Traceability Analysis of Carbon Storage at Duke Forest and Harvard Forest /	149
Transient Traceability Analysis of Land Carbon Storage in Model Intercomparison Projects /	152
Summary /	156
Suggested Reading /	156
Quizzes /	156

This chapter introduces the traceability analysis of transient carbon storage, which is a modification of the original traceability framework that relies on an assumption of steady state. Transient traceability analysis is particularly useful to address the origin and drivers of carbon flow of a system in a state of transition towards a new steady state – a transient system. We will illustrate how the transient traceability can be applied to address scientific questions with two examples. One tracks the differences in modeled carbon storage between two forest ecosystems, the second compares and contrasts outcomes of a set of Model Intercomparison Projects (MIPs) encompassing multiple land carbon models.

INTRODUCTION

Simulations by Earth system models in the Coupled Model Intercomparison Phase 5 project (CMIP5) showed that differences among the models entail large uncertainty in land C storage (Jones et al. 2013). The spread of simulated future land C change across the models is even greater than the spread across the four radiative forcing scenarios, when the ensemble averages for each scenario are compared. Arora et al. (2020) compared model results

from two CMIP phases (CMIP6 vs. CMIP5) and they found that the model mean values of the carbon-concentration and carbon-climate feedback parameters and their multi-model spread under a 1% per year CO₂ increase experiment has not changed significantly across the two CMIP phases for both land and ocean. Moving forward, a big challenge is how to understand and reduce the uncertainty across models to achieve more reliable predictions.

Although land C models have become increasingly complex in recent decades with more and more processes incorporated, most current models have the same theoretical foundation and therefore share some of the same general properties, as described in chapters 1 and 2. These shared theoretical foundations and properties enable many land C models to be represented or approximated in matrix form, as demonstrated in chapter 5. With the matrix representations of the C cycle models, we are able to decompose the modeled land C storage into different traceable components, which is the unified diagnostic system for uncertainty analysis, an overview of which was provided in chapter 9. Based on the common properties and matrix representations of land C models, Xia et al. (2013) developed a traceability framework to decompose steady-state C storage into traceable components. Chapter 17 describes how the framework could be

applied to investigate differences in carbon storage across biomes and how differences in model structure influence the simulated effects of nitrogen cycling and land use change on the carbon cycle.

The traceability framework in its original form depends on a steady-state assumption for ecosystem carbon stocks. However, due to climate change and disturbance history, most ecosystems are not at steady state. The non-steady state is called transient state and the challenge is how to trace transient C storage dynamics. In this chapter, we first introduce a general framework for transient traceability analysis. Then, we illustrate how the transient traceability framework can be applied to address scientific questions using two examples. Our first example uses transient traceability analysis to track differences in C storage between two forest ecosystems. The second applies transient traceability analysis to three model intercomparison projects (MIPs) to identify the sources for the uncertainty in modeled carbon storage dynamics within each MIP and across the three MIPs.

A TRACEABILITY FRAMEWORK FOR TRANSIENT LAND CARBON STORAGE DYNAMICS

In order to realize the traceability of transient C storage, Luo et al. (2017) conducted a theoretical analysis to extend the steady-state traceability framework to work for systems in a transient state. The key was to add another term called C storage potential. It was introduced in chapter 9 and is further discussed below. The core equation for traceability of transient C storage is from Luo et al. (2017) and as follows:

$$X(t) = (-A\xi(t)K)^{-1}B(t)u(t) - (-A\xi(t)K)^{-1}X'(t) \quad (18.1)$$

where $X(t)$ is individual pool size at time t , which is a vector in a multi-pool model; A is a matrix of transfer coefficients between C pools; $\xi(t)$ is a diagonal matrix of environmental scalars to reflect the control of physical and chemical properties, e.g., temperature, moisture, nutrients, litter quality, and soil texture, on C cycle processes; K is a diagonal matrix of exit rates from donor pools, which encapsulates mortality rates for plant pools and decomposition coefficients for litter and soil pools; B is a vector of allocation coefficients of C input to each pool; $u(t)$ is C input, i.e., NPP or GPP;

and $X'(t)$ is net change of any individual C pool at time t , which is a vector for a multi-pool model. The sum of X' of all individual C pools corresponds to net ecosystem production (NEP), or the sign opposite of net ecosystem exchange (NEE).

In this equation, the inverse of the product of $-A$, $\xi(t)$ and K , i.e., $(-A\xi(t)K)^{-1}$, is named chasing time, which is a matrix representing the timescale for the net C pool change to be redistributed in the network consisting of all C pools. The product of chasing time and the allocation coefficient B , i.e., $(-A\xi(t)K)^{-1}B(t)$, is the residence time of individual pools. The product of the residence time and C input is the maximum C that individual pools or the whole ecosystem can store at a time, which is defined as C storage capacity, X_c , that is, $X_c = (-A\xi(t)K)^{-1}B(t)u(t)$. The second term in equation 18.1 – the product of chasing time and net C pool change $((-A\xi(t)K)^{-1}X'(t))$ – represents redistribution of net change of individual C pools in the network. This redistribution of net C pool change indicates the potential of an individual pool or the whole ecosystem to gain or lose C. Therefore, it is named as C storage potential, X_p . So, we can derive another equation from the above descriptions as follows:

$$X(t) = X_c(t) - X_p(t) \quad (18.2)$$

This is the overall equation of the transient traceability framework of land carbon storage. To demonstrate how this transient traceability framework is a useful tool, we will walk through two example applications. The first application uses the framework exactly as described above to investigate the differences in carbon storage between two forest ecosystems, Duke Forest and Harvard Forest, and analyze the underlying mechanisms that explain these differences. The second application uses this general framework in combination with some other methods to decompose modeled land carbon storage in three MIPs: CMIP5, Trends in Net Land-Atmosphere Carbon Exchange (TRENDY), and Multiscale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP). Decomposing differences in land carbon storage across models within each MIP and between the three MIPs helps us to understand why the models simulate different outcomes, and what features of the model structure or parameters may be responsible for these differences.

TRANSIENT TRACEABILITY ANALYSIS OF CARBON STORAGE AT DUKE FOREST AND HARVARD FOREST

In this case study, we will go over an application of the matrix approach to trace differences in carbon storage dynamics between Duke Forest and Harvard Forest. Duke Forest, located in North Carolina, USA ($35^{\circ}58'41''\text{N}$, $79^{\circ}5'39''\text{W}$), is evergreen needleleaf forest and the dominant tree species at this site is *Pinus taeda* (loblolly pine). Harvard Forest, located in Massachusetts, USA ($42^{\circ}32'16''\text{N}$, $72^{\circ}10'17''\text{W}$), is deciduous broadleaf forest dominated by *Quercus rubra* (red oak) and *Acer rubrum* (red maple). These two study sites have contrasting ecosystem types and there are plenty of measured data at both sites available for calibrating our model.

The framework for transient traceability of land C storage dynamics is shown in Figure 18.1. Using this framework, we can decompose transient C storage into C storage capacity and C storage potential. C storage capacity is the product of NPP and C residence time. C storage potential is the product of chasing time and net C pool change. Further, chasing time is jointly determined by environmental scalars for temperature and precipitation,

transfer coefficients and exit rate. C residence time is jointly determined by the environmental scalars, transfer coefficients, exit rate, and allocation coefficients. Environmental scalars can be derived from climate forcing.

The model used is the TECO model, which has been described in chapters 2 and 5. The procedure for this transient traceability analysis is as follows. We first calibrate the TECO model with GPP data for the two sites downloaded from the AmeriFlux website (available at <http://ameriflux.lbl.gov/>). We then run TECO to steady state by recycling ten years of forcing data from 1850 to 1859. Climate forcing data, including air and soil temperature, precipitation, photosynthetically active radiation, vapor-pressure deficit, and relative humidity are derived from an offline run of the Community Land Model 4.5 (CLM4.5) for both historical (1850–2005) and RCP8.5 future (2006–2100) simulations. After that, we run the model in forward simulation mode from 1850 to 2100. We output each component (X , A , ξ , K , B) in Figure 18.1 and calculate transient C storage using equations 18.1 and 18.2. We then verify the calculated transient C storage. That is, we compare direct model output of C storage with C storage calculated by the transient traceability framework.

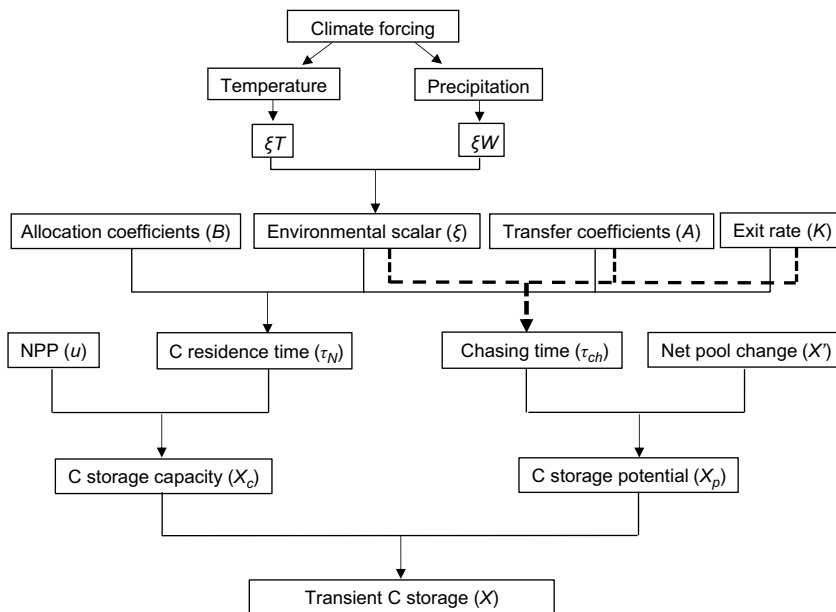


Figure 18.1. Schematic diagram of the traceability framework to analyze transient carbon storage dynamics of terrestrial ecosystems. ξ_W and ξ_T are water and temperature scalars, respectively. Dashed lines show the components that determine chasing time τ_{ch} , (adapted from Jiang et al. 2017).

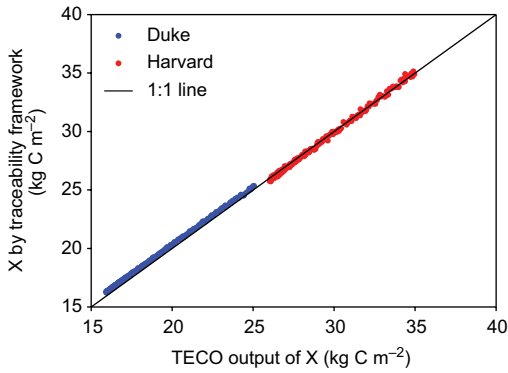


Figure 18.2. Correlation between direct model output of carbon storage (X , the sum of all carbon pools) by the Terrestrial ECOsystem (TECO) model and calculated X by the traceability framework in Duke and Harvard Forests (adapted from Jiang et al. 2017).

They are almost identical in both forests (Figure 18.2), confirming that the transient traceability framework works very well to reproduce the full model simulations.

The results for transient C storage, C storage capacity and C storage potential are shown in Figure 18.3. The trajectories of transient C storage (X , or rather, the sum of the elements [pools] of the vector X), C storage capacity, X_c , and C storage potential, X_p , over time are similar between the two ecosystems, all increasing with time. Moreover, X closely tracks X_c in both ecosystems and X_p only accounts for a very small proportion of X , which indicates that transient C storage in these two ecosystems is predominated by the maximum C storage, i.e., carbon storage capacity (X_c), while carbon storage in response to climate change is relatively

small. The most important difference between the sites is that Harvard Forest has higher X and X_c than Duke Forest (Figure 18.3a). Panel b shows the change of the three variables in the two ecosystems by the end of 2100, which are averages of the last ten years' C storage minus averages of the first ten years' values.

The components of transient C storage are shown in Figure 18.4. The two components of C storage capacity, that is, NPP and C residence time, are shown in panels a and b, respectively. NPP is similar between the two ecosystems, but residence time shows different trends. In Duke Forest, C residence time increases over time, but in Harvard Forest, C residence time decreases. Panel c shows the changes of NPP and C residence time in these two ecosystems at the end of the 21st century compared to 1850.

Environmental scalars increase in both forests (Figure 18.4d and e), which signifies steadily decreasing environmental limitations on C processes. Allocation coefficients show different trends between the two ecosystems (Figure 18.4f). In Duke Forest, b_1 (allocation to leaf) and b_3 (allocation to root) both decline with time, but b_2 (allocation to wood) increases greatly. In Harvard Forest, allocation to leaves and wood both slightly increase, but allocation to roots decreases. The substantial increase in wood allocation at Duke Forest may explain why C residence time in this ecosystem increases; wood usually has longer residence time than leaves and roots. Similarly, panel g shows the changes of allocation coefficients by the end of the 21st century in these two ecosystems.

As shown in equation 18.1 and Figure 18.1, C storage potential, X_p , is codetermined by the chasing time and net C pool change, X' , which equates

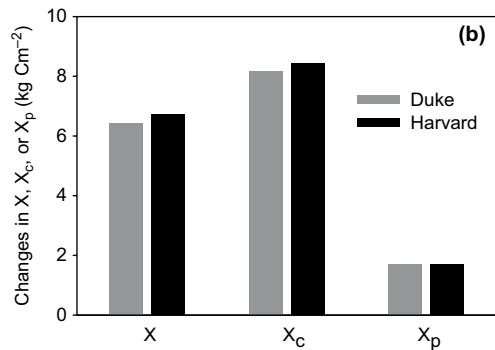
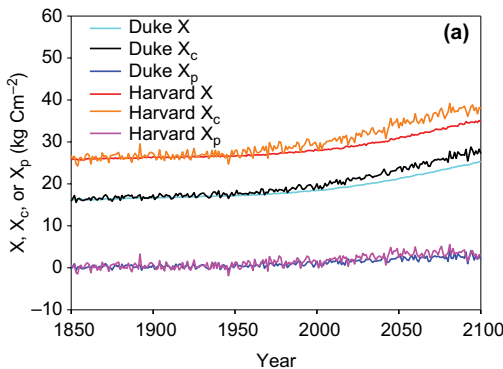


Figure 18.3. Transient total carbon storage (X , the sum of all carbon pools), carbon storage capacity (X_c), carbon storage potential (X_p) in Duke Forest and Harvard Forest and their changes by the end of the 21st century under the RCP8.5 scenario (adapted from Jiang et al. 2017).

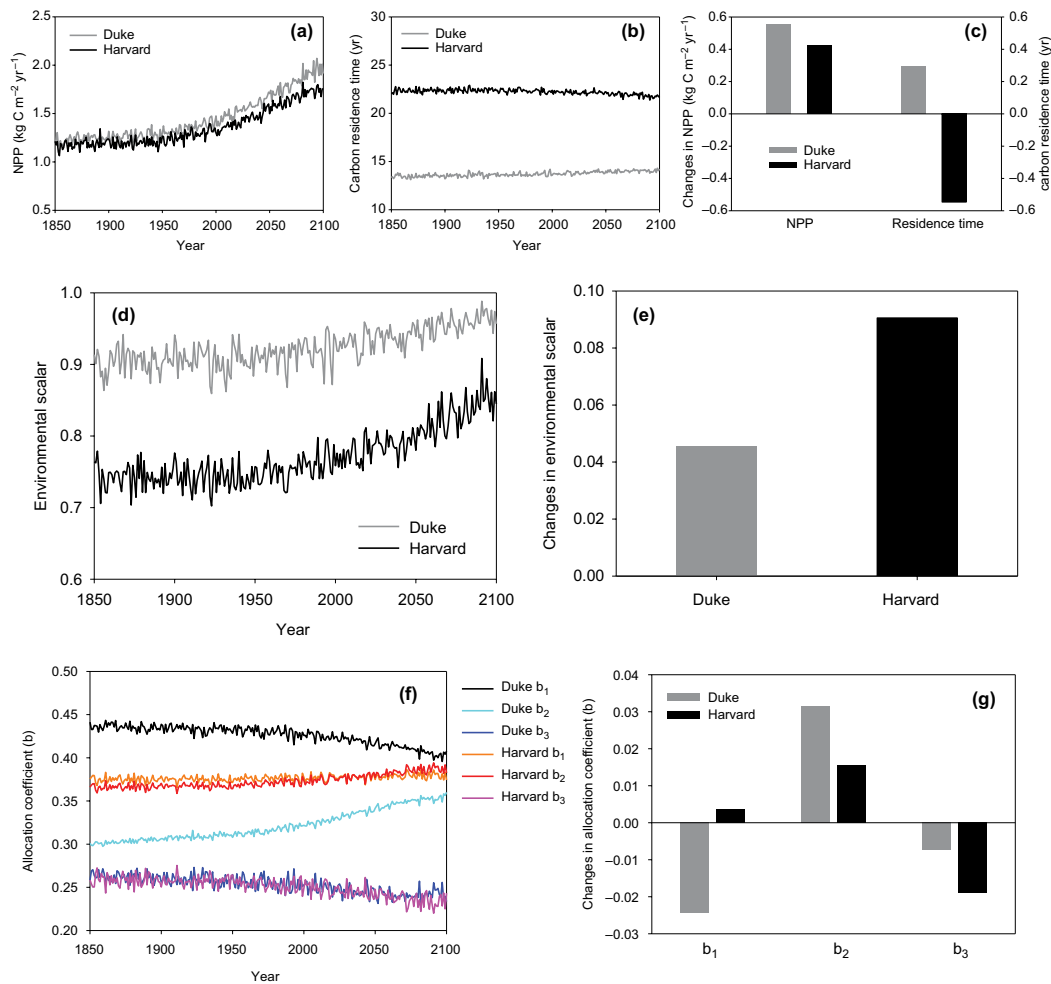


Figure 18.4. Net primary production (NPP), ecosystem carbon residence times, environmental scalars, and allocation coefficients of NPP to leaf (b_1), wood (b_2) and root (b_3) in Duke Forest and Harvard Forest, and their changes by the end of the 21st century (adapted from Jiang et al. 2017).

to NEP or NEE at ecosystem scale. Figure 18.5 shows the correlation between NEP and C storage potential in these two ecosystems. The correlation coefficients, R^2 , are high in both Duke Forest (0.80) and Harvard Forest (0.79). This indicates that X_p is mostly determined by NEP rather than chasing time. Chasing time, represented by the slopes of the linear regressions between X_p and NEP, is an indicator of approximate time needed for transient C storage to reach C storage capacity. Chasing time in Duke Forest is shorter than that in Harvard Forest. In Duke Forest, it takes approximately 19 years for the changed carbon pool to be redistributed in the network. In Harvard Forest, this time is around 28 years. Having shorter chasing time, X_p in Duke Forest is lower than that in Harvard Forest.

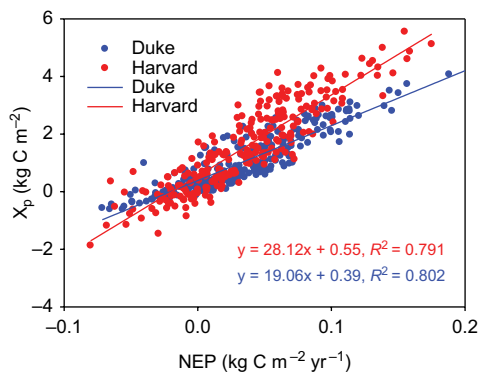


Figure 18.5. Correlation between net ecosystem production (NEP) and C storage potential (X_p) in Duke Forest and Harvard Forest (adapted from Jiang et al. 2017).

To summarize this case study, the transient traceability framework can decompose modeled transient C storage into a few traceable components. This helps us to understand the mechanisms for modeled C dynamics in response to climate change in Duke Forest and Harvard Forest. For example, the difference in carbon storage capacity (X_c) between the two ecosystems is mostly caused by the difference in inherent C residence time. In addition, the contrasting responses of C residence time to climate change between the two ecosystems can be attributed to the different responses of allocation of NPP to plant parts (leaves, wood, and roots). This application demonstrates that the traceability framework can be used to understand how and why different ecosystems respond to climate change differently. Similarly, it can also be used to address how other global change drivers (such as land use change and elevated CO_2) affect land C storage dynamics across ecosystems in simulations with ecosystem models. When applied to global land models, it can also help investigate the differences across biomes under different environmental scenarios.

TRANSIENT TRACEABILITY ANALYSIS OF LAND CARBON STORAGE IN MODEL INTERCOMPARISON PROJECTS

Another important application of the transient traceability framework is to be used in MIPs to identify sources of uncertainty and thereby help improve model development. For example, Zhou

et al. (2018) applied the transient traceability framework but with a modified method to diagnose the causes of uncertainty in modeled global annual land carbon storage within and across three MIPs: CMIP5, TRENDY, and M5TMIP.

The transient traceability analysis of carbon storage at Duke Forest and Harvard Forest introduced above is called authentic transient traceability analysis, that is, the modeled differences of C storage among ecosystems or among models can be traced back to differences in respective components in equation 18.1, and finally to individual processes or parameters in the models. However, the application of authentic transient traceability analysis to MIPs requires much time to figure out the structures and parameterizations of all involved models in order to recode each in matrix form for a thorough model intercomparison. Due to the challenge in acquiring all the details of the involved models, in their analysis, Zhou et al. (2018) applied the transient traceability framework in combination with another technique, variance decomposition, to identify the underlying causes for uncertainty in simulated land carbon storage within and across three MIPs. We called this kind of analysis post-MIP transient traceability analysis to be distinguished from the authentic transient traceability analysis. Figure 18.6 shows the schematic diagram of this transient traceability analysis within and among the three MIPs.

In this post-MIP traceability analysis, they found that models differ a lot in the global

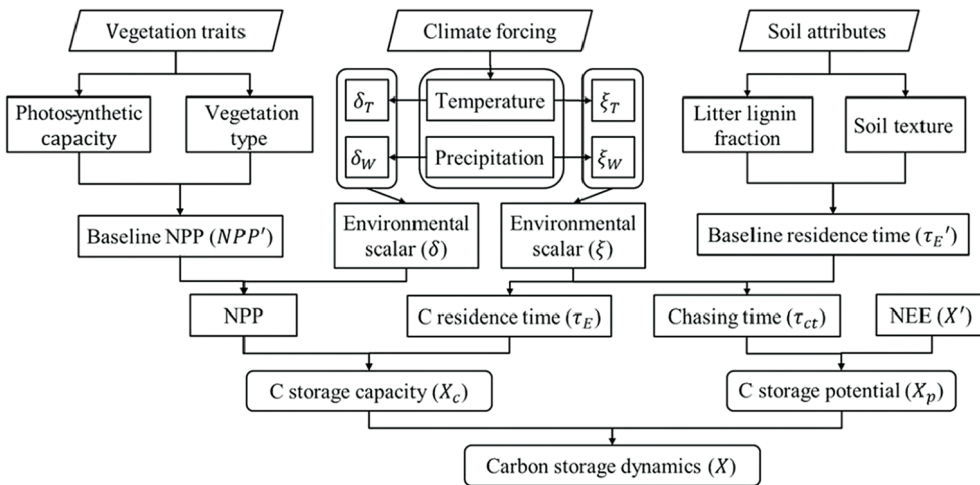


Figure 18.6. Schematic diagram of the transient traceability framework by Zhou et al. (2018) © American Meteorological Society. Used with permission. This framework traces the modeled transient carbon storage dynamics to carbon residence time, NPP, carbon storage potential, and their source factors.

annual carbon residence time, NPP, and carbon storage potential (Figure 18.7a–c). Usually, within each model, relative year to year variation in carbon residence time is much smaller than that of NPP. In addition, NPP in those models with a coupled nitrogen cycle (e.g., BNU-ESM, CESM1(BGC), and NorESM1-Me in CMIP5, and CLM4, CLM4VIC, ISAM, and DLEM in MSTMIP) is lower than other ESMs without a coupled

nitrogen cycle. Carbon residence time and NPP show smaller variations across the nine dynamic global vegetation models in TRENDY in comparison to those models in CMIP5 and MSTMIP. As a result, the variations in carbon storage capacity in TRENDY are likewise not as large as in CMIP5 and MSTMIP.

The global annual carbon storage and carbon storage capacity also vary considerably across the

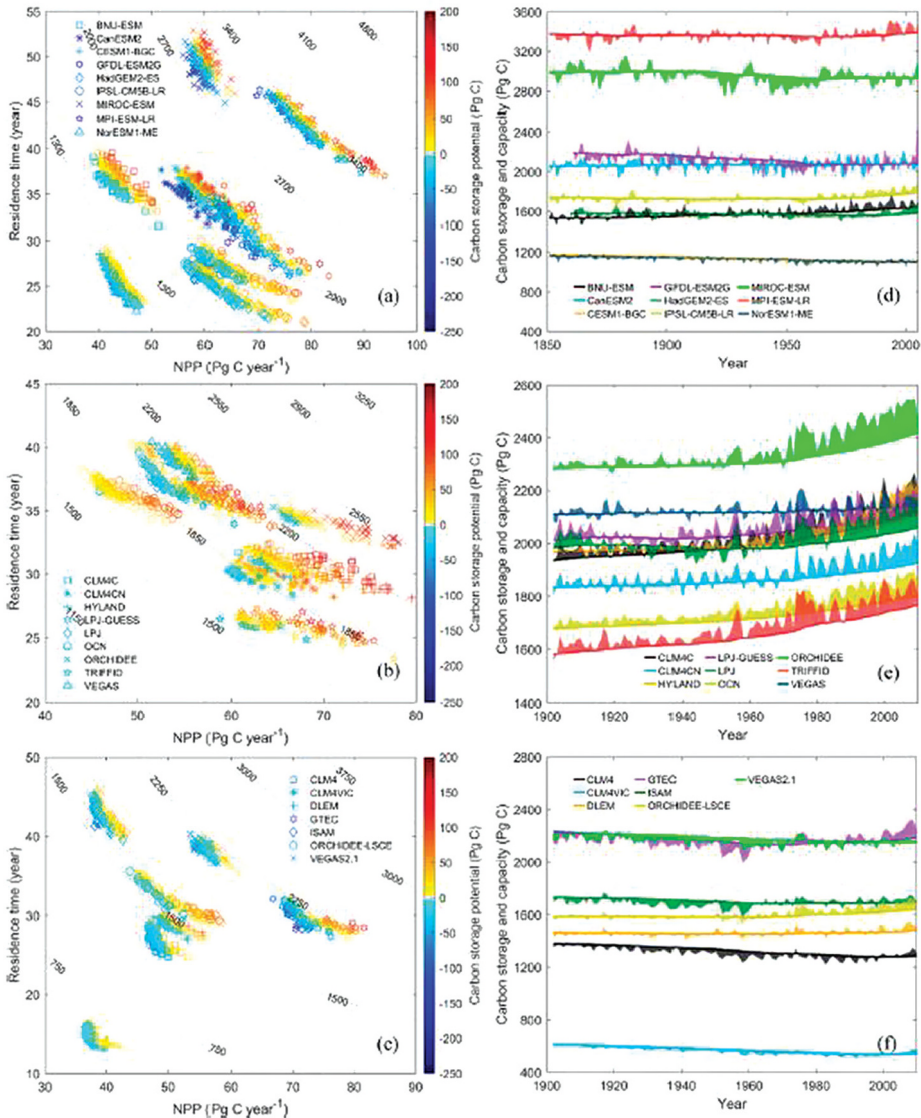


Figure 18.7. The 3D model output space (carbon residence time, NPP, and carbon storage potential), and time series of annual carbon storage (solid lines) with the shaded outlines indicating the year-to-year fluctuations due to changes in carbon storage capacity for the models in CMIP5 (a and d), TRENDY (b and e), and MSTMIP (c and f). The points in (a)–(c) represent the global annual values for the three variables. The contour lines in (a)–(c) represent the carbon storage capacity. Shading in (d)–(f) shows the values of the carbon storage potential for the models (positive above the solid lines, and negative below the solid lines). Reproduced from Zhou et al. 2018. © American Meteorological Society. Used with permission.

models and the temporal dynamics of carbon storage and carbon storage capacity of the models are highly diverse (Figure 18.7d–f). The large range of carbon storage across those models is closely related to that of carbon storage capacity. The interannual trends of carbon storage in the models of the three MIPs are mainly affected by the carbon storage potential, because the sign and magnitude of carbon storage potential determine the direction

and rate of carbon storage change, respectively. The interannual variability of carbon storage in TRENDY is much smaller than that in CMIP5 and MsTMIP.

Carbon residence time and NPP are further traced to their baseline values and environmental scalars. The differences in carbon residence time (or NPP) across the models are codetermined by the differences in baseline carbon residence time (or baseline NPP) and the environmental scalars

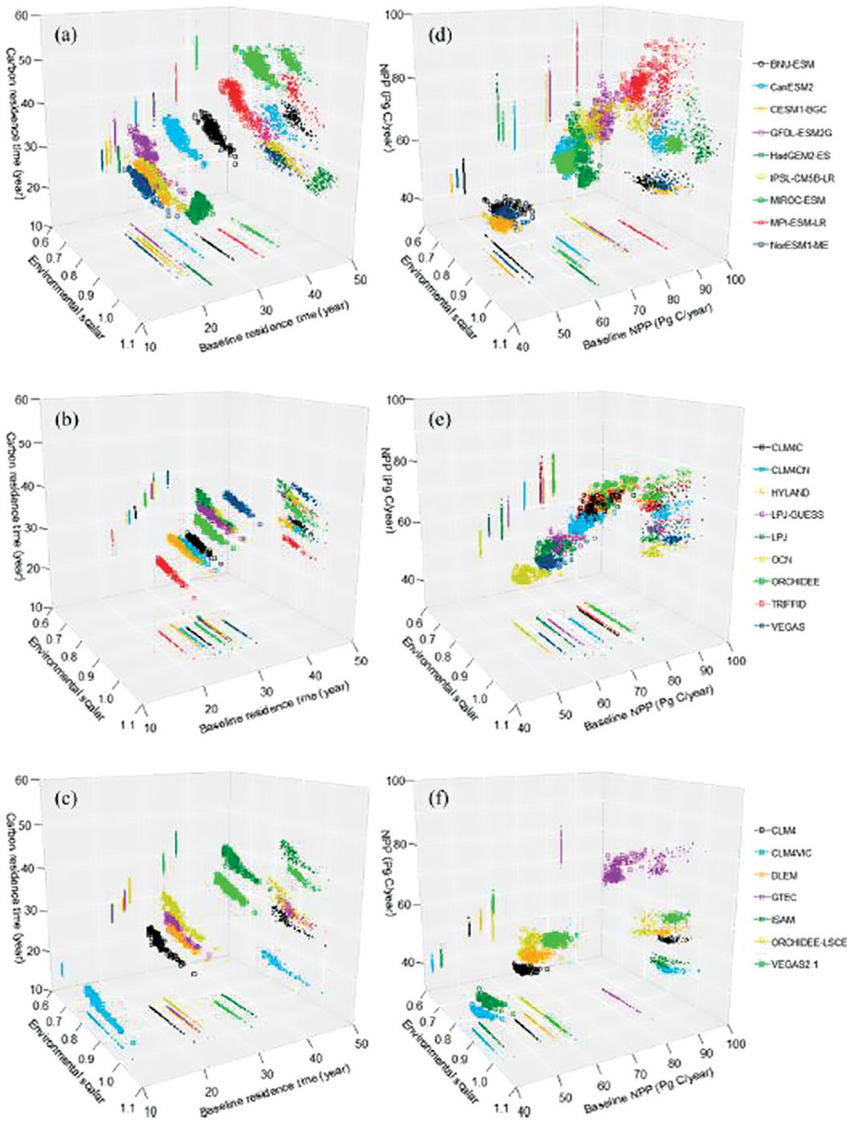


Figure 18.8. Decomposition of the carbon residence time into the baseline carbon residence time and the environmental scalar and decomposition of annual NPP into the baseline NPP and the environmental scalar for CMIP5 (a and d), TRENDY (b and e), and MsTMIP (c and f). The environmental scalar is a product of the temperature and water scalars, which convert the baseline carbon residence time and baseline NPP into their actual values (from Zhou et al. 2018 © American Meteorological Society. Used with permission.).

(Figure 18.8). The baseline carbon residence time and baseline NPP among the models in the three MIPs can differ as much as threefold. In contrast, the environmental scalars are more convergent among the models for both carbon residence time and NPP. That indicates that the large differences in carbon residence time and NPP across models are due mainly to the differences in their baseline carbon residence time and baseline NPP, not much being caused by environmental scalars.

Finally, by adopting a variance decomposition method, Zhou et al. (2018) quantified the relative contribution of each component to the simulated global annual carbon storage for each MIP and for all the three MIPs together. The results revealed that variations in transient carbon storage are dominated by carbon residence time and NPP, and carbon storage potential only contributes less than 1% (Figure 18.9). Moreover, the baseline carbon residence time and baseline NPP contribute more than

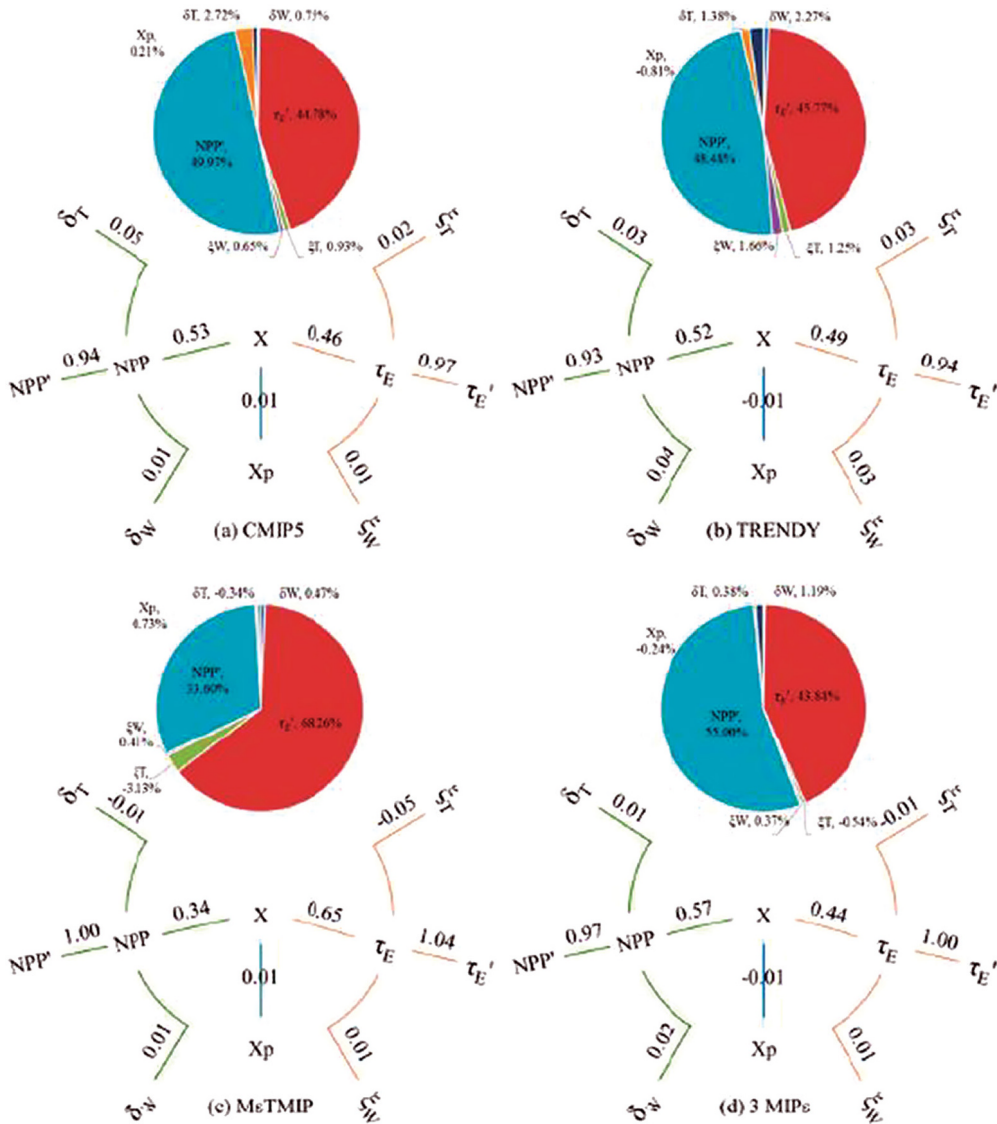


Figure 18.9. Variance decomposition of the carbon storage based on global annual data from models in the three MIPs. First, the variation of the carbon storage X is decomposed into that of the carbon residence time τ_E , NPP, and the carbon storage potential X_p . Second, variations of the carbon residence time and NPP are decomposed into their baseline values (τ_E' and NPP') and the temperature (ξ_T and δ_T) and water (ξ_W and δ_W) scalars. Positive/negative values mean positive/negative contributions of the variables to the variation of carbon storage (from Zhou et al. 2018 © American Meteorological Society. Used with permission.).

90% to the variations in carbon residence time and NPP, respectively. In contrast, the contributions by temperature and water scalars to the variations in the carbon residence time and NPP are both less than 5%. As a consequence, the variations in simulated transient carbon storage across the models can be primarily attributed to the differences in models' baseline carbon residence time and baseline NPP.

The post-MIP approach adopted by Zhou et al. (2018) is a novel approach that provides an alternative way to understand the causes of the uncertainty of multiple models when authentic traceability analysis is not able to be realized due to the effort required to accommodate detailed information on the original models. Such post-MIP traceability analysis can be automatically run with a public platform, TraceME (v1.0), which is an online traceability analysis system for model evaluation on land carbon dynamics (Zhou et al. 2021). While post-MIP traceability analysis offers useful insight, it would be helpful to apply the authentic traceability analysis, as in the first case of this chapter, to MIPs to help identify the specific model components and assumptions that dominate model uncertainty and focus attention on those issues in need of closer scrutiny to improve model behavior.

After identification of the causes by which the models differ in their behavior, modelers can then use observational data to determine which models are more accurate than others in representing the actual processes. This is the realm of benchmark analysis, which will be introduced in detail in chapter 19. In this way, model performance can be greatly improved towards more realistic projections.

SUMMARY

This chapter has demonstrated how the transient traceability framework can be applied to address

different scientific questions with two case studies. This recently developed framework has the potential to be used more broadly in ways similar to the overview of steady-state traceability analysis of land carbon storage in chapter 17. The ultimate goal of the transient traceability framework is to enhance our understanding of how terrestrial ecosystems respond to various environmental changes and to better incorporate such understanding in models to predict the future status of land carbon storage.

SUGGESTED READING

- Jiang LF, Shi Z, Xia JY, Liang JY, Lu XJ, Wang Y, Luo YQ (2017) Transient traceability analysis of land carbon storage dynamics: procedures and its application to two forest ecosystems. *J Adv Model Earth Syst* 9:2822–2835.
- Zhou S, Liang JY, Lu XJ et al. (2018) Sources of uncertainty in modeled land carbon storage within and across three MIPs: Diagnosis with three new techniques. *J Clim* 31:2833–2851.

QUIZZES

1. Is transient C storage determined by C storage capacity and C storage potential? Why/why not?
2. Is carbon storage potential always positive? Why/why not?
3. Carbon storage potential is co-determined by:
 - Carbon residence time
 - Chasing time
 - Net C pool change
 - NPP
4. What scientific questions do you think the transient traceability framework can address? How can it be applied to this purpose?

CHAPTER NINETEEN

Benchmark Analysis

Yiqi Luo

Cornell University, Ithaca, USA

Forrest M. Hoffman

Oak Ridge National Laboratory, Oak Ridge, USA

CONTENTS

Introduction / 157
Aspects of Land Models to be Evaluated / 158
Reference Data Sets as Benchmarks / 159
Benchmarking Metrics / 160
Performance of Three CLM Versions and Future Improvements / 160
Conclusions / 162
Suggested Reading / 162
Quizzes / 162

Tremendous progress has been achieved in the development of land models and their inclusion in Earth system models (ESMs). However, we still have very limited knowledge on the performance skills of these land models. This chapter introduces benchmark analysis, which is a procedure to measure performance of models against a set of defined standards. The benchmark analysis includes: (1) defining targeted aspects of model performance to be evaluated; (2) testing model performance in comparison with a set of benchmarks; (3) measuring model performance skill through quantitative metrics; and (4) evaluating model performance and offering suggestions for future model improvement.

INTRODUCTION

Over the past decades, tremendous progress has been achieved in the development of land models and their inclusion in Earth system models (ESMs). State-of-the-art land models now account for biophysical processes (exchanges of water and energy)

and biogeochemical cycles of carbon, nitrogen, and trace gases. They also simulate vegetation dynamics and disturbances. When coupled as components in ESMs, land models now allow simulation of land-atmosphere biophysical interactions and climate-carbon feedbacks. These models are now widely used for policy-relevant assessment of climate change and its impact on ecosystems or terrestrial resources. However, there is still very limited knowledge of the performance skills of these land models, especially when embedded in ESMs. Quantifying the performance skills of land models would promote confidence in their predictions of future states of ecosystems and climate, and identify those models whose predictions are more likely to be accurate, where ensemble members diverge.

Model performance has traditionally been evaluated via comparison with observed data sets. ‘Validation’ by plotting model data side-by-side with observed data, or computing mismatch metrics such as root-mean-square-error, is traditionally the most common approach to model evaluation (Oreskes, 2003; Rykiel, 1996; see also

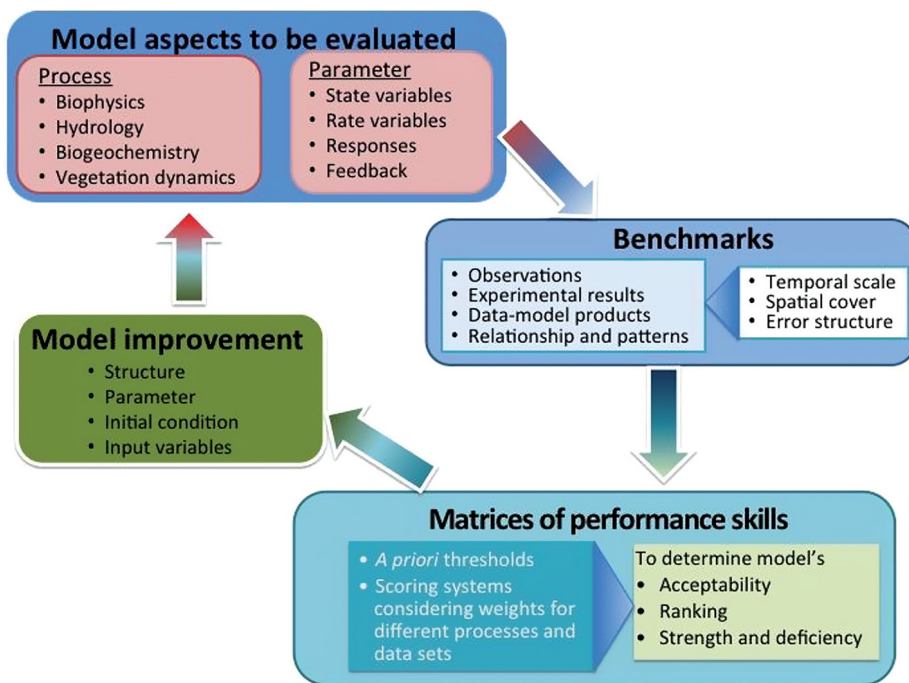


Figure 19.1. Schematic diagram of the benchmarking framework for evaluating land models. The framework includes four major components: (1) defining model aspects to be evaluated, (2) selecting benchmarks as standardized references to test models, (3) developing a scoring system to measure model performance skills, and (4) stimulating model improvement. (Adopted from Luo et al., 2012).

Chapter 2). However, a land model typically simulates hundreds of biophysical, biogeochemical, and ecological processes at regional and global scales over hundreds of years. It would be unrealistic to undertake validation of so many processes at all spatial and temporal scales, even if observations were available. The complex behavior of these interacting processes can be realistically understood only if we holistically assess land models and their major components. Benchmark analysis is an approach that has been recently developed to evaluate the performance of land models.

Benchmark analysis is a standardized evaluation of one system's performance against defined reference data (i.e., benchmarks) that can be used to diagnose the system's strengths and deficiencies for future improvement (Luo et al., 2012). Benchmark analysis has been recently applied to evaluate land models against observations (Collier et al., 2018). A benchmark analysis has four elements: (1) targeted aspects of model performance to be evaluated; (2) benchmarks as defined reference data to evaluate model performance; (3) a scoring system of metrics to measure relative

performance among models; and (4) evaluated performance of models and future improvement (Figure 19.1).

ASPECTS OF LAND MODELS TO BE EVALUATED

Land models typically simulate many processes. Although individual studies may assess only a few aspects of model performance, a comprehensive benchmark analysis is required to evaluate all these major components when land models are integrated with ESMs. The performance of a model should be evaluated for its baseline simulations over broad spatial and temporal scales, and modeled responses of land processes to global change.

The baseline state for biogeochemical cycles includes simulated global totals, spatial distributions, and temporal dynamics of gross primary production, net primary production, vegetation and soil carbon stocks, ecosystem respiration, litter production, litter mass, and net ecosystem production. For example, the International Land Model Benchmarking (ILAMB) project evaluated biomass, burned area, gross primary productivity, leaf

area index, global net ecosystem carbon balance, net ecosystem exchange, ecosystem respiration, and soil carbon (Collier et al., 2018).

To reliably predict future states of ecosystems under a changing environment, land models have to realistically simulate responses of land processes to disturbances and global change. Major global change factors include rising atmospheric CO₂ concentration, increasing land use and surface air temperature, altered precipitation amounts and patterns, and changing nitrogen (N) deposition. The direct effects of these global change factors are relatively easily benchmarked since we have direct knowledge of how ecosystems respond to rising atmospheric CO₂ concentration, increasing temperature, altered precipitation, and changing nitrogen deposition. However, indirect effects of these factors on ecosystem carbon processes are not well understood, although many field experiments have been conducted. Thus, it is more difficult to benchmark model performance in predicting future states of ecosystems.

REFERENCE DATA SETS AS BENCHMARKS

A comprehensive benchmarking analysis usually uses a set of benchmarks, against which land model performance can be evaluated (Table 19.1). Benchmarks could consist of direct observations, results from manipulative experiments, data-model products, or data-derived functional relationships. Direct observations and experimental results are generally accepted to be the most reliable benchmarks for model performance and are typically referred to as reference data. Reference data that are often used for benchmarking biogeochemical cycle models include global data products of gross primary production (GPP), net primary production (NPP), soil respiration, ecosystem respiration, plant biomass, and soil carbon. When they are used in a benchmarking analysis, reference data sets are usually assessed and weighted for their degree of certainty, scale appropriateness, and overall importance of the constraint or process to model predictions (Collier et al., 2018). The ILAMB project evaluates eight variables using a variety of reference data as listed in Table 19.1.

Land models can also be evaluated on their simulated variable-to-variable relationships in comparison with relationships in observations. For example, model representations of the relationships that GPP exhibits with precipitation,

TABLE 19.1
Reference data sets used to measure ecosystem and carbon cycle performance

Variables	Reference data sets	Description
Biomass	Tropical (Saatchi et al., 2011)	forest carbon stocks in tropical regions across three continents
	NBCD2000 (Kelldorfer et al., 2013)	aboveground biomass and carbon baseline data in north America
	USForest (Blackard et al., 2008)	U.S. forest biomass
Burned area	GFED4S (Giglio et al., 2010)	variability and long-term trends in burned area
GPP	Fluxnet (Lasslop et al., 2010)	net ecosystem exchange, photosynthesis, and respiration
Leaf area index	AVHRR (Myneni et al., 1997)	global land cover, LAI and FPAR
	MODIS (De Kauwe et al., 2011)	leaf area index product for a region of mixed coniferous forest
Global NECB	GCP (Le Quéré et al., 2016)	global carbon budget 2016
Net ecosystem exchange	Fluxnet (Lasslop et al., 2010)	net ecosystem exchange, photosynthesis, and respiration
Ecosystem respiration	Fluxnet (Lasslop et al., 2010)	net ecosystem exchange, photosynthesis, and respiration
Soil carbon	HWSD (Todd-Brown et al., 2013)	Harmonized World Soil Data
	NCSCDV22 (Hugelius et al., 2013)	organic carbon storage to 3m depth in soils of the northern circumpolar permafrost region.

NECB = net ecosystem carbon balance

evapotranspiration, and temperature are often assessed. Such variable-to-variable relationships are quantified over a time period from reference data sets and used as benchmarks for the relationships diagnosed in models. This approach is particularly effective to understand the consistency between the observed and simulated sensitivity of ecosystem responses to climate change.

BENCHMARKING METRICS

A comprehensive benchmarking study usually uses a suite of metrics across several variables to holistically assess model performance at the relevant spatial and temporal scales. Many statistical measures are available to quantify mismatches between multiple modeled and observed variables. Five metrics were developed for ILAMB to evaluate model performance. The five metrics are to measure bias, root-mean-square-error (RMSE), phase shift, interannual variability, and spatial distributions (Collier et al., 2018).

The bias measures differences between the mean value of the reference data and that of the model over the same time period and the same spatial area. For example, the bias of gross primary productivity between the reference data and the model (e.g., Community Land Model version 4.5, CLM4.5) is calculated between their respective means in each grid cell where both reference data and modeled values are available. To account for the bias due to the variability at any given spatial location, the bias is nondimensionalized as a relative error to measure the bias score.

RMSE is computed as the square root of the mean square error between modeled values and the reference data over a time period. The RMSE is normalized by the centralized RMSE of the reference data set to get a relative error as a score. By scoring the centralized RMSE, the bias is removed from the RMSE, allowing the RMSE score to be focused on an orthogonal aspect of model performance.

The phase shift is evaluated for the annual cycle of many data sets that have monthly variability by comparing the timing of the maximum of the annual cycle of the variable at each spatial cell across the time period of the reference data set. The phase shift is calculated as the difference between the reference and model data sets by subtracting their respective maximum values in days.

The interannual variability in model simulations is evaluated by removing the annual cycle from both the reference data and the model. A score is then computed as a function of their differences over space.

The spatial distribution of any time-averaged variable is evaluated by computing the standard deviation of modeled values over space normalized by the standard deviation of the reference data. The spatial correlation is also calculated for the period mean values of reference data and modeled values. A score is assigned by the penalty for large deviation of the normalized standard deviations and the spatial correlation from a value of 1.

The overall score for a given variable and data product is a weighted sum of the five metrics, producing a single scalar score for each variable for every model or model version. Readers who are interested in details of these metrics may study the paper by Collier et al. (2018).

PERFORMANCE OF THREE CLM VERSIONS AND FUTURE IMPROVEMENTS

The metrics for bias, RMSE, seasonal cycle phase, spatial distribution, interannual variability, and variable-to-variable assessments were applied to evaluate three CLM versions (CLM4 vs. CLM4.5 vs. CLM5) under two forcing data sets (GSWP3v1 vs. CRUNCEPv7) (Lawrence et al., 2019). The quality of the simulations across model generations was found to be generally improving. CLM5 outperforms CLM4 for the majority of assessed variables (Figure 19.2). The improvements from CLM4.5 to CLM5 were relatively subtle in that several variables show improvement (e.g., biomass, burned area, LAI, net ecosystem carbon balance, net ecosystem exchange, and ecosystem respiration) but others show degradation (e.g., soil carbon).

The functional relationships were also assessed between two variables (e.g., precipitation vs. GPP or LAI) (Figure 19.3). CLM5 performed better than CLM4 or CLM4.5 for the relationships between GPP and climate variables. However, the relationship between GPP and surface air temperature slightly degraded from CLM4.5 to CLM5.

The ILAMB benchmark analysis provides some insights into model development. An improvement or degradation trend between two CLM versions can result from a mix of scores for individual metrics. The degradation in the simulations of

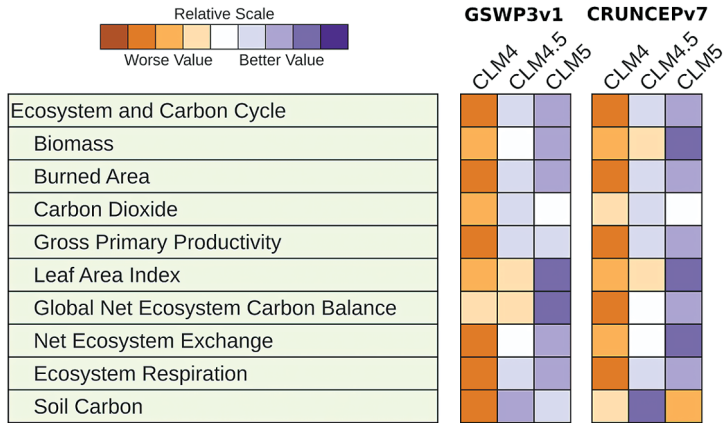


Figure 19.2. Evaluation of performance of CLM4, CLM4.5, and CLM5 under two sets of forcing, GSWP3v1 and CRUNCEPv7. A stoplight color scheme is used to indicate aggregate performance for each model by variable. (Adopted from Lawrence et al., 2019).

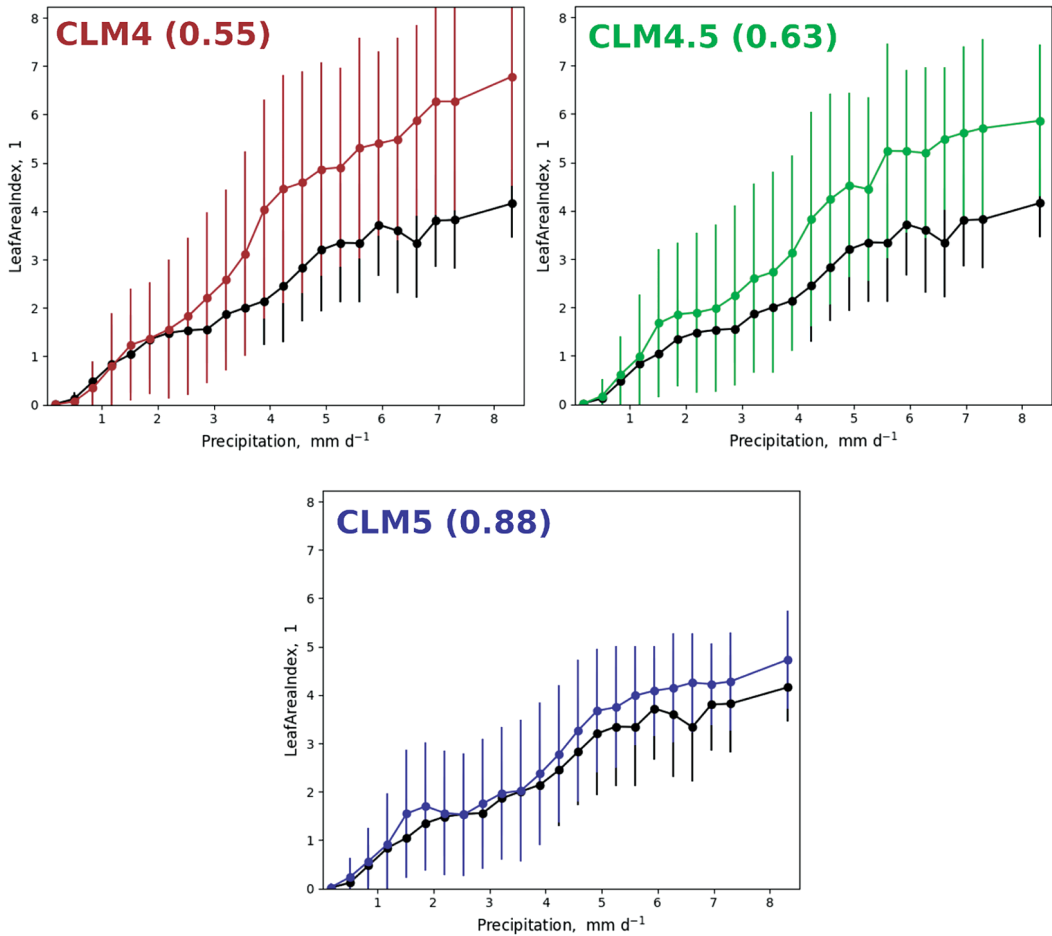


Figure 19.3. Variable-to-variable comparison between annual precipitation and LAI for CLM4, CLM4.5, and CLM5 under the GSWP3v1 forcing. The black line is the observationally derived relationship. Error bars indicate the ± 1 standard deviation of LAI for all grid cells that lie within that precipitation bin. Values in parentheses are the scores for that comparison. (Adopted from Lawrence et al., 2019).

soil carbon stocks from CLM4.5 to CLM5 is partially due to high uncertainty in the observational estimates. Another metric evaluates the models against apparent soil carbon turnover time, showing an improvement from CLM4.5 to CLM5. The disagreement between two metrics of soil carbon may suggest the need for future improvement of observationally constrained estimates.

Model performance depends on three elements: model structure, parameterization, and forcing (see Chapters 21 and 33). The model structure that simulates soil carbon dynamics in CLM is primarily based on first-order kinetics. Although this model structure has been questioned, almost all data sets from studies of litter decomposition and soil incubation suggest the structure may be highly reliable (Chapter 1). Model parameterization is likely the main cause of the model-data mismatch. Chapter 37 discusses methods to improve model parameterizations of CLM5 to improve model performance.

CONCLUSIONS

A four-component benchmark analysis was outlined: (1) identification of aspects of models to be evaluated; (2) selection of benchmarks as standardized references to evaluate models; (3) a scoring system to measure model performance skills; and (4) evaluation of model performance to inform model improvement. The International Land Model Benchmarking (ILAMB) project has developed an open-source model benchmarking software package to score model performance. ILAMB has developed a suite of reference data sets as benchmarks, five metrics plus variable-to-variable relationships as the scoring system to evaluate models or model versions. The ILAMB package has been applied to

perform comprehensive model assessment across a wide range of land variables. Such benchmark analysis offers insights into strengths and weaknesses of different models or model versions for identifying future improvements.

SUGGESTED READING

- Collier, N., F. M. Hoffman, D. M. Lawrence, G. Keppel-Aleks, C. D. Koven, W. J. Riley, M. Mu, and J. T. Randerson (2018), The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation, *J. Adv. Model. Earth Syst.*, 10(11):2731–2754, doi:10.1029/2018MS001354.
- Luo, Y. Q., J. T. Randerson, G. Abramowitz, C. Bacour, E. Blyth, N. Carvalhais, P. Ciais, D. Dalmonech, J. B. Fisher, R. Fisher, P. Friedlingstein, K. Hibbard, F. Hoffman, D. Huntzinger, C. D. Jones, C. Koven, D. Lawrence, D. J. Li, M. Mahecha, S. L. Niu, R. Norby, S. L. Piao, X. Qi, P. Peylin, I. C. Prentice, W. Riley, M. Reichstein, C. Schwalm, Y. P. Wang, J. Y. Xia, S. Zaehle, and X. H. Zhou (2012), A framework for benchmarking land models, *Biogeosci.*, 9(10):3857–3874, doi:10.5194/bg-9-3857-2012.

QUIZZES

1. What are the similarities and differences between model validation and benchmark analysis?
2. How does benchmark analysis evaluate model performance?
3. What variables in carbon cycle models would you choose to be evaluated by a benchmark analysis?
4. What data sets do you think would be important to be used as benchmarks to evaluate models?
5. What five metrics does the ILAMB package use to score model performance?

CHAPTER TWENTY

Practice 5

TRACEABILITY ANALYSIS FOR EVALUATING TERRESTRIAL CARBON CYCLE MODELS

Jiayang Xia and Jian Zhou

East China Normal University, Shanghai, China

CONTENT

Introduction / 163

The practice is designed to help you learn traceability analysis to identify sources of model uncertainty in predicting terrestrial carbon (C) storage. All practices are performed in the training software CarboTrain. With this tool, you will apply traceability analysis to simulation results from a matrix form model (called authentic traceability analysis) and to model intercomparison projects (MIP) without matrix models (i.e., post-MIP traceability analysis). The authentic traceability analysis will show you how simulation results from a matrix model are explained by traceable components over space and among biomes. The post-MIP traceability analysis can help you understand the sources of uncertainty among different models.

INTRODUCTION

Traceability analysis provides an approach to divide the simulated land carbon dynamic into several traceable components, such as carbon (C) storage capacity, gross primary productivity (GPP), C residence time, and environmental scalars. This practice offers two exercises. The first uses authentic

traceability analysis in which simulation results by the CABLE matrix model are explained by traceable components hierarchically over space and among biomes. The authentic traceability analysis for the first exercise is described in detail in Chapter 17, based on the study by Xia et al. (2013). The second exercise uses post-MIP traceability analysis to attribute variations in modeled land C storage among three CMIP6 models (i.e., CESM2, CNRM-ESM2-1, and IPSL-CM6A-LR) over 1980–2000 to different sources. The post-MIP traceability analysis is described in detail in Chapter 18. More information is available in papers by Zhou et al. (2018, 2021). The authentic traceability analysis can pinpoint model uncertainty to individual processes and/or specific parameter values but requires matrix models. In comparison, the post-MIP traceability analysis can be applied to any modeling results to understand sources of uncertainty but may not be able to trace uncertainty to specific processes and/or parameters.

The exercises are performed in the training software CarboTrain by selecting Unit 5 and Exercises 1–2. Click *Run Exercise* to generate the figures.

EXERCISE 1: Authentic traceability analysis

This exercise helps you learn to do traceability analysis with a matrix model, which derives from the Community Atmosphere-Biosphere-Land Exchange (CABLE) model. CABLE is one of the most widely used global land models for simulating terrestrial biogeochemical and biophysical processes. The C cycle diagram of the CABLE model is shown in Chapter 14, Figure 14.1. CABLE has nine carbon pools, including the plant pools (leaf, root and wood), the litter pools (metabolic and structural litter as well as coarse woody debris) and three soil pools (microbial biomass, slow and passive soil organic matter). The matrix form of the CABLE model we will use here was derived by Xia et al. (2013). In this exercise, the spatial resolution is

Here you can use the CarboTrain software to do this exercise. In the main window of CarboTrain (Figure 20.1), choose unit 5 and exercise 1, and then set the output path of the running results by clicking the “**Set Output Folder**” button. After you have done all the above steps, click ‘**Run Exercise**’ to start the exercise.

A pop-up window will show up with the message “Task submitted!” after you click the ‘Run Exercise’ button as shown in Figure 20.2. After clicking ‘OK’, the task will run automatically in your computer.

The command window shows the running processes of the task and you can see which step of the program is complete as shown in Figure 20.3. When the task is finished, another pop-up window will show up with the message



Figure 20.1. Steps to run Exercise 1.

$1 \times 1^\circ$. The land grid cells are categorized into nine biomes including Evergreen Needleleaf Forest (ENF), Evergreen Broadleaf Forest (EBF), Deciduous Needleleaf Forest (DNF), Deciduous Broadleaf Forest (DBF), C_3 Grassland (C3G), C_4 Grassland (C4G), Tundra, and Barren/sparse vegetation (Barren).

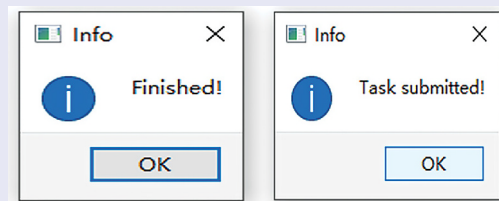


Figure 20.2. Tips at the beginning and end of the task.

```

###==== Start to draw carbon storage ...
###==== End drawing carbon storage.
###==== Start to draw npp and residence time ...
###==== End drawing npp and residence time.
###==== Start to draw the decomposition of residence time ...
###==== End drawing the decomposition of residence time.
###==== Start to draw environmental scalars ...
###==== End ===###

```

Figure 20.3. Prompt in CMD interface when the program is running.

<< unit5-output > output-ex1 > unit5_exercise1





Name	Date modified	Type
 .vscode	4/26/2021 10:50 AM	File folder
 dataSources	5/4/2021 4:49 PM	File folder
 output_figs	4/26/2021 10:50 AM	File folder
 practisce_cable	12/25/2020 9:45 AM	Python File

Figure 20.4. Output results after the task is complete.

“Finished!” as shown in Figure 20.2. The software will generate the running results to the output path you set before.

Figure 20.4 shows an example of the output results after the task is completed. Enter the output path you set before, you will see a directory named `unit5_exercise1`. Enter this directory and find the files as shown in Figure 20.4.

There are four data files in the folder `unit5_exercise1/dataSource`. The simulated global distribution of total ecosystem C storage capacity by the CABLE model is recorded in `data_global_ctot.csv`. The ‘nan’ values represent ocean grids or non-vegetated lands. The global map of the total ecosystem C storage capacity is shown in Figure 20.5a.

The file `data_npp_resTime.xlsx` provides the simulated net primary productivity (NPP) and ecosystem C residence time in each land grid cell. Figure 20.5b shows how NPP and ecosystem C residence time together determine the spatial difference in ecosystem C storage capacity in the CABLE model. For example, ENF has an intermediate NPP ($0.39 \text{ kg C m}^{-2} \text{ yr}^{-1}$) and a relatively long C residence time (86.4 years), leading to the largest total ecosystem carbon storage capacity (34.1 kg C m^{-2}) among the nine biomes. By contrast, Tundra has a small ecosystem C storage capacity

(8.7 kg C m^{-2}) due to the low NPP ($0.1 \text{ kg C m}^{-2} \text{ yr}^{-1}$), though its C residence time is long (141.2 years).

The file `residence_components.xlsx` contains simulated results of total ecosystem C residence time, baseline residence time, and environmental scalar in each land grid cell. Figure 20.5c shows how baseline C residence time and environmental scalar jointly determine the global distribution of ecosystem C residence time in the CABLE model. It is clear that the order of ecosystem C residence time among biomes is different from that of baseline C residence time. In CABLE, ecosystem C residence time changes with the biome types as DNF (163.3 years) > Tundra (141.2 years) > ENF (86.4 years) > Shrubland (52.6 years) > DBF (33.3 years) > C3G (26.6 years) > EBF (26.3 years) > Barren (20.4 years) > C4G (17.5 years).

The `environmentalScalars.xlsx` file provides the mean annual precipitation, mean annual temperature, water scalar, and temperature scalar in each land grid cell. As shown in Figure 20.5d–e, the environmental scalars link the climate forcings directly to terrestrial C cycle processes in the CABLE model. Figure 20.5e shows that the temperature scalar varies systematically among biomes in the CABLE model, whereas the mean water scalar is

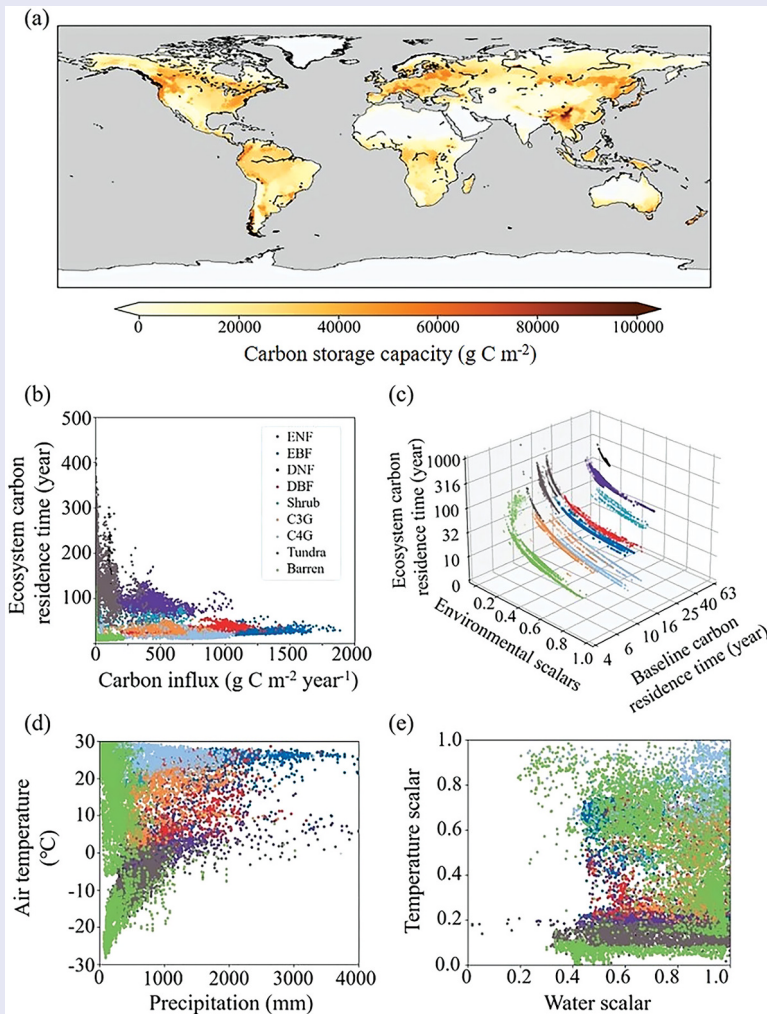


Figure 20.5. Simulated spatial distribution of terrestrial carbon storage capacity and its traceable components by the CABLE model. (a) Spatial distribution of total ecosystem C storage capacity; (b) determination of the ecosystem carbon storage capacity by NPP and ecosystem residence time in various biomes; (c) dependence of ecosystem C residence time on baseline C residence time and the environmental scalar in various biomes; (d) global distribution of major biomes in relation to annual temperature and precipitation; (e) global distribution of major biomes in relation to water and temperature scalars.

distributed in a narrow range from 0.65 in EBF to 0.87 in DNF.

The figures of this exercise can be found in `unit5_exercise1/output_figs`. More details about the CABLE model and simulations in this exercise are provided in Chapter 17.

QUESTIONS:

1 Which environmental factor, water or temperature, contributes more to the difference

in ecosystem C residence time among biomes in the CABLE model? Why?

2 Which biome has the longest baseline C residence time? Why?

3 Would you expect the incorporation of nitrogen cycling to influence the simulated ecosystem C storage compared to the C-only model? How can this question be explored by applying the authentic traceability analysis with the CABLE model?

EXERCISE 2: Post-MIP traceability analysis

In this exercise we will practice performing uncertainty analysis of carbon cycle modeling after simulations were done without models being converted to matrix equations (Chapter 18). Because the outputs of transient land C storage rather than long-term steady state carbon storage (i.e., carbon storage capacity) are available from CMIP6 models, we compare the simulation results of the three CMIP6 models over 1980–2000. Here you also use the CarboTrain software to do this exercise. Launch CarboTrain and select unit 5 and exercise 2. Open the Config 5 tab, where you can customize the spatial and temporal ranges of the running results by entering latitude and longitude and temporal ranges. Please note that the data entered here should be within the allowable range (−90 to 90 for latitude, 0 to 360 for longitude, 1980 to 2000 for the temporal range). Set the output path by clicking Set

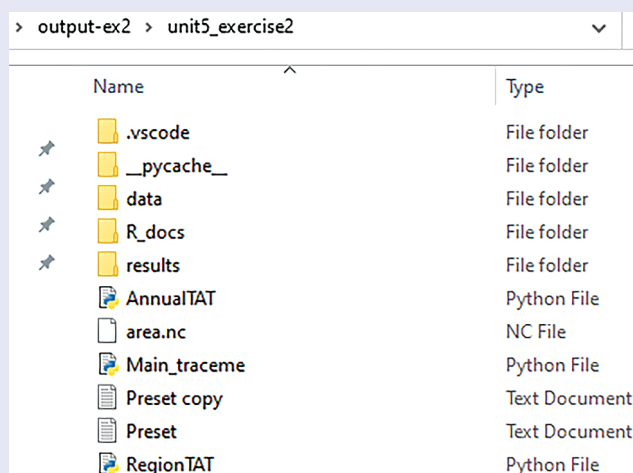
Output Folder. Finally, click Run Exercise to start the traceability analysis of the three CMIP6 models. (Note that the task running time is related to the selected ranges and the configuration of the computer).

The command window shows the running processes of the task as shown in Figure 20.6. When the task is finished, results will appear in the output path you set before (Figure 20.7).

Here are descriptions of the files and the workflow of the model evaluation system in CarboTrain. All files can be found in or under the `unit5_exercise2` folder in the output directory you specified earlier. The package consists of a file named `Preset.txt`, three Python scripts (`Main_traceme.py`, `AnnualTAT.py`, `RegionTAT.py`), and three subfolders (`data`, `results`, `R_docs`). `Preset.txt` contains a specification of the CMIP6 model outputs stored in the `data` folder. Traceability analysis requires the outputs of each

```
#####Temporal Traceability analysis has finished#####  
#####Start to run Spatial Traceability analysis...  
#####Start to read data#####  
#####Start to calculate baseline residence time #####  
##### Start to draw #####  
##### Save nc-file #####  
#####Temporal Traceability analysis has finished#####  
#####The program is end#####
```

Figure 20.6. Prompt in CMD interface when the program is running.



Name	Type
.vscode	File folder
__pycache__	File folder
data	File folder
R_docs	File folder
results	File folder
AnnualTAT	Python File
area.nc	NC File
Main_traceme	Python File
Preset copy	Text Document
Preset	Text Document
RegionTAT	Python File

Figure 20.7. Output results after the task is done.

model to include total C storage (vegetation C, soil C and/or litter C), GPP, NPP, temperature, and precipitation. After the task is submitted, the `Main_traceme.py` script reads the information in `Prset.txt` to preprocess the data. The preprocessed data is transferred to the `AnnualTAT.py` and `RegionTAT.py` scripts to perform temporal and spatial traceability analysis, respectively. The `R_doc` folder contains the R language script used to calculate the variance contribution of different components to C storage in the temporal traceability analysis. The results of the traceability analysis are output in the `results` folder.

We will now step through using the traceability analysis to analyze the differences in land C storage among three CMIP6 models (i.e., CESM2, CNRM-ESM2-1, and IPSL-CM6A-LR). In the `results` folder, you can find the results of the temporal traceability analysis on the three CMIP6 models over 1980–2000. First, you can

find differences in the time series of simulated C storage among the models (`temporal-1-Carbon-Dynamic.png`; Figure 20.8a). It is clear that the IPSL-CM6A-LR model has the lowest land C storage among the three models for the simulation period. The traceability analysis decomposes the C storage into C storage capacity and potential (Figure 20.8a). The result showed that the lowest land C storage in IPSL-CM6A-LR was due to the lowest C storage capacity rather than the C storage potential. Then, ecosystem C storage capacity is further decomposed into NPP and C residence time (`temporal-2-NPP-ResidenceTime.png`; Figure 20.8b). We can see that the lowest ecosystem C storage capacity in IPSL-CM6A-LR was driven by the shortest ecosystem C residence time among the three models. Furthermore, NPP is decomposed into GPP and C use efficiency (CUE) (`temporal-3-GPP-CUE.png`; Figure 20.8c). Ecosystem C residence time

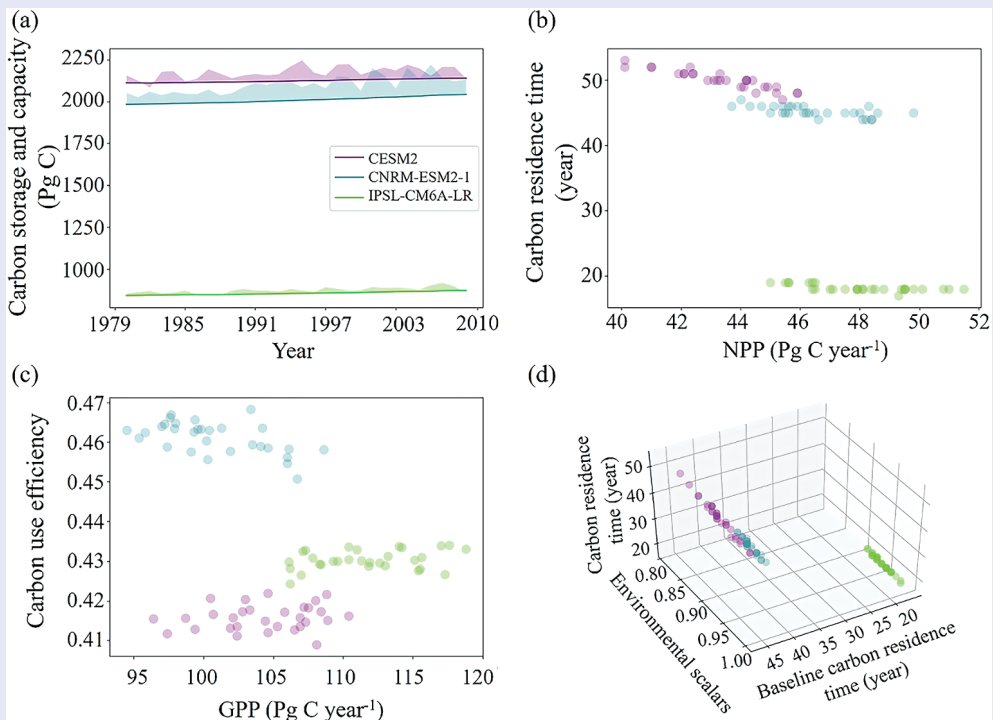


Figure 20.8. Results of the temporal traceability analysis on three CMIP6 models from 1980 to 2000. (a) Land C storage decomposed into C storage capacity and potential. (solid lines: C storage; shaded outlines: C storage capacity; shades: C storage potential (positive above and negative below the solid lines)); (b) C storage capacity decomposed into NPP and residence time; (c) NPP determined by GPP and CUE; (d) residence time decomposed into baseline residence time and environmental scalar.

can be decomposed into baseline C residence time and environmental scalars (`temporal-4-Envs-baselineResidenceTime.png`; Figure 20.8d). The simulated environmental scalars have not been provided for each model, so we cannot here further evaluate how climate forcings influence the baseline C residence time. However, among the three earth system models in this exercise, it is clear that the model parameterization of C allocation, C age, and C transfers among pools are the most important contributors to the large difference in global land C dynamic in Figure 20.8a.

In the `results/nc-files` folder, you can access the NetCDF-format data of the

temporal and spatial traceability analysis for each model. In the `results/figures` folder, you can find the outputs that show the results of the spatial traceability analysis on the three models over 1980–2000. Each figure includes the global distribution of a variable simulated by each model and the standard deviation of that variable. First, you can get the global distribution of the simulated C storage by each model and the global distribution of standard deviation of simulated ecosystem C storage among the models. The global distributions of all traceable components in Figure 20.8 are mapped. Figure 20.9 shows that the lowest baseline C residence time in IPSL-CM6A-LR is

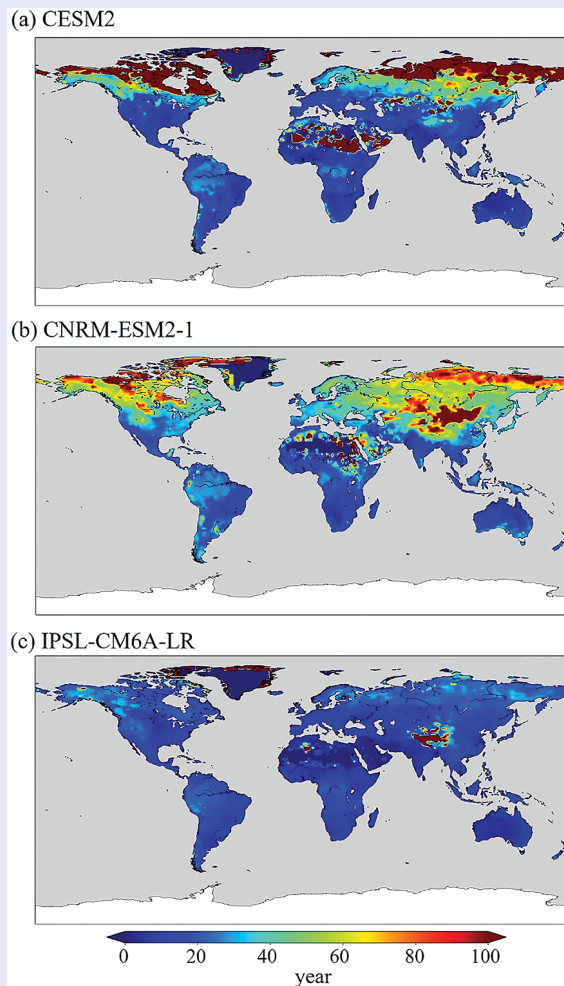


Figure 20.9. Global distribution of baseline carbon residence time in three CMIP6 earth system models over 1980–2000.

widely distributed in the Northern Hemisphere, except for the Tibetan Plateau. This analysis is helpful for informing developers of the IPSL-CM6A-LR model on how they may further improve their parameterization of ecosystem baseline C residence time.

QUESTIONS:

1. Among the three earth system models examined in this exercise, why does the IPSL-CM6A-LR model simulate the lowest land C storage? Which traceable component contributes most to the low values for C storage?
2. Based on the Figure 20.9, can you figure out which region has the largest variation in baseline C residence time among the three models? Based on the output of the post-MIP traceability analysis, can you generate a global map of the standard deviation of baseline C residence time among the three models?
3. If we want to apply the authentic traceability analysis to a model-intercomparison project (e.g., CMIP6), what additional information or modeling outputs are needed from each model?

UNIT SIX

Introduction to Data Assimilation



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER TWENTY-ONE

Data Assimilation

INTRODUCTION, PROCEDURE, AND APPLICATIONS

Yiqi Luo

Cornell University, Ithaca, USA

CONTENTS

Introduction of Data Assimilation /	173
The Need for Data Assimilation /	174
Seven-Step Procedure of Data Assimilation /	176
Scientific Values of Data Assimilation /	178
Suggested Reading /	180
Quizzes /	180

Realistic prediction of ecosystem responses to climate change requires not only a perfect model structure to represent the real-world processes but also parameterization to constrain model specifications and external forcing variables to reflect the environment that an ecosystem experiences to perform its functioning. Data assimilation is a statistical approach to model parameterization. This chapter introduces data assimilation, mainly focusing on concepts, procedure, and applications. We relate data assimilation to regression analysis to show a seven-step procedure: defining a research objective, having data, using one model, measuring data-model mismatches, minimizing the mismatches via global optimization, estimating parameters, and predicting ecosystem changes.

INTRODUCTION OF DATA ASSIMILATION

Data assimilation is a statistically rigorous approach to model parameterization. The latter is one of the three essential elements of realistic model prediction. To realistically predict ecosystem responses to climate change, we need the model structure to represent the real-world processes that control

system functions. We also need model specification through parameterization to constrain model predictions (Figure 21.1). Moreover, the external forcing variables have to reflect the physical, chemical, and biological environment that an ecosystem experiences to perform its functioning. Ideally, all the three elements have to be perfectly aligned before a model can well predict ecosystem responses.

In reality, we spend much more time developing process-based models than thinking about parameterization or forcing. When prediction of a model does not match well with observations, we usually look at model structure and often ignore parameterization and forcing. In fact, parameterization cannot be totally ignored. To make a model work, we have to tune parameters. But we have not learned much from parameter tuning in the past several decades. Data assimilation is a relatively new approach that can be used to rigorously estimate parameter values and offers a new way to learn about model parameterization. We will learn about data assimilation in the training units 6–10.

External forcing will be explored in unit 9. In short, we need to have a data-model consistent

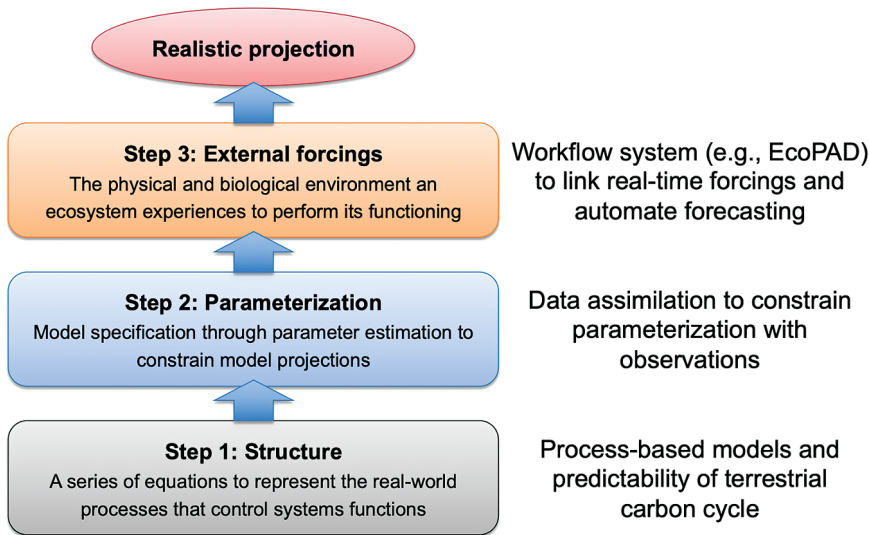


Figure 21.1. Three elements required for realistic model prediction of ecosystem responses to global change. Parameterization is equally important as model structure and forcing in predicting states of ecosystems under global change. Data assimilation offers a statistically rigorous approach not only to fit a model better with data, but also enable it to evaluate which, how, how much, and why parameters change. The workflow system Ecological Platform for Assimilating Data (EcoPAD) to link real-time forcing variables for automating forecasting and predictability of terrestrial carbon cycle is discussed in Chapter 33.

system, which can be realized by an interactive Ecological Platform for Assimilating Data (EcoPAD) or similar workflow systems, to have realistic external forcing variables to drive models for ecological forecasting.

THE NEED FOR DATA ASSIMILATION

To fully understand the need of data assimilation, we should first understand process-based modeling. Traditional modeling is often called simulation modeling or forward modeling. Simulation modeling usually needs to develop a process-based model first. The model usually means a model structure with a series of equations to represent processes in a system and assigned parameter values. Then, we use forcing variables to drive the model. The forcing variables for an ecological model usually include radiation, temperature, and precipitation. When the forcing variables are used to drive the model, the model generates outputs, such as carbon storage or net ecosystem exchange. The model outputs are sometimes compared with data for validation. This simulation approach has been extensively used by ecologists for about 60 years. It is very powerful for exploring ideas and hypotheses in ecology.

Here is an example of a simulation modeling study done at the Free Air CO₂ Enrichment (FACE) Duke Forest project. The project had three FACE rings and three control rings plus one prototype ring. In the FACE rings, the CO₂ concentration was 200 parts per million higher than that in the three ambient rings. One of the ideas for the FACE study was to show how a forest would be when CO₂ concentration increases to the elevated level in the middle of this century. The hypothesis was that forest carbon sequestration in the elevated CO₂ rings would be an estimate when the atmospheric CO₂ concentration was gradually increasing to that level.

Dr. James Reynolds and I worked together to test this hypothesis using a simulation modeling approach (Figure 21.2). We used an ecosystem model to test that hypothesis. We simulated two scenarios. One was to have CO₂ concentration gradually increasing as shown in panel A or abruptly increased as in panel D of Figure 21.2. When the atmospheric CO₂ concentration is gradually increasing, both photosynthesis and ecosystem respiration are gradually increasing as shown in panel B. But respiration lags behind due to the carbon that enters the ecosystem through photosynthesis and has to stay in the ecosystem for a while before it is released via respiration. The difference between photosynthesis and respiration is

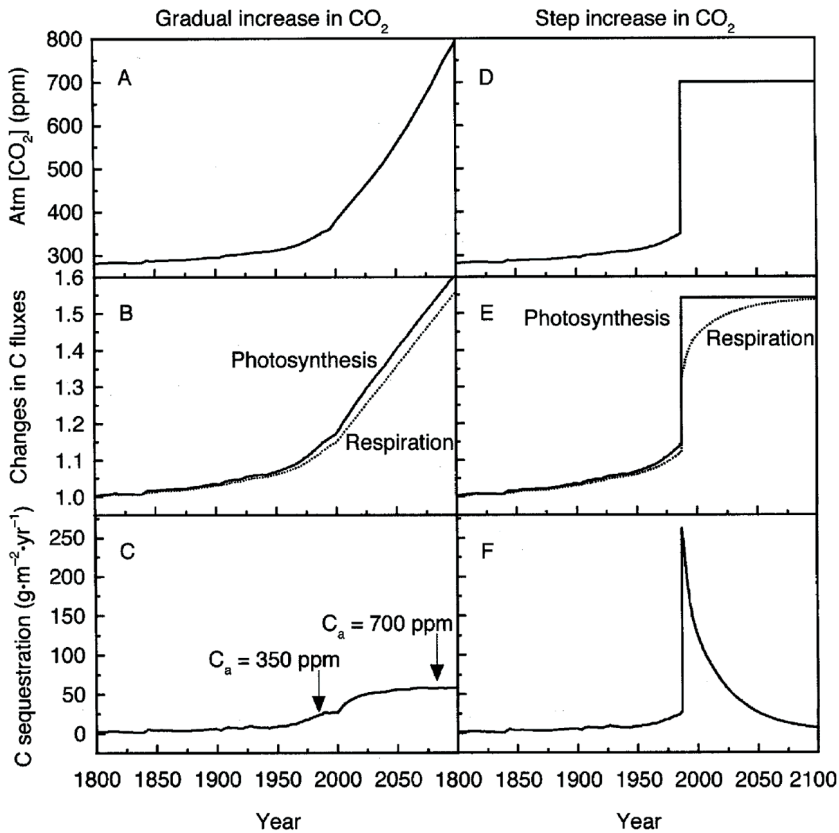


Figure 21.2. Simulation model to evaluate responses of ecosystem carbon (C) processes to a gradual vs. step increase in atmospheric $[\text{CO}_2]$. In response to the gradual increase in $[\text{CO}_2]$ (A), both photosynthesis and respiration increased gradually (B), leading to a gradual increase in carbon sequestration (C). In response to the step CO_2 increase (D), photosynthesis immediately increases, but respiration slowly increases (E), leading to a high rate of carbon sequestration right after the step increase in CO_2 , followed by decline.

the ecosystem carbon sequestration as shown in panel C. The modeled carbon sequestration gradually increases under the gradual CO_2 increase scenario in the real world. In contrast, photosynthesis abruptly increases in response to a step increase in CO_2 concentration as in the FACE experiment, but ecosystem respiration gradually increases, leading to a pulse increase in carbon sequestration followed by a gradual decline as shown in panel F. This is a simulation modeling study. It is quite powerful to contradict the original hypothesis that carbon sequestration observed in FACE plots can be extrapolated to the real world when CO_2 concentration is gradually increasing.

However, when simulation models are used for prediction or forecasting, the model and data usually do not match well. Predictions of soil carbon density by eleven models used in the Coupled Model Intercomparison Project Phase 5, CMIP5,

do not match with the observation-based carbon density from the homogenized world soil carbon database (HWSD) (Luo et al. 2015). None of the model predictions match with observations well. The mismatches between the models and observations are partly due to the lack of data constraints of either model structures or parameterization.

Data assimilation is also needed to integrate information contained in both model and data. Modeling is one approach to scientific inquiry mainly through process thinking. Data are obtained from field or laboratory research, which is also an approach to scientific inquiry through snapshot records of ecosystem states at the time when the observation was made. The two approaches acquire information on different aspects of an ecosystem. The information acquired from recording the state of the ecosystem is highly complementary to that of process understanding. Integration of model and

data will help gain the best knowledge from imperfect data and imperfect models (Luo et al. 2011).

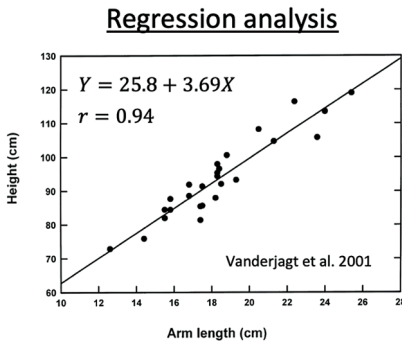
When data assimilation is used for integration of data and models, it starts with data and associated uncertainty (or noise). Data is combined with modeled values to inversely infer model structures and/or parameter values. Therefore, data assimilation is sometimes called inverse analysis, data-model fusion, inverse modeling, multiple constraints, or inference analysis, depending on situations when this technique is applied.

SEVEN-STEP PROCEDURE OF DATA ASSIMILATION

A data assimilation study is usually conducted by following a seven-step procedure. I will explain the seven steps mainly using the example in the study by Xu et al. (2006). You may go back and forth between this chapter and the paper by Xu et al. (2006). This seven-step procedure is explained in the practice of unit 6 and other chapters.

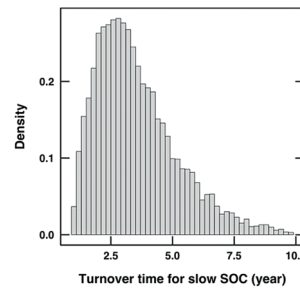
Data assimilation is fundamentally a statistical analysis. It is very similar to regression analysis in terms of the procedure.

When we conduct a regression analysis to understand a relationship, for example, between arm length vs. height, we usually have to go through seven steps. First, we need to decide what the objective is. In this case, we try to predict height from arm length. Second, we need to collect data of the arm length and height of all individuals in a sample of a population we are going to investigate. Third, we need to select a regression equation. In this case, a linear model is adequate. Fourth, we need to measure mismatches between observed heights, y , and estimated heights \hat{y} from the regression line. Fifth, we will minimize the mismatch using the method of least squares. Sixth, we will obtain the optimized estimates of regression coefficients a and b of the linear model. The seventh step is to predict a height from the arm length of one individual. For example, the regression line between heights and arm lengths from a sample in Nigeria reported from a paper by Vanderjagt et al. (2001) shows $y = 25.8 + 3.69x$, where x is the length of arm and y is the height (Figure 21.3). In short, the seven steps of conducting regression analysis are: (1) setting up an objective; (2) collecting data; (3) selecting an equation; (4) measuring



1. Using arm length (X) to predict height (Y) (**objective**)
2. Collect **data** of X and Y of all people in a group
3. Select an **equation** $Y = a + bX$
4. Measure **mismatch** ($y - \hat{y}$), \hat{y} is an estimate of y
5. Minimize $\sum(y - \hat{y})^2$ to **optimize estimation** of a and b
6. Obtain the optimized **parameters** a and b values
7. Use the optimized equation to do **prediction**

Data assimilation



1. **Objective:** To predict ecosystem responses to global change
2. **Data sets**
3. **Model**
4. Cost function as a measure of **mismatches**
5. Global **optimization**
6. Estimated **parameters**
7. **Prediction**

Figure 21.3. A seven-step procedure for both regression analysis and data assimilation. Data assimilation is a statistical approach and has a similar procedure with regression analysis. A key measure is fitness between data and model values in regression and posterior probability distributions of estimated parameters in data assimilation, although the data-model fitness is also important for data assimilation.

mismatches; (5) optimization by minimizing the sum of squares of the residuals; (6) estimation of parameters; and (7) prediction (Figure 21.3).

The procedure to conduct a data assimilation study is similar to a regression analysis. To conduct a data assimilation study, we also need to: first, set up an objective, such as predicting ecosystem response to global change; second, have data sets; third, have a model; fourth, develop a cost function to measure mismatches between observations and model values; fifth, have global optimization to minimize the cost function; sixth, estimate parameter values; and seventh, do prediction.

We use a study done by Xu et al. (2006) to show the seven steps. First, the objective of the study was to predict terrestrial carbon sequestration in response to elevated CO₂ concentration. Second, the study used six data sets at the ambient CO₂ treatment and six data sets at the elevated CO₂ treatment. Third, the study used the Terrestrial Ecosystem (TECO) model. Fourth, mismatches between observed and modeled values were measured by a cost function. Fifth, the optimization method used in the study was the Markov Chain Monte Carlo with Metropolis-Hasting algorithm. Sixth, the estimated parameters were decomposition coefficients of litter and soil organic carbon. Seventh, prediction was on carbon sequestration in nine pools of the TECO model at both the ambient and elevated CO₂ treatments.

Although the procedure is very similar to the regression analysis, there are some new concepts, such as mapping functions, a cost function, Markov Chain Monte Carlo (MCMC) sampling series, and the Metropolis-Hasting criterion. Those new concepts are explained in Chapter 22 and illustrated with examples in the practice session of unit 6.

Let us go over each of the seven steps (Figure 21.3). Note that cited symbols, page numbers, figures, and tables in this section all refer to those in the paper by Xu et al. (2006).

1. Step 1 is to define an objective. The general objective of the study was to predict ecosystem responses to elevated CO₂ concentration with uncertainty quantified. You can find this objective in paragraph 2 on page 1. A more specific objective of the study was to estimate parameter values, $c_1 \dots c_7$, in Table 1 on page 3 of the paper by Xu et al. (2006) and then predict carbon pool changes at ambient and elevated CO₂ treatments in Duke

Forest as described in paragraph 4 on page 2. Parameters, $c_1 \dots c_7$, represent decomposition of organic carbon from litter and soil pools.

2. Step 2 is to have data sets. The data sets used in the study are soil respiration, woody biomass, foliage biomass, litterfall, soil carbon, and mineral carbon at the ambient and elevated CO₂ treatments. Those data sets are described in paragraph 6 on page 3 and Table 2 on page 4 of the paper by Xu et al. (2006). We also need to use standard deviations of the six data sets in data assimilation.
3. Step 3 is to have a model. This study uses the Terrestrial Ecosystem (TECO) model. The TECO model is depicted in Figure 1 and described in equation 1 on page 2 of the paper by Xu et al. (2006). Please note that that figure has a typo. Pool X_1 should be a non-woody pool, which includes foliage and fine root biomass. One more point is that the TECO model in this paper was described by the matrix equation, which is the same equation as we studied in units 1–5 with slightly different notations.
4. Step 4 is to define a cost function. The cost function is to measure mismatches between observed and modeled values. The mismatches are described by equation 5 for individual data sets and equation 6 for all the six data sets on page 4 of the paper by Xu et al. (2006). The modeled values have to be mapped to observations via a mapping function Φ , which has six elements, respectively, for six data sets, as described in equation 3 on page 4. The mapping function is further illustrated in the practice session in unit 6 of this book. The cost function is equivalent to the likelihood function in equation 8 on page 4 of the paper by Xu et al. (2006).
5. Step 5 is to minimize the cost function via global optimization. The optimization was done with Markov Chain Monte Carlo (MCMC) sampling series. The MCMC has two phases, a proposing phase and a moving phase. The proposing phase is to generate new parameter sets. The moving phase is to examine if the newly proposed parameters should be accepted or not using a Metropolis-Hastings criterion. The two phases of MCMC are described in paragraphs 12 and 13 on pages 4 and 5 of the

paper by Xu et al. (2006). The study uses five parallel runs to test convergence of sampling series with the Gelman-Rubin method as described in paragraphs 14 and 15 on page 5 of the paper by Xu et al. (2006).

6. Step 6 is to estimate parameter values after the cost function is minimized in step 5. The estimated parameter values are described in paragraph 17 on page 5, depicted in Figures 3 and 4, and Table 3 on pages 6–8 of the paper by Xu et al. (2006). Estimated parameters have three types of distributions, or histograms, as shown in the mid-columns in Figures 3 and 4. The three types of histograms are bell-shaped for parameters c_1 , c_2 and c_4 , edge-hitting for c_6 , and flat for c_3 , c_5 and c_7 in Figure 4 on page 7 of the paper by Xu et al. (2006). The bell-shaped histograms indicate that those parameters are well constrained by data. The flat-shaped histograms mean that those parameters are not constrained by data. In other words, the six data sets do not have any information that is relevant to those parameters. The edge-hitting histograms mean that those parameters may be strongly correlated with other parameters or caused by other reasons.
7. Step 7 is to use the estimated parameters from step 6 in the TECO model to predict future states of ecosystems. The predicted carbon pool changes in the year 2010 at the ambient and elevated CO_2 treatments are described in paragraph 17 on page 5, Figure 9 on page 11, and Table 4 on page 12 of the paper by Xu et al. (2006). Uncertainty in model prediction is quantified with cumulative density functions in Figure 9 and confidence intervals in Table 4. CO_2 effects are represented by the predicted changes in pool sizes as indicated by the differences between the solid lines at elevated CO_2 and dashed lines at ambient CO_2 in Figure 9. As you can see from the paper by Xu et al. (2006), elevated CO_2 had the largest effects on woody biomass but no effects on the passive soil carbon pool.

Again, the seven steps to conduct a data assimilation study are setting up an objective, collecting data sets, having a model, and defining a cost function, using a global optimization method to

minimize the cost function, estimating parameter values, and predicting.

The seven-step data assimilation is essential to estimate parameter values and constrain model predictions through parameterization. The parameterization, in turn, is essential toward realistic model prediction.

SCIENTIFIC VALUES OF DATA ASSIMILATION

Data assimilation is a statistical tool. Its scientific values are realized when it is used to estimate parameter values, select alternative model structures, quantify uncertainty, and/or evaluate values of different data sets to constrain model prediction. The study by Xu et al. (2006) has illustrated how data assimilation was used to estimate parameters and quantify uncertainty of estimated parameters and model predictions. Chapter 31 further explains how data assimilation is used to quantify uncertainty associated with different model structures. Briefly, the chapter shows a study conducted by Shi et al. (2018) evaluating three soil carbon models: a classic or conventional model, a microbial model, MIMICS, and a vertically resolved model, CLM4.5, with three data sets: topsoil organic carbon in 0–30 cm, subsoil organic carbon in 30–100 cm, and microbial biomass carbon. The three data sets can constrain subsets of parameters for all the three models while complex models generate larger uncertainties in predictions even with the same data sets to constrain parameters.

Chapters 29 and 32 (i.e., practice in unit 8) explore how data assimilation can be used to evaluate values of different data sets to constrain model predictions. The values of data sets are measured by information content according to the Shannon information index and quantified from probability density functions of predicted changes. In principle, data sets contain information mainly on the parameters that are related to relevant processes. For example, flux data, such as net ecosystem exchange and gross primary production, can usually constrain parameters related to flux processes, such as leaf area index and maximal carboxylation rate. In comparison, pool data, such as plant biomass and soil carbon content, usually have information to constrain pool-related parameters, such as carbon transfer coefficients between pools.

Data assimilation is often used to test alternative model structures. For example, four alternative

models were evaluated for their representation of the priming effect on soil organic carbon decomposition with 84 data sets from 26 studies (Liang et al. 2018). Scientifically, the study evaluates whether the soil priming effect leads to net loss or net gain of soil organic carbon by adding new carbon input. Priming is a term to describe stimulation of old soil organic carbon (SOC) decomposition by new carbon addition. This priming effect usually results from stimulated microbial growth by new carbon addition.

The study requires data to be collected from isotope-labeled carbon addition experiments. Those studies have to provide SOC content, the added amount of new carbon, multiple measurements of CO₂ emission rates from total SOC samples and from labeled new carbon substrate, from which the CO₂ emission rates from old SOC can be estimated.

The study uses four models, including conventional soil carbon dynamic model, new-old carbon interactive model, Michaelis-Menten model, and reverse Michaelis-Menten model. Data assimilation is used to integrate the 84 data sets with the four models to evaluate data-model matches by comparing observed with modeled cumulative CO₂ emission rates from the total SOC, old SOC, and new carbon substrate (Figure 21.4). The conventional model can fit observed carbon releases from new, old, and total SOC extremely well. The interactive model can fit observations quite well, too. However, the fitting is not good for Michaelis-Menten and reverse Michaelis-Menten models, especially with observed carbon releases from old and total SOC.

The model-data fitting also shows that the conventional model fits the best and the Michaelis-Menten

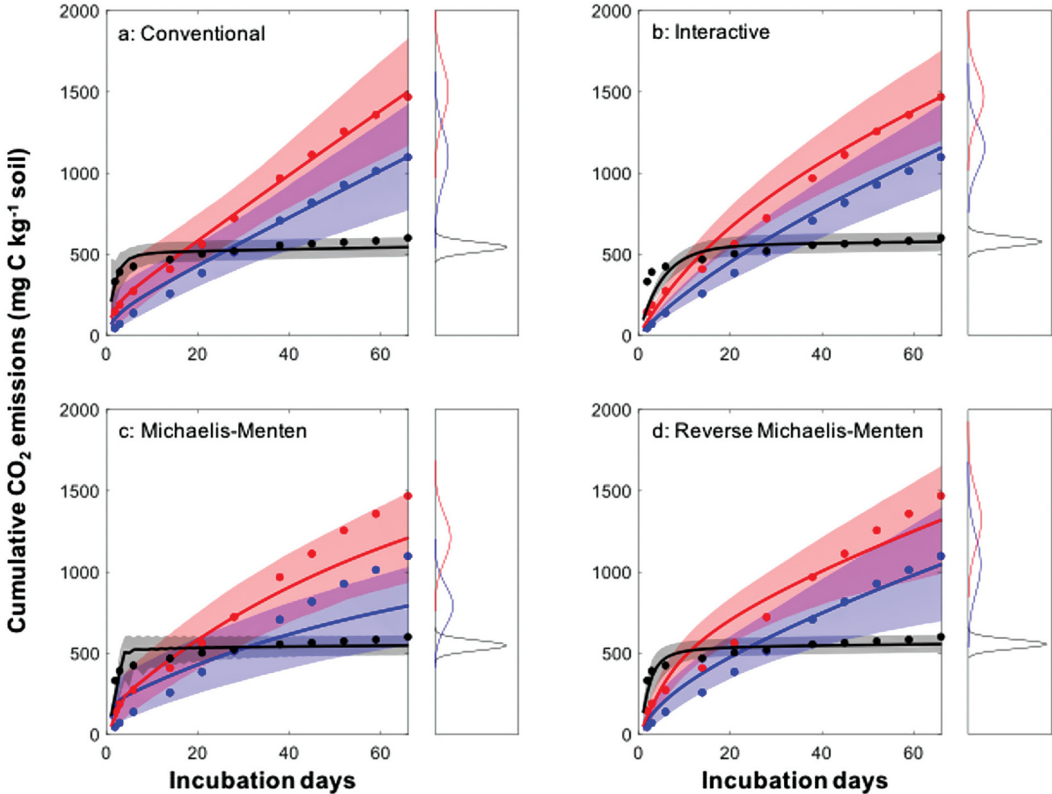


Figure 21.4. An example showing the performances of different models in simulating cumulative CO₂ emissions from old and new C substrates. Dots and lines are observations and model simulations, respectively. Shaded areas are the simulated ranges from 2.5th to 97.5th percentiles (i.e., 95% range). Blue and red are CO₂ emissions from old C at the control and new C addition treatments, respectively; Black is CO₂ emissions from added new C. The distributions of simulated cumulative CO₂ emissions at the end of experiment are also shown in each panel.

model fits the worst. A deviance information criterion (DIC) was used to select alternative model structures. DIC penalizes the model severely by the number of parameters. Among the four models, DIC is the smallest for the interactive models but largest for the conventional model as the latter has 12 parameters although the fitting of the model with data is the tightest. Based on the DIC, the most parsimonious model is the interactive model.

Then, the interactive model was used to estimate priming effect, replenishment, and net change in carbon after adding new carbon to the soil incubation experiments. Nearly 54% of the newly added carbon stays in the soil, stimulated old carbon decomposition via the priming effect is nearly 10% of the newly added carbon. As a consequence, SOC has a net gain of 32% of the added new carbon. This analysis indicates that priming does happen but is unlikely to lead to a net loss of SOC.

This study used data assimilation to address a highly controversial issue on priming effect. The study shows that adding new carbon to soil does result in a priming effect but eventually results in net carbon gain. The increase in SOC is related to nitrogen content in the added substrates. The study shows how data assimilation was used to select alternative model structures and suggests that a two-pool interactive model is the most parsimonious model to represent a SOC priming effect.

Overall, data assimilation is a statistically rigorous approach to model parameterization, which is essential to generate realistic model prediction. Data assimilation is usually conducted by following a seven-step procedure. The seven steps are: (1) setting up an objective; (2) collecting data sets; (3) having a model; (4) defining a cost function; (5) using a global optimization method to minimize the cost function; (6) estimating parameter values; and (7) predicting. The scientific values of data assimilation can be realized when it is used to estimate parameters, select alternative model structures, evaluate values of data sets in constraining model predictions, and quantify uncertainty in model prediction.

SUGGESTED READING

Xu, T., L. White, D. Hui, and Y. Luo. 2006. Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model

prediction. *Global Biogeochemical Cycles*, 20, GB2007, doi:10.1029/2005GB002468.

QUIZZES

1. What are the three elements, which all have to be perfectly lined up to realistically forecast future states of an ecosystem? Why?
2. Data assimilation is a method
 - a. to integrate data with model
 - b. to calibrate a model with data
 - c. to use statistical principles of analyzing data
 - d. often used for parameter estimation
3. Why is simulation modeling good for exploring ideas but not for prediction or forecasting?
4. Why are field research and modeling scientifically complementary? It is because:
 - a. one makes measurement and the other uses computer
 - b. they are two approaches to scientific inquiry of a research subject in different ways
 - c. the field research records the state of an ecosystem whereas the modeling explores relationships among processes
 - d. data from field research can be better interpreted from process understanding whereas model forecast can be better constrained with data
5. What are the seven steps of conducting a data assimilation study?
6. A cost function is
 - a. a function to calculate cost of training
 - b. to measure mismatches between modeled and observed values for all the data sets
 - c. a data set to be used in data assimilation
 - d. a model to be optimized through data assimilation
7. Posterior probability density function is
 - a. to indicate carbon density after field measurement
 - b. to indicate probability of a parameter value before data assimilation
 - c. a function of time from prior to posterior
 - d. to indicate a relative likelihood of an estimated parameter value after data assimilation
8. Why do we need a global instead of local optimization method for data assimilation?

CHAPTER TWENTY-TWO

Bayesian Statistics and Markov Chain Monte Carlo Method in Data Assimilation

Feng Tao

Tsinghua University, Beijing, China

CONTENTS

Introduction /	181
Bayes' Theorem /	181
Markov Chain Monte Carlo Method /	183
Convergence of MCMC Results /	186
Suggested Reading /	187
Quizzes /	187

Data assimilation is an effective way to integrate observations into models. We will demonstrate how parameters in a model may be estimated by data assimilation in such a way that model simulations best fit observations. Data assimilation based on Bayesian inversion is used to retrieve posterior distributions of model parameters from observations. The Markov Chain Monte Carlo (MCMC) method is applied as a numerical method to home in on the parameter set that maximizes goodness of fit between model outputs and measurements.

INTRODUCTION

In this chapter, we will first give a brief introduction to Bayesian statistics, which is the theoretical foundation of data assimilation. By learning the Bayes' theorem, you will understand why we can get knowledge from the data. After familiarizing ourselves with the theory, we will then learn how to apply the theory to an algorithm, namely the Markov Chain Monte Carlo (MCMC) method, which is useful for optimization problems, like fitting a model with multiple parameters to observational datasets. At the end of the chapter, we will discuss how to achieve stable optimization results with the MCMC algorithm, the so-called

simulation convergence problem. This chapter is a brief introduction to these topics to help readers build an intuitive understanding of data assimilation. If you want to know more about rigorous theoretical proofs of theorems and equations mentioned in this chapter, you are encouraged to refer to the suggested reading at the end.

BAYES' THEOREM

The Bayes' theorem, in plain words, is a method for calculating the validity of thinking (Horgan 2016). The "thinking" can be one's hypotheses, claims, or propositions. The results (i.e., the updated validity of thinking) given by Bayes' theorem is based on the best available evidence, such as the data, observation, or any information one can obtain. Bayesian thinking happens every day in our lives. Before we collect any related information and update our thinking over a particular question, we all do some initial thinking on the validity of the question, no matter if our attitude is initially skeptical or credulous. When we get new related data or information from the real world, we will naturally update our thinking over the question. This evidence-based updated thinking is the primary character of Bayes' theorem.

We can mathematically express Bayes' theorem as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \propto P(B|A)P(A) \quad (22.1)$$

If we take events happening in our lives alphabetically as A and B , the probability of event A happening can be expressed as $P(A)$. The probability of event B happening can be expressed as $P(B)$. We use $P(A|B)$ to indicate the probability of event A happening given the fact that event B has happened. Analogously, $P(B|A)$ represents the probability of event B happening given the fact that event A has happened.

The Bayes' theorem intuitively expresses that the probability of event A to happen given event B has already happened is proportional to the possibilities of event B happening where event A has already happened, and, at the same time, event A happening. When we calculate the probability of event A happening given event B has already happened (i.e., $P(A|B)$), we can assume event B has been well observed. The probability of event B happening (i.e., $P(B)$) then becomes a constant. $P(A|B)$ is now proportional to a product of $P(B|A)$ and $P(A)$. We may know some information about A and can transfer such information into an initial thinking (i.e., $P(A)$). Based on the initial thinking, we can also get a sense of how likely it is for event B to happen given A has happened (i.e., $P(B|A)$). Mathematically we call $P(A)$ our prior knowledge, while $P(B|A)$ is termed the likelihood. Combining the prior knowledge and likelihood, we update our thinking as posterior knowledge (i.e., $P(A|B)$).

We will use an example to illustrate how Bayes' theorem works. Suppose you are in an international meeting and receive a flyer advertising the training course "New Advances in Land Carbon Cycle Modeling", which will train people in carbon cycle modeling and data assimilation. You are interested and prepared to apply. Before the application, however, you feel concerned that your lack of experience in simulation modeling may hinder your success in the training course. You may wonder what the probability of your success in the training course will be, given you have no previous modeling experience. By emailing the course coordinator, you learn that according to previous records, 80% of past trainees attending the training course succeeded in understanding the training material and finished the practices. In terms of

the background of all trainees, half of them had no experience in modeling before the course. Among the successful trainees, 43.75% had no experience in modeling before the course.

We can conceptualize this issue into Bayes' theorem. We can define the event A as one's success in the training course and the event B as the trainee having no modeling experience before the training course. We now know $P(B) = 0.5$ (half of the trainees have no modeling experience before the training course); $P(A) = 0.8$ (80% of the trainees succeeded in the training course); and $P(B|A) = 0.4375$ (among the successful trainees, 43.75% had no experience in modeling before the training). By substituting all the factors into Equation 22.1, we get $P(A|B) = 0.7$. The result implies that even if a trainee did not have any experience in modeling before the training course, the chance of success is 70%, which is still pretty large.

The question could go the other way. Suppose you have prior experience in modeling, you may ask what the possibility of your being successful in the training course is. To answer this question, we need to calculate $P(A|\neg B)$, where $\neg B$ indicates the event of B not happening, and $P(\neg B) = 1 - P(B)$. According to Equation 22.1,

$$\begin{aligned} P(A|\neg B) &= \frac{P(\neg B|A)P(A)}{P(\neg B)} \\ &= \frac{[1 - P(B|A)]P(A)}{1 - P(B)} \end{aligned} \quad (22.2)$$

when we substitute all the probability numbers we obtained from the coordinator into Equation 22.2, we get $P(A|\neg B) = 0.9$. So, if you have previous experience in modeling, the chance of success in the training course increases from 70% to 90%.

We used discrete events and their possibilities in the above example to illustrate Bayes' theorem. In simulation modeling, we use parameters that are continuous values to represent ecological processes. We assume that if we choose the correct parameter values, observations in the real world (e.g., soil organic carbon content) can then be predicted by the model. In this context, we use probability distributions of parameters (as expressed by the probability density function) to show all possible values of one parameter and the likelihood of occurrence for different values within that overall distribution. The Bayes' theorem will help us, in a reverse way,

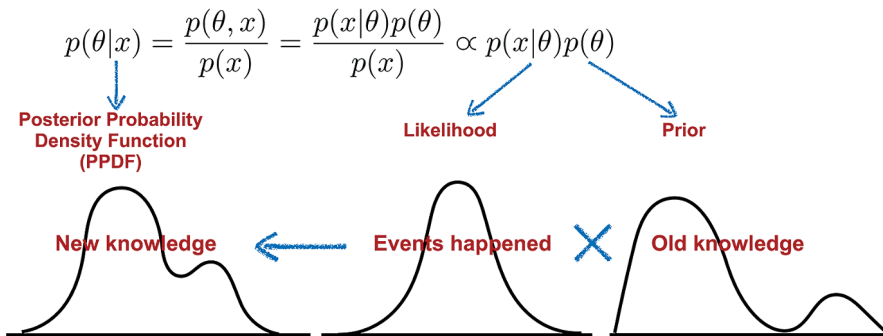


Figure 22.1. Schematic diagram of the Bayesian inference process.

find the most likely distribution of parameter θ to best describe the observed data x in a model:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta) \quad (22.3)$$

where the $p(\theta)$ is the prior probability density function of parameter θ ; $p(x|\theta)$ represents the likelihood that we use the model with θ set to different values from its distribution, and correctly predict the observation x . Combining the prior distribution $p(\theta)$ with the likelihood $p(x|\theta)$ will give the posterior probability density function of θ (i.e., $p(\theta|x)$).

Bayes' theorem explicitly offers the possibility of using observational data to refresh our prior knowledge. The term $p(x|\theta)$ in Equation 22.3 shows how likely we are to observe data x under different values of parameter θ from its distribution. This indicates if we propose different θ values and record them with their likelihoods (i.e., $p(x|\theta)$), we will eventually obtain the posterior distribution describing the optimized parameter values that best fit the observations (Figure 22.1).

MARKOV CHAIN MONTE CARLO METHOD

In data assimilation, the Markov Chain Monte Carlo (MCMC) method offers an algorithm that screens a range of possible parameter values under specific prior knowledge and retrieves the posterior distribution of the parameter. The MCMC method is composed of two parts, namely a Markov Chain and the Monte Carlo method. A Markov Chain is a stochastic model that describes a sequence of possible events. The probability of each event in a Markov Chain depends only on the state attained in the previous event. The Monte Carlo method, on the other hand,

represents a class of algorithms that sample events from a Markov Chain to fit an optimization target. Various algorithms exist for sampling. Here, we will focus on the Metropolis-Hastings algorithm.

We begin with an example to show the workflow of the MCMC method. Suppose that we are organizing the training course “New Advances in Land Carbon Cycle Modeling”. The admission process is critical to the final outcome of the course. By scanning the profiles of applicants, we need to select those applicants who will most likely benefit from the training. The question at this stage is then formulated as “through which algorithm shall we select the participants from among all the candidates?”

First, we may need to specify the topical scope of the training course so that we can define the target group who may be of interest (Figure 22.2). A clear and well-defined scope will help define the “perfect candidate” who will definitely benefit from the training. For example, we may accept people who are doing research in the carbon cycle field, familiar with basic simulation modeling, and who would like to integrate models and data to enhance their understanding of ecological processes.

After defining the scope of the training course, we may send out the flyers to advertise the course. We will then receive applications from people who are interested in and believe that they can benefit from the course. All the applicants now become the potential candidates as trainees. We now need to evaluate the match between the backgrounds of candidates and our training scope. The characteristics of the “perfect candidate” in our mind will function as a reference in such evaluation.

We make decisions based on the assessments of candidates. Three possibilities will be on the table. We may find that the background of the applicant is very close to that of our “perfect candidate”, which means the applicant will most likely benefit from

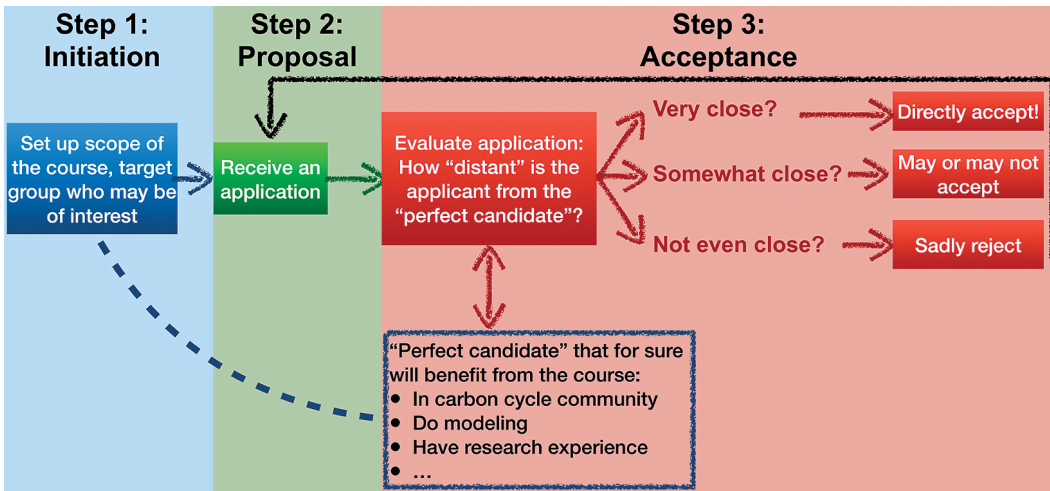


Figure 22.2. An example of using the MCMC method to admit candidates based on their application details.

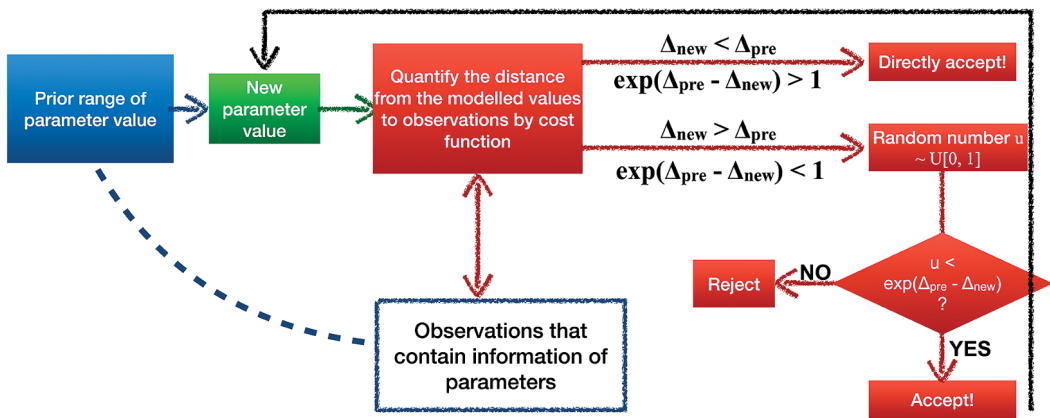


Figure 22.3. Workflow of the MCMC method. The Metropolis-Hastings algorithm is used to accept or reject proposed parameter values in a Markov Chain.

the training course. In this case, we will admit the applicant without hesitation. Conversely, we may also receive applications where the background of the applicant does not fit the scope of the training at all. We will directly reject those applications. Another possibility is that the background of the applicant does not perfectly fit our “perfect candidate”, but is still close enough to get likely benefits from our course. Instead of direct rejections, we may consider an algorithm to decide whether to accept or reject the applicant. For example, we may set new criteria to evaluate the fitness of these applicants to the training course and eventually admit those who meet the criteria.

Four steps together make up the admission process. The first step initiates the process. We

need to set the scope of the training course and determine the target group who will be of interest. In the second step, applications provide all the possibilities we can choose from to fit our target of maximizing the overall benefit of the training. We then make decisions in the third step to accept or reject the applications according to some standards. Finally, in step four, we repeat the procedure from steps two and three for each candidate.

The MCMC method follows the same four steps to optimize parameters to fit model simulations with observations (Figure 22.3). In step one, we initiate the algorithm by first determining the optimization target (i.e., the observations). We set a prior range of the possible parameter values that remains consistent with the ecological meaning

of the investigated process, and simultaneously, allows sufficient flexibility in model simulations.

The second step generates the Markov Chain. We propose a sequence of candidates of parameter values so that we can select the candidates by a certain algorithm (in our example, the Metropolis-Hastings algorithm). A proposal distribution will be selected and serves to generate new candidate parameter values. We then apply the proposed parameter values in the model and get the simulation results.

We use two proposal distributions in step two, which are the uniform distribution in the test run and the normal distribution in the formal run (Haario et al. 2001). When we have no prior knowledge about the parameter, a uniform distribution will be a reasonable start. The uniform distribution does not contain any shape information of the parameter distribution, but only sets the upper and lower limits of parameter values, which we determined based on our general understanding of the parameter in step one. The k^{th} proposed parameter value is expressed as:

$$\theta^k = \theta^{k-1} + r \frac{\theta^{\max} - \theta^{\min}}{D} \quad (22.4)$$

where r is a random value drawn from the uniform distribution in the interval between 0 and 1; D indicates the maximum step size in the proposal. For example, $D = 5$ indicates that the maximum step of the k^{th} proposed parameter value (θ^k) from the $(k-1)^{\text{th}}$ proposed parameter value (θ^{k-1}) is 20% of the prior range (i.e., $\theta^{\max} - \theta^{\min}$).

After the test run of MCMC with a uniform proposal distribution, we obtain some preliminary information on the investigated parameter. The following formal run will then use the results of the test run as its prior knowledge so that we can more efficiently get the stable posterior distribution of the parameter. A normal distribution will be assumed in the formal run. When we investigate multiple parameters, we assume that the parameters follow a multivariate normal distribution. The parameter values will be proposed as:

$$\theta^k = \begin{cases} \theta^{k-1} + \mathcal{N}\left(0, \text{cov}\left(\theta^{\text{test run}}\right)\right) & k \leq k_0 \\ \theta^{k-1} + \mathcal{N}\left(0, \text{cov}\left(\theta^{\text{formal run}}\right)\right) & k > k_0 \end{cases} \quad (22.5)$$

where we calculate the covariance of parameters by the results of the test run at the starting stage

of the formal run (i.e., before the k_0^{th} iteration). After some iterations of the formal run, we continuously update the covariance information from the formal run results. The point k_0 is an empirical value we may set based on experience, depending on how fast we think the formal run can fully utilize the prior information in the test run.

In step three, we apply the Metropolis-Hastings algorithm to accept or reject proposed parameter values. We evaluate how the simulated results based on the proposed parameter values fit the observations and decide whether to accept the proposed parameter values or not. A cost function serves to quantify the degree of discrepancy of the predicted values of the target variable x in the simulations relative to the observations of x . We express the cost function of the k^{th} proposed parameter (θ^k) as:

$$\Delta_k = \frac{1}{2\sigma^2} \left[\sum_{i=1}^n (x_{i,\text{mod}} - x_{i,\text{obs}})^2 \right]_{\theta^k} \quad (22.6)$$

where $\left[\sum_{i=1}^n (x_{i,\text{mod}} - x_{i,\text{obs}})^2 \right]_{\theta^k}$ describes the distance of the simulation results from observations with the sample size of n (e.g., ten-year record of net primary productivity of a grassland site); σ^2 is the variance of observations, which we can either calculate from measurements or from empirical knowledge. Equation 22.6 expresses the cost function for a single data set. When multiple data sets are used (e.g., ten-year record of net primary productivity, soil organic carbon content, and soil respiration of a grassland sites), the mismatches of individual data sets are added together, optionally with weightings to emphasize the importance of certain variables relative to the others, to obtain the cost function.

Now we make use of Bayes' theorem. From Equation 22.3, we can define the likelihood of the parameter value (θ^k) in fitting model simulations ($x_{i,\text{mod}}$) to observations ($x_{i,\text{obs}}$) as:

$$\mathcal{L}(\theta^k) = p(\theta^k | x) \propto p(x | \theta^k) p(\theta^k) \quad (22.7)$$

We use the cost function in a monotonically decreasing exponential form to describe $p(x | \theta^k)$:

$$\mathcal{L}(\theta^k) \propto \exp(-\Delta_k) \quad (22.8)$$

Using Equation 22.8 to quantify the likelihood requires several assumptions in Bayesian statistics (Craiu and Rosenthal 2014). From the data optimization perspective, Δ_k expressed in Equation 22.6 offers a measurement of the deviation of model simulations from observations. We connect such a metric with the likelihood of the proposed parameter values: we assign a high likelihood to proposed parameter values when the cost function result is small and vice versa.

We use the likelihood ratio of consecutively proposed parameter values in the Markov Chain to make the acceptance decision. The probability of accepting the k^{th} proposed parameter value (θ^k) based on the results of $k-1^{\text{th}}$ proposed parameter value (θ^{k-1}) can be formulated as:

$$P(\theta^{k-1}, \theta^k) = \min \left\{ 1, \frac{\mathcal{L}(\theta^k)}{\mathcal{L}(\theta^{k-1})} \right\} \quad (22.9)$$

Substituting Equation 22.8 into Equation 22.9 gives:

$$P(\theta^{k-1}, \theta^k) = \min \{ 1, \exp(\Delta_{k-1} - \Delta_k) \} \quad (22.10)$$

When the cost function with θ^k is smaller than that with θ^{k-1} , this tells us that the results of the present simulation are closer to the observations than the previous one. By Equation 22.10, we get $P(\theta^{k-1}, \theta^k) = 1$, which indicates we will definitely accept the proposed parameter value θ^k . Conversely, if the cost function calculated from proposed parameters θ^k is higher than that of θ^{k-1} , this tells us that the simulation results by θ^k , in comparison to observations, is worse than that by θ^{k-1} . We then get a $P(\theta^{k-1}, \theta^k)$ result in the interval between 0 and 1. In this case, we may or may not accept the proposed θ^k , depending on a random chance. In practice, we compare $P(\theta^{k-1}, \theta^k)$ with a random value u that follows the uniform distribution $u \sim U(0, 1)$. We accept the proposed θ^k when $P(\theta^{k-1}, \theta^k) > u$. Otherwise, we will reject the proposed θ^k .

It may seem strange that we sometimes accept the proposed θ^k , instead of a direct rejection when a proposed set of parameters yields simulation worse than the previous set (i.e., the new cost larger than the previous one, $\Delta_k - \Delta_{k-1} < 0$). In optimization, the “bad” results can be helpful in finding the global minimum of the cost function.

The response surface of the cost function can be extremely nonlinear for a multivariate, high-dimension model. If we do not give a chance to accept a “bad” set of parameters, we may easily get trapped in a local optimum instead of the global optimum. Therefore, we assign a possibility of acceptance, subject to chance, when $P(\theta^{k-1}, \theta^k)$ is in the interval between 0 and 1.

CONVERGENCE OF MCMC RESULTS

When we do MCMC, we need to confirm that the results are the same even if we start from different points along the Markov Chain. This is called testing for convergence (Gelman et al. 2014). In theory, the MCMC should converge as long as the Markov Chains are long enough. That means that the parameters from different independent MCMC simulations will share the same or very similar posterior distributions.

The estimated parameters in the MCMC simulation, however, do not necessarily resemble each other at the very beginning among different Markov Chains. In the test run and starting stage of the formal run, accepted parameter values may experience a different pathway to find the global optimum. However, after sufficient iterations of MCMC simulation (e.g., 10,000 iterations), the observations will eventually constrain the parameter to the same space, where the model simulates most closely to the measurements entering the cost function. By plotting the sampling series, we can visually determine when the accepted parameter values are stably constrained and set the early stage of the sampling series as the burn-in period. The exact length of burn-in period can be empirical. The first half of the accepted parameter chain is a safe setting for burn-in period given sufficient iterations in MCMC simulation (e.g., 100,000 iterations). In the final analysis, we exclude the burn-in results and only use the results after convergence to generate the posterior distribution.

Both visual and statistical assessment can be used to determine the convergence of MCMC results. The degree of overlap among different MCMC sampling series is a rough indicator of the convergence. We can also quantify the convergence statistically. The Gelman-Rubin (G-R) statistics are one option. The G-R statistics quantify the differences among different runs (B_j) and the fluctuation of accepted parameter values within the same run (W_j):

$$\left\{ \begin{array}{l} B_i = \frac{N}{K-1} \sum_{k=1}^K (\bar{c}_i^{\cdot,k} - \bar{c}_i^{\cdot\cdot})^2 \\ W_i = \frac{1}{K(N-1)} \sum_{k=1}^K \sum_{n=1}^N (c_i^{n,k} - \bar{c}_i^{\cdot,k})^2 \end{array} \right. \quad (22.11)$$

where i denotes parameters investigated in the study; K is the number of parallel runs; N is the length of each run; $c_i^{n,k}$ represents the n^{th} accepted value of parameter i in the k^{th} parallel run after the burn-in period. The G-R statistic is then defined as:

$$GR_i = \sqrt{\frac{W_i(N-1)/N + B_i/N}{W_i}} \quad (22.12)$$

Once convergence is reached, GR_i should approximately approach 1.

In summary, this lecture introduces two fundamental underpinnings of data assimilation. The first, the Bayesian inference, sets the theoretical foundation of improving model performance by observations. The second, the Markov Chain Monte Carlo method, provides a numerical method to assimilate observational data into a model based on an optimization of the goodness of fit of the model output to the observational data. In the next chapter, using examples from different ecological studies we will show you how data assimilation

can be deployed to advance understanding of the land carbon cycle.

SUGGESTED READING

- Haario, H., E. Saksman, and J. Tamminen. 2001. An adaptive Metropolis algorithm. *Bernoulli* 7:223–242.
- Xu, T., White, L., Hui, D. and Luo, Y., 2006. Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction. *Global Biogeochemical Cycles*, 20(2).

QUIZZES

1. Briefly describe the Bayes' theorem.
2. Why do we need a test run before the formal run when we know little about the prior?
3. Explain why we need to sometimes accept parameters that have a larger cost function value (Δ) than the previously accepted ones.
4. Why do we discard the results from the burn-in period?
5. We need to generate a random value (u) to compare with the results $\exp(\Delta_{\text{pre}} - \Delta_{\text{new}})$ when the cost function value is larger than the previous one ($\Delta_{\text{new}} > \Delta_{\text{pre}}$). Why do we accept the proposed parameter values only when $u < \exp(\Delta_{\text{pre}} - \Delta_{\text{new}})$, instead of $u > \exp(\Delta_{\text{pre}} - \Delta_{\text{new}})$?



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER TWENTY-THREE

Application of Data Assimilation to Soil Incubation Data

Junyi Liang

China Agricultural University, Beijing, China

Jiang Jiang

Nanjing Forestry University, Nanjing, China

CONTENTS

Soil Incubation Experiments /	189
Soil Carbon Models /	190
Application of Data Assimilation to Soil Incubation Data /	191
Summary /	196
Suggested Reading /	196
Quizzes /	196

Soil incubation is a widely used technique in studying soil organic carbon cycling. Integrating soil incubation data with soil carbon models can potentially reveal mechanisms of soil carbon dynamics underlying observations. This chapter is to illustrate how data assimilation is applied to analyze data from soil incubation experiments using soil carbon models. After a brief introduction to soil incubation experiments and soil carbon models, a three-pool model is used to illustrate the seven-step procedure of data assimilation for the analysis of soil incubation data. As two critical aspects, different cases in the optimization step and the dependence of parameter acceptance rate on cost function are described with detailed examples.

SOIL INCUBATION EXPERIMENTS

Soil incubation experiments are commonly carried out for studying soil organic carbon and nitrogen cycling processes. Typically, fresh soil samples are collected from the field, crushed, sieved, and mixed before a certain amount of soil is placed in a container (e.g., a Mason jar). The container is then

exposed to different treatments of, for example, temperature and moisture. For carbon decomposition studies, carbon dioxide (CO₂) emission rate (or respiration rate) is usually measured repeatedly during the incubation period. Data from soil incubation studies are usually plotted by either CO₂ emission rates or cumulative CO₂ emission over time. Compared with in situ observations in the field, soil incubation experiments allow ready control of environmental factors and reduce heterogeneity and confounding effects from many processes and factors. Thus, results from such experiments can facilitate mechanistic understanding of soil carbon processes, such as the turnover rates and temperature sensitivity of soil organic carbon decomposition. However, incubation experiments usually isolate the soil from plants and other components in ecosystems, and thus may show different behavior from soils in a full ecosystem setting.

To illustrate how soil incubation data may be used to inform modeling via data assimilation, this chapter takes as an example the study by Liang et al. (2015), which compares different methods for

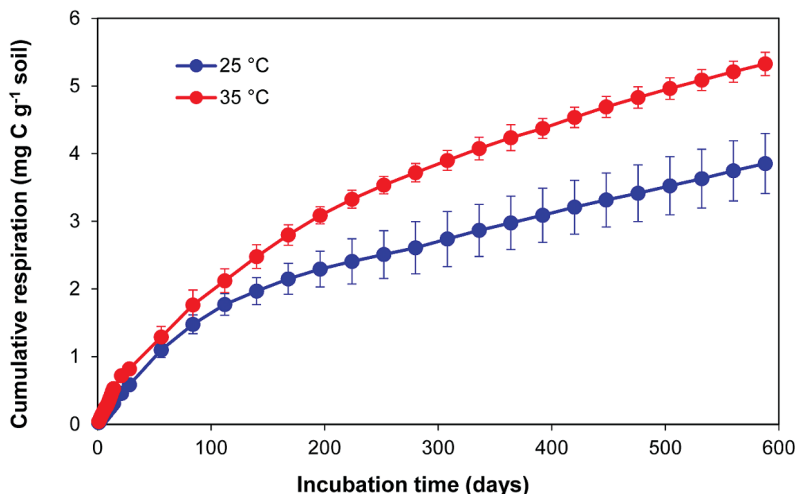


Figure 23.1. Cumulative CO₂ emission (R) at 25°C and 35°C. Data originally from Haddix et al. (2011) and used in Liang et al. (2015). Modified with permission from *Soil Biology and Biochemistry*: Liang et al. (2015).

estimating temperature sensitivity of soil organic carbon decomposition. The data are from a study by Haddix et al. (2011). Fresh soils were sampled from the field and transported to the laboratory. In the laboratory, the soil samples were sieved, and visible roots and rocks were picked out. From each soil sample, subsamples were taken and incubated in jars at different temperatures. The first seven days were treated as pre-incubation to minimize the possible artificial influences of field sampling, sieving and transportation. After the pre-incubation, CO₂ emission rates were measured as the rate of CO₂ concentration increase in the head space of the jars over time. CO₂ emission rates were measured daily during the first two weeks of incubation, weekly for the next two weeks, and every four weeks thereafter. Overall, there were 36 sampling occasions over the 588-day incubation period. Cumulative CO₂ emissions can be calculated by adding together the daily amounts of emitted CO₂ from day 0 (Figure 23.1). These example data will be used for the following illustration in this chapter.

Conventionally, total CO₂ emissions and/or CO₂ emission rates during the incubation period are directly compared to reveal the effect of different treatments, such as temperature or moisture levels. For example, temperature sensitivity impacts the decomposition of soil organic carbon under climate warming. In the past, the temperature sensitivity of soil organic carbon decomposition was directly calculated as the CO₂ emission rate at a higher temperature, R_{high} , divided by that at a lower temperature,

R_{low} (Rey and Jarvis 2006). This estimate usually underestimates the temperature sensitivity after the initial incubation stage because greater decomposition results in less substrate at high than low temperatures at the same point of incubation time. In addition, the direct comparisons of total CO₂ emissions and/or CO₂ emission rates may not reveal processes underlying soil carbon dynamics, such as turnover rates of different carbon components and dependences of temperature sensitivity on substrate quality. When soil carbon models, which explicitly represent such processes, are integrated with data from soil incubation experiments via data assimilation, we can potentially learn more about the process responses underlying the observed soil carbon dynamics.

SOIL CARBON MODELS

The general framework and underlying principles of many current soil carbon models were presented in chapters 1 and 2. Generally, first-order kinetics are used to simulate soil carbon cycling (Stanford and Smith 1972, Andren and Paustian 1987). The simplest soil carbon decomposition model simulates soil carbon dynamics as a single pool (Figure 23.2a). However, one-pool models usually perform poorly, since soil organic matter is compartmentalized into pools of different lability, due to various structures of substrates and protection mechanisms. As a result, multiple-pool models are most commonly used to simulate soil carbon dynamics.

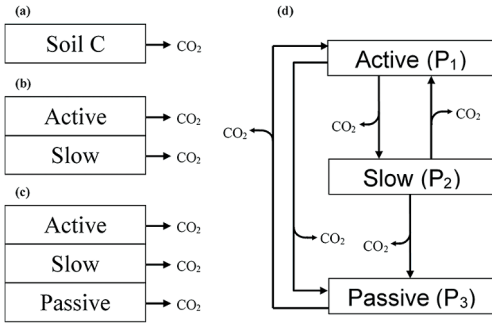


Figure 23.2. Schemes of soil carbon models using first-order kinetics. Modified with permission from *Soil Biology and Biochemistry*: Liang et al. (2015).

For example, a two-pool model divides soil carbon into active and slow pools (Figure 23.2b), while a three-pool model divides soil carbon into active, slow and passive pools (Figure 23.2c). The two- and three-pool models simulate soil carbon cycling without transfers among pools. Another type of model simulates transfers among pools (Figure 23.2d). Similar to ecosystem models discussed in previous chapters, soil carbon models can be described in a matrix form (Chapter 5). The only difference when using soil carbon models to represent incubation experiments is that they have initial soil carbon pools size(s) at the very beginning of the experiment, but do not have carbon inputs.

Here we will take the three-pool model without transfer (Figure 23.2c) as our example to illustrate the procedure of data assimilation for analysis of soil incubation data. The three-pool model can be described as:

$$\frac{dX(t)}{dt} = -KX(t) \quad (23.1)$$

where

$$K = \text{diag}(k_i) = \begin{pmatrix} k_1 & 0 & 0 \\ 0 & k_2 & 0 \\ 0 & 0 & k_3 \end{pmatrix}, \quad (23.2)$$

and

$$X(t) = \begin{pmatrix} X_1(t) \\ X_2(t) \\ X_3(t) \end{pmatrix} \quad (23.3)$$

k_1, k_2, k_3 are the turnover rates of the active, slow and passive pools, and X_1, X_2, X_3 are their pool sizes, respectively.

APPLICATION OF DATA ASSIMILATION TO SOIL INCUBATION DATA

Our data assimilation example follows the seven steps introduced in Chapter 21. Before we start, let us revisit the seven-step procedure again: (1) defining an objective; (2) preparing data; (3) choosing a model; (4) using a cost function; (5) applying an optimization method; (6) estimating parameters; and (7) generating predictions.

Step 1: defining an objective. The objective of the study by Liang et al. (2015) is to compare different methods for estimating temperature sensitivity of soil organic carbon decomposition. They first reviewed published methods for temperature sensitivity. The temperature sensitivity is usually expressed as Q_{10} , measuring the proportional change in soil carbon decomposition rate for a 10 K warming. They found that many studies directly compare CO_2 emissions at different incubation temperatures, estimating an apparent temperature sensitivity of soil organic carbon decomposition. However, this method does not provide the intrinsic temperature sensitivity of different soil carbon components. As introduced earlier, soil carbon models consider soil carbon dynamics in different pools depending on their turnover rates (k), which can be used to represent substrates with different lability. How the parameter k changes with temperature can inform the intrinsic temperature sensitivity of different soil carbon components. Therefore, a specific objective of the study is to estimate the intrinsic temperature sensitivities of different carbon pools in soil models.

Step 2: preparing data. As mentioned above, the data come from an incubation experiment by Haddix et al. (2011). For the illustration of this chapter, data from the 25°C and 35°C treatments (Figure 23.1) are used. Means and standard deviations of measurements from the incubation experiment are needed for data assimilation. The data are organized as shown in Table 23.1.

TABLE 23.1
*Cumulative CO₂ emission during the incubation
 organized for data assimilation*

Incubation time (days)	25°C		35°C	
	Mean	SD	Mean	SD
1	0.025	0.002	0.039	0.003
2	0.048	0.004	0.076	0.001
3	0.072	0.003	0.112	0.003
4	0.099	0.007	0.147	0.004
5	0.121	0.007	0.182	0.006
6	0.142	0.009	0.214	0.008
7	0.165	0.007	0.250	0.009
8	0.180	0.006	0.267	0.023
9	0.212	0.005	0.306	0.019
10	0.226	0.004	0.347	0.014
11	0.244	0.003	0.389	0.021
12	0.271	0.010	0.434	0.010
13	0.288	0.011	0.487	0.015
14	0.312	0.016	0.528	0.015
21	0.459	0.026	0.718	0.014
28	0.585	0.055	0.819	0.053
56	1.099	0.113	1.290	0.155
84	1.478	0.139	1.764	0.219
112	1.770	0.161	2.121	0.176
140	1.968	0.200	2.477	0.175
168	2.149	0.229	2.797	0.151
196	2.294	0.265	3.087	0.126
224	2.406	0.334	3.326	0.132
252	2.508	0.353	3.535	0.129
280	2.609	0.386	3.719	0.135
308	2.738	0.407	3.900	0.146
336	2.865	0.384	4.076	0.168
364	2.977	0.395	4.235	0.192
392	3.088	0.400	4.373	0.148
420	3.207	0.398	4.535	0.150
448	3.314	0.399	4.691	0.155
476	3.414	0.418	4.829	0.159
504	3.525	0.429	4.962	0.159
532	3.630	0.435	5.087	0.154
560	3.745	0.445	5.210	0.156
588	3.853	0.444	5.326	0.172

Step 3: choosing a model. Models are chosen dependent on the study objective. Liang et al. (2015) uses four models to compare different estimates of Q_{10} . In practice, the length of the incubation experiment is an important aspect to consider when choosing which model to use. If the length of experiment is relatively short, for example, days to months, the one-pool or two-pool models may be appropriate. If the incubation lasts longer, for example, years, the three-pool model may be better to fit data. Here, our purpose is to illustrate how to use data assimilation to analyze soil incubation data, and the example experiment lasts for 588 days. Therefore, the three-pool model is chosen. The three-pool model has a total of eight to-be-determined parameters, including the initial fractions of active and slow pools, f_1 and f_2 , decomposition rates of organic carbon in three pools at 25°C, k_1 , k_2 , k_3 , and corresponding temperature sensitivity parameters, q_1 , q_2 and q_3 . Details of these parameters are shown in Table 23.2. The initial fraction of the passive pool, f_3 , does not need to be estimated as it can be directly calculated as $1 - f_1 - f_2$. The turnover rates of carbon pools at 35°C can be calculated as $k_i \times q_i$. If your experiment does not have treatments at different temperatures, just ignore those temperature sensitivity parameters. In that case, the model has five parameters to be estimated. To conduct data assimilation, the prior probability density functions (PDFs) of the parameters are needed, which represents the prior knowledge ahead of data assimilation. When there is not much prior knowledge about the parameter distributions, the prior PDFs can be specified as uniform distributions over parameter ranges, for example based on available literature (Table 23.2).

Step 4: cost function. Before the cost function is estimated, a mapping function is needed to relate the simulation results to their corresponding measurements. Looking at Table 23.1, the 15th measurement is conducted in day 21. However, the 15th model simulation is CO₂ emission in day 15, and the model

TABLE 23.2

A description of model parameters and the ranges of their prior uniform distributions of the three-pool models used in the chapter

Parameter	Unit	Description	Prior uniform distribution	
			Minimum	Maximum
f_1	Unitless	Initial fraction of the active pool	1.00×10^{-2}	1.00×10^{-1}
f_2	Unitless	Initial fraction of the slow pool	1.00×10^{-1}	6.00×10^{-1}
k_1	d^{-1}	Turnover rate of the active pool	1.00×10^{-3}	2.00×10^{-2}
k_2	d^{-1}	Turnover rate of the slow pool	1.00×10^{-5}	5.0×10^{-4}
k_3	d^{-1}	Turnover rate of the passive pool	1.00×10^{-6}	5.0×10^{-5}
q_1	Unitless	Temperature sensitivity of the active pool	1.00	3.00
q_2	Unitless	Temperature sensitivity of the slow pool	1.00	4.00
q_3	Unitless	Temperature sensitivity of the passive pool	1.00	5.00

simulation for day 21 is the 21st result. Our mapping function will gather a subset of model results that correspond to each measurement occasion. We can build a function between the incubation time, t , and the rank of measurements, i : $t = f(i)$, using the first column of Table 23.1. The incubation time corresponding to a specific measurement can be derived via the mapping function. For example, the incubation time of the 36th measurement is day 588 ($588 = f(36)$). With this mapping function (i.e., $t = f(i)$), you can easily locate the model results corresponding to the 36th measurement is in day $f(36)$ (i.e., day 588).

After deriving the subset of model results that correspond with observations, the cost function, J , which measures the magnitude of mismatch of the simulation results from observations, can be specified. We will adopt the formula for J introduced in Chapter 22. Figure 23.3 shows a snapshot of the J values from iteration 101 to 200 when running data assimilation using the three-pool model and data in Table 23.1. The cost function is used in the optimization step to determine whether to accept or reject proposed parameter sets, which is further discussed in the next step.

Step 5: optimization. Let us follow the introduction to the Markov Chain Monte Carlo (MCMC) method in Chapter 22 to conduct the optimization. The value of the cost function in iteration m is compared with that in iteration $m-1$. There are basically two cases when comparing the values of cost function, $J_m \leq J_{m-1}$ and $J_m > J_{m-1}$. Examples below are used to show the two cases.

Case 1: In Figure 23.3, $J_{112} = 21.3$ and $J_{111} = 221.2$, respectively. 21.3 is smaller than 221.2, meaning the mismatch between model results and measurements is reduced, i.e., the model simulation is improved. In this case, the proposed parameters are accepted.

Case 2: $J_{173} = 114.1$ and $J_{172} = 69.2$, respectively (Figure 23.3). 114.1 is bigger than 69.2, meaning the mismatch between model results and measurements is increased, i.e., the model simulation is worsened. In this case, the exponential of

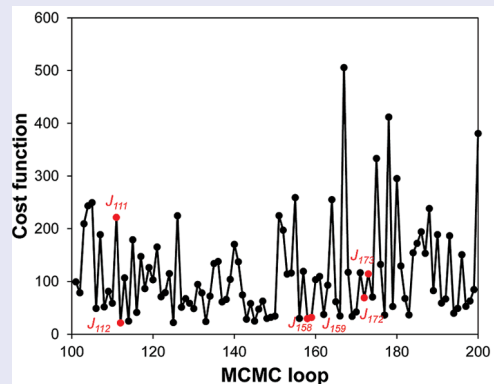


Figure 23.3. A snapshot of the cost function values from iteration 101 to 200.

$(J_{172}-J_{173})$ is calculated. Here the value is 3×10^{-20} . The value is then compared with a randomly number between 0 and 1. If the randomly selected number is smaller than 3×10^{-20} , the proposed parameters are accepted. Otherwise, the proposed parameters are rejected. Given the number to be compared is a random value between 0 and 1, the chance for accepting proposed parameters is $(3 \times 10^{-18})\%$. Let us take a look at another example, $J_{159} = 31.9$ and $J_{158} = 29.6$, respectively (Figure 23.3). 31.9 is slightly bigger than 29.6. Therefore, we calculate the exponential of $(J_{158}-J_{159})$, which is 1×10^{-2} , suggesting there is a chance of 1% to accept the proposed parameters.

Recall that the purpose of data assimilation is to derive global optimizations for parameters. Allowing a chance to accept proposed parameters even with the increased mismatch between simulations and observations can avoid the parameter estimations getting trapped in local optimizations. But the chance to accept proposed parameters decreases steeply with the increase in simulation-observation mismatch (Figure 23.4).

Step 6: *estimating parameters*. Now we have all the accepted parameters after finishing 100,000 MCMC iterations. Excluding results in the

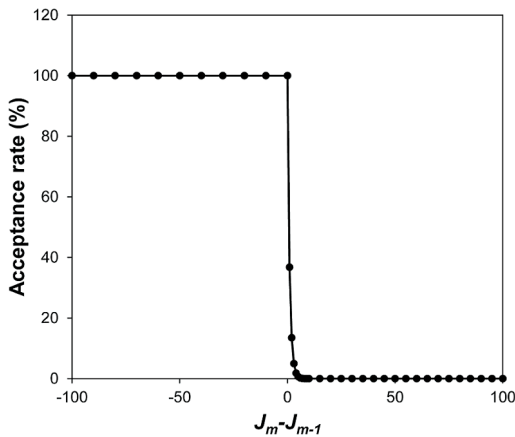


Figure 23.4. Dependence of parameter acceptance rate on the change of cost function ($J_m - J_{m-1}$). If the mismatch between simulations and observations decreased ($J_m - J_{m-1} \leq 0$), the acceptance rate is 100%. If $J_m - J_{m-1} > 0$, there is a chance to accept proposed parameters, and the acceptance rate decreases steeply with the mismatch increase.

burn-in period (Chapter 22), we have posterior distributions for the selected parameters (Figure 23.5). In the figure, the distributions of parameters related to the active and slow pools have significant single peaks, which means these parameters are well constrained. The distributions of parameters related to the passive pool, by contrast, are not as good, suggesting the data may not have enough information to constrain them.

From the posterior distributions, we can derive the maximum likelihood estimates (MLEs) of parameters. MLEs represent the parameter value at which the distribution peaks. In the example, the Q_{10} distributions peak at 1.22 and 1.76 for the active and slow pools. We can use these values to represent the intrinsic temperature sensitivities at 25°C of the corresponding pools. The Q_{10} distribution of the passive pool is not well constrained, and we can use the mean value, 2.67, as a reasonable guess of its intrinsic temperature sensitivity at 25°C. The temperature sensitivity increases with the decrease of substrate lability, suggesting that carbon pools with longer turnover times are more vulnerable to warming.

Step 7: *generating predictions*. predictions of soil carbon decomposition can be generated with the derived parameters and the model. With the generated predictions, we can evaluate the model performance by comparing observations and simulations. In Figure 23.6, if all dots are right on the 1:1 line (i.e., $y = x$), it means the model simulations are exactly equal to observations, which of course is very unlikely. In practice, we can use a regression line to describe the comparison of model simulations and observations. The regression lines at 25°C and 35°C are $y = 0.9919x$ and $y = 0.9998x$, or very close to $y = x$ which would represent perfect agreement. This indicates that the constrained model fits the data very well.

With the constrained model, we can also simulate the dynamics of different carbon pools during the incubation. Figure 23.7 shows that the active pool dominates the CO_2 emission at the early stage and is depleted very fast. At the end of incubation, the cumulative CO_2 emission from the active pool is similar between the 25°C and 35°C incubations. The contributions of the slow and passive pools to CO_2 emissions gradually increase after the active pool is depleted.

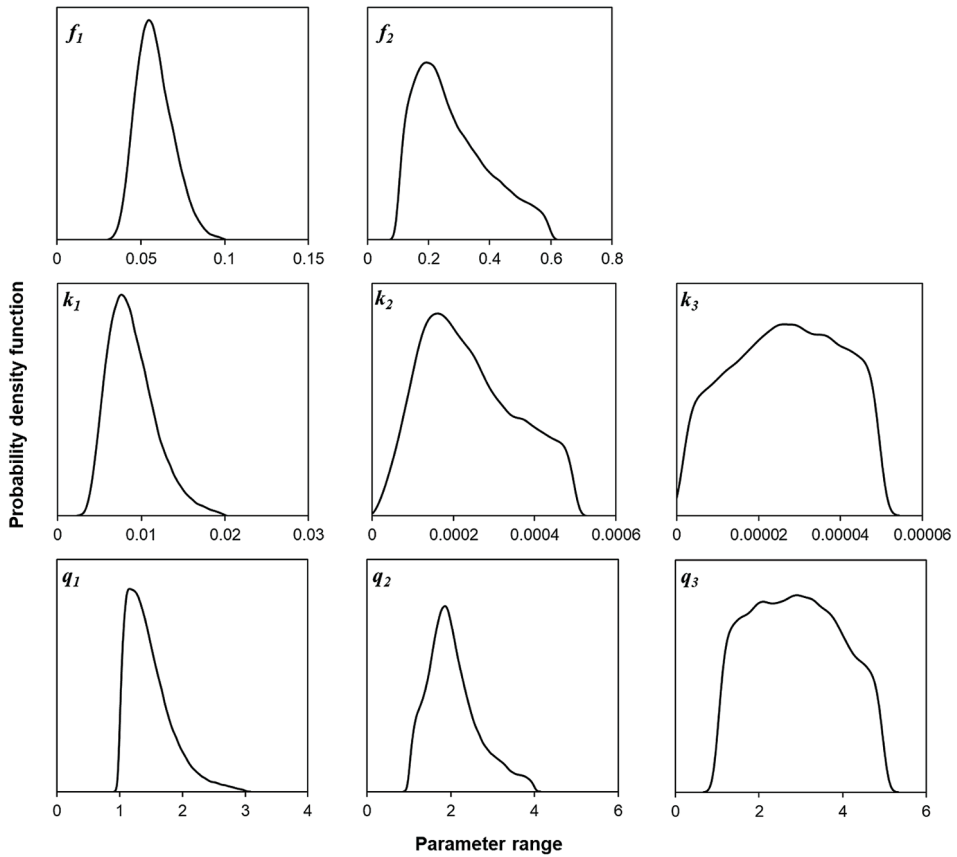


Figure 23.5. Posterior probability density functions of the parameters in the three-pool model. Modified with permission from Soil Biology and Biochemistry: Liang et al. (2015).

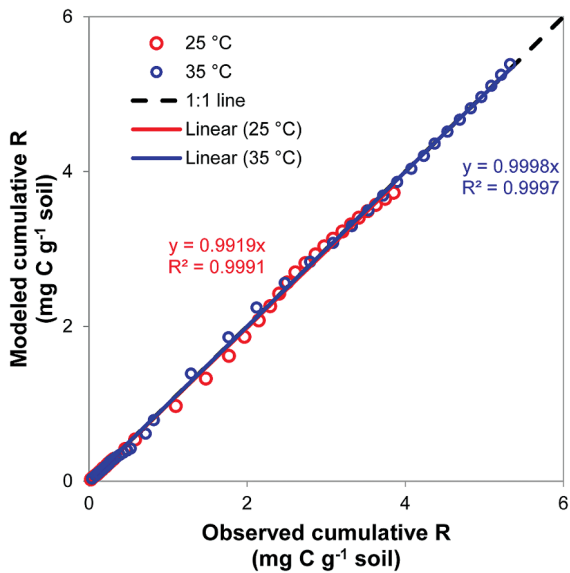


Figure 23.6. Comparison of observed and modeled cumulative CO₂ emissions (R) in the incubation experiment. All the dots distribute around the 1:1 line, suggesting the model simulations are well matched with observations.

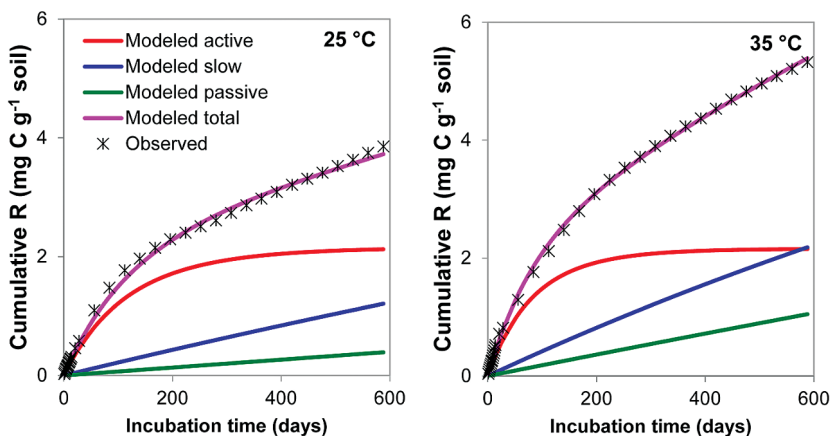


Figure 23.7. Observed and modeled cumulative CO₂ emissions (R) from individual and total pools at two incubation temperatures. Modified with permission from *Soil Biology and Biochemistry*: Liang et al. (2015).

SUMMARY

This chapter introduced the application of data assimilation to soil incubation experiments. A variety of models can be chosen to simulate soil carbon cycling depending on the scientific question and the length of incubation. Assimilating soil incubation data with models can help understand processes underlying the observed soil carbon dynamics, such as turnover rates and temperature sensitivities of substrate with different qualities.

SUGGESTED READING

Liang, J., D. Li, Z. Shi, J. M. Tiedje, J. Zhou, E. A. G. Schuur, K. T. Konstantinidis, and Y. Luo. 2015. Methods for

estimating temperature sensitivity of soil organic matter based on incubation data: A comparative evaluation. *Soil Biology & Biochemistry* 80:127–135.

QUIZZES

1. How can soil carbon models and data assimilation help with the analysis of soil incubation data?
2. Why is the length of incubation experiments an important consideration when choosing a model?
3. Suggest why some parameters can be well constrained and other cannot, and how this information guides future experimental design.

CHAPTER TWENTY-FOUR

Practice 6

THE SEVEN-STEP PROCEDURE FOR DATA ASSIMILATION

Xin Huang

Northern Arizona University, Flagstaff, USA

CONTENTS

Introduction / 197
Step 1: Defining an Objective / 197
Step 2: Preparing Data / 198
Step 3: Model / 198
Step 4: Cost Function / 200
Step 5: Optimization Method / 201
Step 6: Estimated Parameters / 203
Step 7: Prediction / 203
Exercises with CarboTrain Toolbox / 204
Suggested Reading / 206

This chapter will guide you to learn the seven-step procedure of data assimilation by replicating a published study on the assimilation of observational data into the TECO model to calibrate decomposition rate parameters for multiple soil organic matter pools. We will first review the seven steps in Chapter 21 and learn how to program these steps using code examples in Python. Then we will perform three exercises to reproduce figures from an earlier paper on the study with the CarboTrain toolkit. It is recommended that you go over the source code of the three exercises in CarboTrain. You are expected to program a data assimilation algorithm with your own model or data sets after this practice.

INTRODUCTION

The seven steps in data assimilation (DA) are described in detail in Chapter 21. They are: (1) defining an objective; (2) preparing data;

(3) choosing a model; (4) using a cost function; (5) applying an optimization method; (6) estimating parameters; and (7) generating predictions (Figure 24.1). This practice will use an example from a study by Xu et al. (2006) to introduce how to program these seven steps. The example of the DA study by Xu et al. (2006) is programmed in a Python file 'Probabilistic_inversion.py'. We will use this Python program to illustrate each of the seven steps in a DA study.

STEP 1: DEFINING AN OBJECTIVE

Defining an objective usually means to decide target parameters to be estimated by DA. The target parameters chosen in the study by Xu et al. (2006) are decomposition rates to represent fractions of carbon (C) leaving seven soil organic matter (SOM) pools of the Terrestrial ECOSystem (TECO) model (Table 24.1). The TECO model will be described in Step 3.

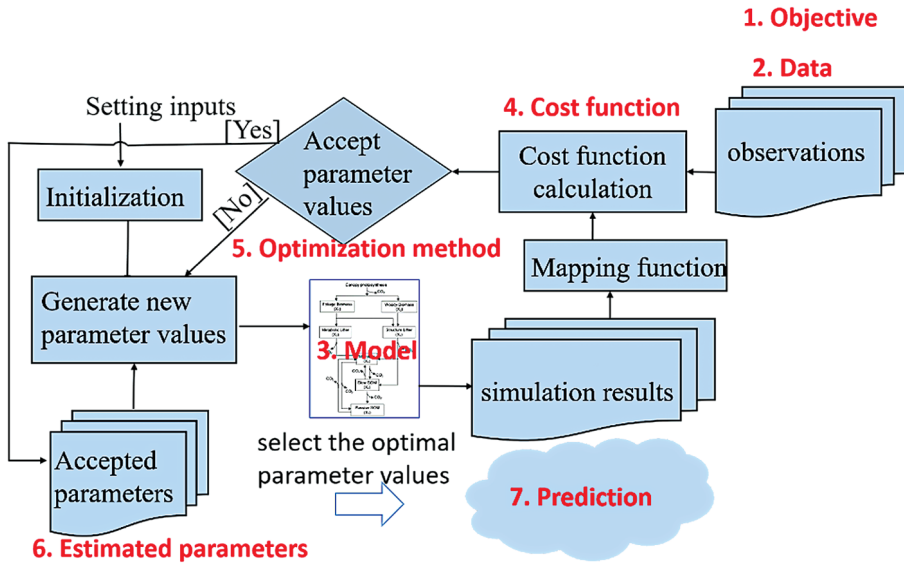


Figure 24.1. Seven-step procedure of the data assimilation (DA) study by Xu et al. (2006).

TABLE 24.1
Seven parameters to be estimated

Parameter	Description	Unit	Min ($\times 10^{-4}$)	Max ($\times 10^{-3}$)
c_1	Fraction of C leaving pool 1	$gCg^{-1}d^{-1}$	1.764	2.95
c_2	Fraction of C leaving pool 2	$gCg^{-1}d^{-1}$	0.548	0.274
c_3	Fraction of C leaving pool 3	$gCg^{-1}d^{-1}$	54.79	27.34
c_4	Fraction of C leaving pool 4	$gCg^{-1}d^{-1}$	5.48	2.74
c_5	Fraction of C leaving pool 5	$gCg^{-1}d^{-1}$	27.4	6.85
c_6	Fraction of C leaving pool 6	$gCg^{-1}d^{-1}$	0.548	0.274
c_7	Fraction of C leaving pool 7	$gCg^{-1}d^{-1}$	0.0137	0.00913

Based on the study by Xu et al., (2006)

STEP 2: PREPARING DATA

The DA study in Xu et al. (2006) integrated six observations (i.e., soil respiration, woody biomass, foliage biomass, litterfall, C in forest floor, and C in forest mineral soil) under ambient and elevated CO_2 treatments. All the 12 data sets are saved in six data files under the ‘/input’ folder. There are three rows in each data file. The first row is the time series, the second is observation under ambient CO_2 treatment and the final row is observation under elevated CO_2 treatment. Figure 24.2 shows the Python code for reading the six data files and calculating the six observation variances. The variable `ninput` with a value of 1 or 2 is to indicate which CO_2 treatment is applied in the Python program. For example, if `ninput` has a value of 1, the variables `soilResp`, `woody`, `foliage`, `litterfall`, `forestFloor`, and `forestMineral` save the six observations from the ambient CO_2 treatment. All these six observations are collected in the `obsList` variable and the corresponding six observation variances are saved in the `varList` variable. These variables will be used in Step 4.

STEP 3: MODEL

A terrestrial ecosystem (TECO) model is used (Xu et al., 2006). The TECO model has a seven-pool structure and the fractions of C exiting each pool each day are the parameters to be estimated (Table 24.1, the variable `c` in Figure 24.3a). This step involves two


```

209  ## Step 2: 6 Data Sets and their variances
210  ## initialize observation-related variables
211  soilResp=np.loadtxt('./Source_code/unit_6/input/SoilRespiration.txt')[[0,ninput],]
212  woody=np.loadtxt('./Source_code/unit_6/input/Woody.txt')[[0,ninput],]
213  foliage=np.loadtxt('./Source_code/unit_6/input/Foliage.txt')[[0,ninput],]
214  litterfall=np.loadtxt('./Source_code/unit_6/input/Litterfall.txt')[[0,ninput],]
215  forestFloor=np.loadtxt('./Source_code/unit_6/input/ForestFloor.txt')[[0,ninput],]
216  forestMineral=np.loadtxt('./Source_code/unit_6/input/ForestMineral.txt')[[0,ninput],]
217  ## collect the 6 data sets into an observation list
218  obsList=[woody[1,],foliage[1,],litterfall[1,],forestFloor[1,],forestMineral[1,],soilResp[1,]]
219  ## variance of 6 observations
220  # ddof=1 non-bias estimator for sample variance
221  varList=[np.var(obs, ddof=1) for obs in obsList]

```

Figure 24.2. The Python code for reading six observations from text files and calculating their variances.

(a)

```

50  ## Step 3: TECO model
51  def run_model(c):
52      """Forward run model.
53      Global args: tau, b, u, x0, Nt, cbnScale"""
54      ## x stores 7 simulated carbon pools simulated in 5 years (daily scale)
55      x=np.zeros([7,Nt+1], dtype=float)
56      ## multiple matrix A and diagonal matrix C
57      AC = np.matrix([[ -c[0], 0, 0, 0, 0, 0, 0],
58                    [ 0, -c[1], 0, 0, 0, 0, 0],
59                    [ 0.7123 * c[0], 0, -c[2], 0, 0, 0, 0],
60                    [ 0.2877 * c[0], c[1], 0, -c[3], 0, 0, 0],
61                    [ 0, 0, 0.45 * c[2], 0.275 * c[3], -c[4], 0.42 * c[5], 0.45 * c[6]],
62                    [ 0, 0, 0, 0.275 * c[3], 0.296 * c[4], -c[5], 0],
63                    [ 0, 0, 0, 0, 0.004 * c[4], 0.03 * c[5], -c[6]]])
64      x[:,0] = x0
65      ## simulate for 5 years
66      for i in range(1,Nt+1):
67          ## Eq. 1 in Xu et al., (2006)
68          x[:,i]=np.dot((np.eye(7)+AC*tau[i-1]),x[:,i-1])+np.asarray(b)*u[i-1]*cbnScale
69          ## simulated carbon pools with 7 rows and Nt columns
70          xsimu=x[:,range(1,Nt+1)]
71          .....
90      return [woody_simu, foliage_simu, litterYr, forestFloor_simu, forestMineral_simu,
              soilResp_simu]

```

(b)

```

249      ## Step 3: TECO model
250      ## running model, return [woody_simu, foliage_simu, litterYr, forestFloor_simu,
251      forestMineral_simu, soilResp_simu]
251      simuList = run_model(c_new)

```

Figure 24.3. The Python code for (a) defining a function of TECO model simulation; (b) calling the function to run the model.

program fractions in `Probabilistic_inversion.py`: defining model simulation, and running model simulation (Figure 24.3). A function `run_model` (Figure 24.3a) defines the TECO model simulation with a set of parameter values (variable `c`). The core codes of model simulation (lines 65–68) are to program the differential equation:

$$\frac{dX(t)}{dt} = \xi(t)ACX(t) + BU(t) \quad (24.1)$$

where $X(t)$ is a 7×1 vector to represent the seven C pool sizes at time t ; $\xi(t)$ is a scaling function describing environment effects at time t ; A is a 7×7 matrix representing the C transfer coefficients among the seven pools; C is a 7×7 diagonal matrix with diagonal elements describing the fraction of C leaving each pool or the parameters to be estimated; $B = (0.25, 0.3, 0, 0, 0, 0, 0)^T$ is a vector accounting for the partitioning coefficients of C input to non-woody and woody biomasses; $U(t)$ is the C input from photosynthesis at time t .

The variable `Nt` (Figure 24.3a) is five-year simulation time on a daily time step, with a value of 1825. The variable `xsimu` will accrue the seven C pool sizes over the `Nt` simulation days. Rather than the seven C pool sizes, the return values of the `run_model` function are six simulated data sets (i.e., `woody_simu`, `foliage_simu`, `litterYr`, `forestFloor_simu`, `forestMineral_simu`, and `soilResp_simu`) to match with the six observational datasets. Mapping operators are used for this purpose and we will learn about these operators in Step 4.

To run the model simulation, a new set of parameter values (variable `c_new`) needs to be passed to the `run_model` function (Figure 24.3b). During this process, the parameter values in variable `c_new` will be assigned to variable `c` in Figure 24.3a. Variable `simuList` saves the return values of the model simulation (i.e., the six simulated data sets). In Step 4, the cost function will calculate the mismatch between the simulated data sets (`simuList`) and the observed ones (`obsList` in Step 1).

STEP 4: COST FUNCTION

The cost function quantifies the discrepancy between simulated data sets (i.e., `simuList`) and

observed ones (i.e., `obsList`). In the TECO model simulation, function `run_model` calculates the seven C pool sizes over the course of five years, with results saved in variable `xsimu`. To be comparable with six observed data sets (`obsList`), `xsimu` needs to be converted to six simulated data sets (`simuList`) with mapping operators (i.e., `phi_litterfall`, `phi_slResp`, `phi_woodBiom`, `phi_foilageBiom`, `phi_cForestFloor`, `phi_cMineral` in Figures 24.4 and 24.5). Before mapping, function ‘`run_model`’ needs to update the values of `Phi_slResp` and `Phi_litterfall` mapping operators with parameter values `c` (Figure 24.5b). After mapping, the mismatch between `simuList` and `obsList` is calculated according to:

$$P(Z|c) \propto \exp \left\{ - \sum_{i=1}^6 \frac{1}{2\sigma_i^2} \sum_{t \in \text{obs}(Z_i)} [Z_i(t) - \phi_i X(t)]^2 \right\} \quad (24.2)$$

where $P(Z|c)$ is conditional probability density of observations Z on parameters c , i.e., the likelihood function of parameter c ; σ_i is the ith

```

50 ## Step 3: TECO model
51 def run_model(c):
    : : : : :
69 ## simulated carbon pools with 7 rows and Nt columns
70 xsimu=x[:,range(1,Nt+1)]
71 ## use mapping function to covert 7 carbon pools to 6 simulated data sets
72 ## get yearly simulated litterfall
73 litterDaily=tau[:-1] * np.dot(phi_litterfall, xsimu)
74 litterYr = [sum(litterDaily[range((i-1)*365, i*365)]) for i in range(1,6)]
75 ## get simulated soilResp
76 # convert float to int. In python, index starts from 0.
77 soilTime=soilResp[0,:].astype(int)-1
78 soilResp_simu=np.asarray(tau)[soilTime]*np.dot(phi_slResp,xsimu
    [:(soilTime)]+0.25*(1-b[0]-b[1])*u[soilTime]
79 ## get simulated woody biomass
80 woodyTime=woody[0,:].astype(int)-1
81 woody_simu = np.dot(phi_woodBiom, xsimu[:, woodyTime])
82 ## get simulated foliage biomass
83 foliageTime=foliage[0,:].astype(int)-1
84 foliage_simu=np.dot(phi_foilageBiom, xsimu[:, foliageTime])
85 ## get simulated Forestfloor biomass
86 forestFloorTime = forestFloor[0, :].astype(int) - 1
87 forestFloor_simu = np.dot(phi_cForestFloor, xsimu[:, forestFloorTime])
88 ## get simulated ForestMinearal biomass
89 cMineralTime = forestMineral[0, :].astype(int) - 1
90 forestMineral_simu = np.dot(phi_cMineral, xsimu[:, cMineralTime])
91 return [woody_simu, foliage_simu, litterYr, forestFloor_simu, forestMineral_simu,
    soilResp_simu]

```

Figure 24.4. Python code for mapping seven pool sizes (`xsimu`) to six simulated data sets (`woody_simu`, `foliage_simu`, `litterYr`, `forestFloor_simu`, `forestMineral_simu`, and `soilResp_simu`) from the TECO model simulation.

```

(a)
224     ## The mappings
225     phi_slResp=[0.25*c[0],0.25*c[1],0.55*c[2],0.45*c[3],0.7*c[4],0.55*c[5],0.55*c[6]]
226     phi_woodBiom=[0,1,0,0,0,0,0]
227     phi_foilageBiom=[0.75,0,0,0,0,0,0]
228     phi_litterfall=[0.75*c[0],0.75*c[1],0,0,0,0,0]
229     phi_cMineral=[0,0,0,0,1,1,1]
230     phi_cForestFloor=[0,0,0.75,0.75,0,0,0]

(b)
247     ## only update 2 mapping when new parameter values generate
248     phi_slResp=[0.25*c_new[0],0.25*c_new[1],0.55*c_new[2],0.45*c_new[3],0.7*c_new
      [4],0.55*c_new[5],0.55*c_new[6]]
249     phi_litterfall=[0.75*c_new[0],0.75*c_new[1],0,0,0,0,0]

```

Figure 24.5. The Python code for (a) defining the mapping operators; (b) updating mapping operators according to parameter values (c_{new}).

```

250     ## Step 3: TECO model
251     ## running model, return [woody_simu, foliage_simu, litterYr, forestFloor_simu,
      forestMineral_simu, soilResp_simu]
252     simuList = run_model(c_new)
253
254     ## Step 4: Cost Function
255     J_new=sum([sum((simuList[i]-obsList[i])**2)/(2*varList[i]) for i in range(6)])

```

Figure 24.6. Python code for calculating the mismatch between simulated ($simuList$) and observed datasets ($obsList$) based on cost function.

observation variance; $obs(Z_i)$ is the time series of i th observation; $Z_i(t)$ is the i th observation at time t ; φ_i is the mapping operator; $X(t)$ is the simulated seven pool sizes at time t ; $\varphi_i X(t)$ is the i th simulated data set after mapping.

The value of mismatch is saved in the variable J_{new} (Figure 24.6). Step 5 (Optimization) will use the value of J_{new} to decide whether the set of parameter values should be accepted or not.

STEP 5: OPTIMIZATION METHOD

Our study uses the Metropolis-Hasting method to draw parameter samples from their prior distribution. This method iteratively executes two phrases (i.e., proposing phase and moving phase) until a preset iteration number (e.g., 20,000) is reached. The proposing phase is implemented by function *GenerateParamValues*, which generates new parameter values (i.e., variable c_{New}) based the current accepted values (i.e., variable c_{opt}) (Figure 24.7a). During this process, function *GenerateParamValues* will call another function *isQualified* to assess whether the newly proposed parameter values (c_{New}) are in the reasonable parameter range or not. If the

isQualified function returns TRUE, c_{New} is a reasonable new set of parameter values and the generation of new parameter values will stop. Otherwise, function *GenerateParamValues* will discard this set of parameter values, generate new parameter values, assign these values to c_{New} , and call function *isQualified* again. Following the proposing phase, c_{New} will be used to update mapping operators (i.e., Φ_{slResp} and $\Phi_{litterfall}$) and run model simulations through calling function *run_model* (Figure 24.7b). The mathematical mechanism behind the proposing phase is available in Xu et al. (2006).

The moving phase decides whether the new parameter values (i.e., c_{New}) are accepted or not (Figure 24.8).

The value of J_{new} (i.e., mismatch between the simulated (c_{New}) and observed datasets) is compared with $J_{record}[record]$ (i.e., the mismatch using the previously accepted parameter value $c_{record}[record]$). The c_{New} will be accepted if J_{new} is smaller than $J_{record}[record]$, or the value $\exp(J_{record}[record] - J_{new})$ is larger than a random number ($randNum$) from the uniform distribution between 0 and 1. If accepted, the new parameter values are saved into an array, c_{record} . The

(a)

```
28 def GenerateParamValues(c_op):
29     """Generate new parameter values based on eigvalue and eigvectors
30     Global args: eigD, eigV, cmin, cmax, paramNum"""
31     flag = True
32     while (flag):
33         # Normally distributed pseudorandom numbers
34         randVector = np.random.randn(paramNum)
35         cT = randVector * np.sqrt(eigD)
36         cNew = np.dot(eigV, (np.dot(eigV.T, c_op) + cT))
37         if (isQualified(cNew)):
38             flag = False
39     return cNew
40
41 def isQualified(c):
42     """Decide whether the new parameter values exist in [cmin, cmax] value interval
43     Global args: paramNum, cmin, cmax"""
44     flag = True
45     for i in range(paramNum):
46         if(c[i] > cmax[i] or c[i] < cmin[i]):
47             flag = False
48             break
49     return flag
```

(b)

```
245 ## Proposing step: generate a new set of parameter values based on current accepted
      parameter values
246 c_new = GenerateParamValues(c_record[:, record])
247 ## only update 2 mapping when new parameter values generate
248 phi_slResp=[0.25*c_new[0],0.25*c_new[1],0.55*c_new[2],0.45*c_new[3],0.7*c_new
      [4],0.55*c_new[5],0.55*c_new[6]]
249 phi_litterfall=[0.75*c_new[0],0.75*c_new[1],0,0,0,0,0]
250 ## Step 3: TECO model
251 ## running model, return [woody_simu, foliage_simu, litterYr, forestFloor_simu,
      forestMineral_simu, soilResp_simu]
252 simuList = run_model(c_new)
```

Figure 24.7. Python codes for (a) defining the function of proposing phase in Metropolis-Hasting algorithm; (b) calling the function to generate parameter samples.

```
254 ## Step 4: Cost Function
255 J_new=sum([sum((simuList[i] - obsList[i])**2) / (2 * varList[i]) for i in range(6)])
256 delta_J = J_record[record] - J_new
257 ## Moving step: to decide whether the new set of parameter values will be accepted or not
258 randNum = np.random.uniform(0, 1, 1)
259 if (min(1.0, np.exp(delta_J)) > randNum):
260     ## accept the new set of parameter values and update relevant variables
261     record += 1
262     c_record[:, record] = c_new
263     J_record[record] = J_new
264     ## print out the acceptance rate
265     print('simu=' + str(simu) + ' accepted=' + str(record))
```

Figure 24.8. The Python code of the moving phase in Metropolis-Hasting algorithm.

corresponding mismatch will be saved in another array, J_record . The count of accepted parameter values, $record$, is then increased by 1. Therefore, $c_record[record]$ and $J_record[record]$ indicate the

current last element in the two arrays, which also represent the currently accepted parameter values and corresponding mismatch. If the new parameter values from this iteration are not accepted, they

are discarded. Whether *cNew* is accepted or not, the next iteration always uses the currently accepted parameter values (*c_record[record]*) for the proposing phase to generate a new set of parameter values.

STEP 6: ESTIMATE PARAMETERS

The outputs of DA are accepted parameter values (*c_record*), corresponding mismatches (*J_record*), the number of accepted parameter values (*record*), the optimal parameter values through maximum likelihood estimation (*bestC*) and the simulated data sets given the optimal parameters (*bestSimu*). All these outputs are saved to text files through calling function *write_io_file* (Figure 24.9). All parameter values accepted are saved in 'param_accepted.txt', all mismatches with parameter values accepted in 'mismatch_accepted.txt', the number of accepted parameter values in 'accepted_num.txt', the optimal parameter values in 'bestParam.txt', and simulation outputs given the optimal parameter values in 'xxx_bestSimu.txt' (e.g., *Woody_CestSimu.txt*) in the output directory folder (*outDir*).

Posterior distributions of parameters are the constrained parameter range after DA, which are often used for estimating the parameter uncertainty. An R script is provided to plot the posterior distributions with all accepted parameter values in 'param_accepted.txt'. The peak of the distribution represents the optimal parameter value. The simulated data sets with these optimal parameter values are also compared with observations in the R script. After DA, 'Probabilistic_inversion.py' runs the R script automatically and saves plots into the '/figures' folder.

STEP 7: PREDICTION

Parameter uncertainty as expressed by the posterior distribution of parameters will translate into prediction uncertainty characterized by the cumulative probability distribution of simulated C pool sizes. Prediction in the study of Xu et al. (2006) uses two functions: *prediction* and *forward_run* (Figure 24.10a). The function *prediction* is the start point of the prediction step. First this function generates 12,000 samples (*c_all*) from the accepted parameter values saved under *outDir* folder. Then function *prediction* uses each sample (*c*) of these 12,000 sets of parameter values for model forward simulation through calling function *forward_run*. The variable *x_record* saves all simulated pool sizes (Line 128 in Figure 24.10a). Finally, the results of prediction (i.e., *x_record*) will be written to the file 'prediction.txt' in the *outDir* folder. To call function *prediction*, we only need to provide the directory of DA results (*outDir*) as shown in Figure 24.10b.

With each set of parameter values (*c*) sampled in function *prediction*, function *forward_run* executes the model simulation over the 10 years following the DA period (i.e., 2001 to 2010) using Equation 24.1. This function is similar to function *run_model* in Step 3 except that the simulation times are different. The environmental scalar (*tau_forecast*) and C input from photosynthesis (*u_forest*) from 1996 to 2000 were replicated twice to provide environmental 'forcing' for 2001 to 2010. Variable *x* stores the seven simulated C pool sizes in each daily simulation step. The return value of *forward_run* is the simulated pool sizes at the end of year 2010 (*xsimu*). Line 128 in Figure 24.10a shows an example of calling *forward_run* in function *prediction*.

```

102 def write_io_file(outDir):
103     """write outputs to files.
104     Global args: c_record, J_record, record, bestC, bestSimu"""
105     np.savetxt(outDir+'/mismatch_accepted.txt', J_record[1:record])
106     np.savetxt(outDir+'/param_accepted.txt', c_record[:, 1:record])
107     np.savetxt(outDir+'/accepted_num.txt', [record])
108     np.savetxt(outDir+'/bestParam.txt', bestC)
109     np.savetxt(outDir+'/Woody_bestSimu.txt', bestSimu[0])
110     np.savetxt(outDir+'/Foliage_bestSimu.txt', bestSimu[1])
111     np.savetxt(outDir+'/Litterfall_bestSimu.txt', bestSimu[2])
112     np.savetxt(outDir+'/Forestfloor_bestSimu.txt', bestSimu[3])
113     np.savetxt(outDir+'/ForestMineral_bestSimu.txt', bestSimu[4])
114     np.savetxt(outDir+'/SoilResp_bestSimu.txt', bestSimu[5])

```

Figure 24.9. The Python code for saving DA outputs to text files.


```

(a)
116 def prediction(outDir):
117     nsample = 12000
118     # check whether two files exist
119     if(os.path.isfile(outDir+"/param_accepted.txt") and
        os.path.isfile(outDir+"/accepted_num.txt")):
120         c_record = np.loadtxt(outDir+"/param_accepted.txt")
121         record = int(np.loadtxt(outDir+"/accepted_num.txt"))-1
122         # generate 12,000 random integer ranging from 1 to record
123         sampleId = np.random.randint(1,record,nsample)
124         c_all = c_record[:,sampleId] # 12,000 samples of p(c|Z)
125         x_record = np.zeros([7,nsample], dtype=float)
126         for i in range(nsample):
127             c = c_all[:,i]
128             x_record[:,i] = forward_run(c)
129             print('the '+str(i+1)+'th sampling, total '+ str(nsample))
130         # save the predicted pool sizes
131         np.savetxt(outDir+'/prediction.txt', x_record)
132     else:
133         # Can't find param_accepted.txt and accepted_num.txt in the output folder
134         print('Warning: Please finish exercise 1 first!')
135
136 def forward_run(c):
137     """prediction with the parameter values c
138     Global args: tau, b, u, x0, Nt, cbnScale
139     """
140     ## environmental scalar and C input duplicate to represent 2000 to 2010
141     Nt_forecast = 365*10
142     tau_forecast = np.tile(tau,2) # repeat twice
143     u_forecast = np.tile(u,2)
144     ## x stores 7 simulated carbon pools from 2000 to 2010 (daily scale)
145     x=np.zeros([7,Nt_forecast+1], dtype=float)
146     ## multiple matrix A and diagonal matrix C
147     AC = np.matrix([[ -c[0], 0, 0, 0, 0, 0, 0],
148                    [0, -c[1], 0, 0, 0, 0, 0],
149                    [0.7123 * c[0], 0, -c[2], 0, 0, 0, 0],
150                    [0.2877 * c[0], c[1], 0, -c[3], 0, 0, 0],
151                    [0, 0, 0.45 * c[2], 0.275 * c[3], -c[4], 0.42 * c[5], 0.45 * c[6]],
152                    [0, 0, 0, 0.275 * c[3], 0.296 * c[4], -c[5], 0],
153                    [0, 0, 0, 0, 0.004 * c[4], 0.03 * c[5], -c[6]]])
154     x[:,0] = x0
155     ## simulate carbon pools from 2000 to 2010
156     for i in range(1,Nt_forecast+1):
157         ## Eq. 1 in Xu et al., (2006)
158         x[:,i]=np.dot((np.eye(7)+AC*tau_forecast[i-1]),x[:,i-1])+np.asarray(b)*
            u_forecast[i-1]*cbnScale
159     ## simulated carbon pools at 2010
160     xsimu=x[:,Nt_forecast]
161     return xsimu
(b)
273 if ninput == 1 and sys.argv[2] == "ParamRange.txt" and enable_prediction == 1:
274     prediction(outDir)
275 elif ninput == 2 and sys.argv[2] == "ParamRange.txt" and enable_prediction == 1:
276     prediction(outDir)

```

Figure 24.10. Python code for (a) defining functions for prediction; (b) calling functions for prediction.

EXERCISES WITH CARBOTRAIN TOOLBOX

The following three exercises using the CarboTrain toolbox will help readers become familiar with the DA methodology described above. Exercise 1 is to conduct DA with the TECO model separately for ambient and elevated CO₂ treatments. Based on the results, we will generate figures similar to Figures 3, 4 and 8 in Xu et al. (2006). Exercise 2 uses the

optimised parameter values from Exercise 1 to predict soil C pool sizes. The expected figure is similar to Figure 9 in Xu et al. (2006). The Posterior distributions of parameters c_1 , c_2 and c_4 are bell-shaped in Figures 3 and 4 of Xu et al. (2006). In Exercise 3 we will re-conduct DA under the ambient CO₂ treatment using enlarged prior parameter ranges for parameters c_3 , c_5 , c_7 . Results similar to Figure 5 in Xu et al. (2006) are expected.

EXERCISE 1

1. Open CarboTrain → A dialog appears (Figure 24.11) → Select Unit 6 and Exercise 1 → Choose ambient CO₂ → Choose an output directory on your computer → Click 'Run Exercise' → A dialog appears (Figure 24.12a) → Click 'OK'
2. A series of outputs will be printed in another window (Figure 24.13). The variable *simu* is the number of total executed simulations while the *accepted* variable means the number of simulations with accepted parameter values. These numbers will keep increasing until *simu*

reaches a value approximating 20,000. A dialog will pop up to notify that DA is finished. Execution time is about 20 minutes depending on the specifications of your computer.

3. After DA, a dialog appears to notify that the task is complete (Figure 24.12b). Open the output directory you specified in Step 1. There will be four folders: '/practice_1_ambient', '/practice_1_elevated', '/practice_2', '/practice_3_ambient'. Open the '/practice_1_ambient' folder. You will find text files and figures generated in the subfolder '/'

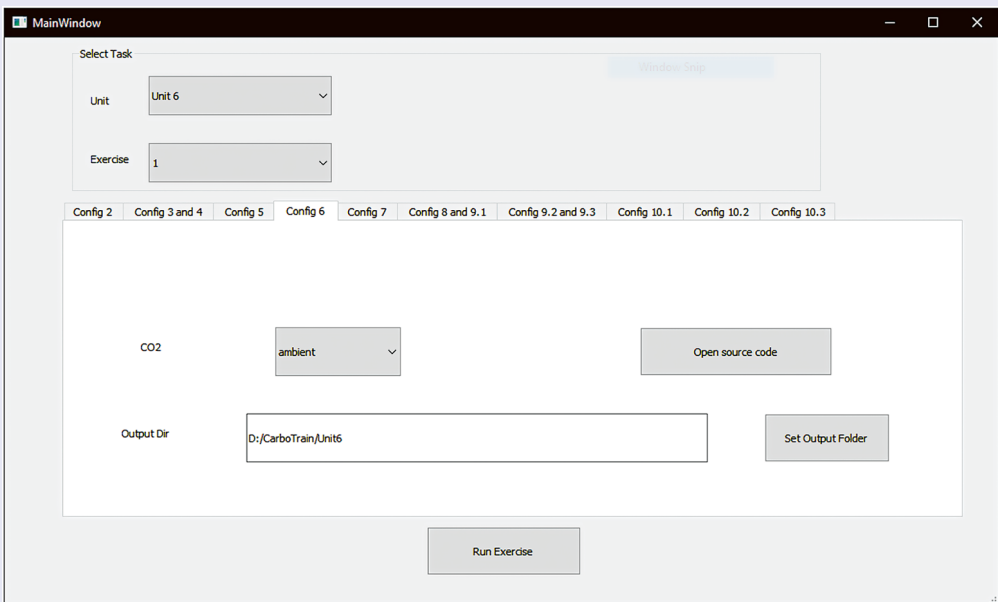


Figure 24.11. The main window in CarboTrain toolbox for this practice.

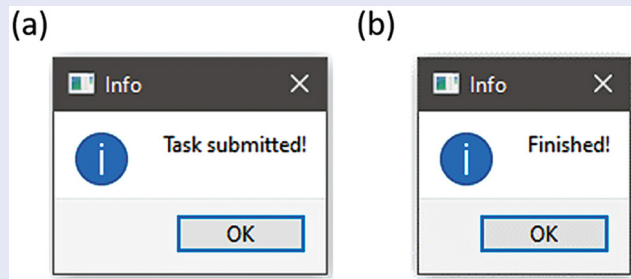


Figure 24.12. The dialog notifies (a) submitting a new task; (b) finishing a task.

```

Administrator: Command Prompt...
simu=350 accepted=151
simu=351 accepted=152
simu=352 accepted=153
simu=353 accepted=154
simu=354 accepted=155
simu=355 accepted=156
simu=359 accepted=157
simu=361 accepted=158
simu=370 accepted=159
simu=371 accepted=160
simu=373 accepted=161
simu=374 accepted=162
simu=377 accepted=163
simu=383 accepted=164
simu=386 accepted=165
simu=388 accepted=166
simu=393 accepted=167
simu=395 accepted=168
simu=396 accepted=169
simu=413 accepted=170
simu=415 accepted=171
simu=419 accepted=172
simu=426 accepted=173
simu=428 accepted=174
simu=429 accepted=175
simu=430 accepted=176
simu=431 accepted=177
simu=432 accepted=178
simu=434 accepted=179
simu=436 accepted=180
simu=437 accepted=181
simu=441 accepted=182

```

Figure 24.13. Window in CarboTrain toolbox to display the progress of the DA experiment’.

figures’. These files and figures are results of DA under the ambient CO₂ treatment.

- Repeat Steps 1–3 but choose elevated CO₂ in Step 1. Do not change the output directory. After DA, open the ‘/practice_1_elevated’ folder in the output directory. The files and figures in this folder are results of DA under the elevated CO₂ treatment. Compare the results with Figures 3, 4 and 8 in Xu et al. (2006).

QUESTIONS:

What is the acceptance rate in each exercise? Which parameters are well constrained? Which observations contribute to the constrained parameter values? How many shapes are there in the posterior distributions (e.g. bell shape)? What are the meanings behind these different shapes?

EXERCISE 2

- In the main window of CarboTrain, select Unit 6 and Exercise 2 → Use the default output folder as for Exercise 1 → Click ‘Run Exercise’.
- Open the output folder and find a figure generated in the ‘/practice_2’

folder. Compare this figure with Figure 9 in Xu et al. (2006).

QUESTIONS:

Does elevated CO₂ positively influence C accumulation (i.e., larger pool sizes)? Do poorly constrained parameters (i.e., c₃, c₅ and c₇) influence the predicted sizes of associated pools (i.e., X₃, X₅ and X₇)?

EXERCISE 3

- Repeat Steps 1–3 of Exercise 1 but choose ambient CO₂ and Exercise 3 in the CarboTrain main window (Figure 24.11).
- After DA, go to the output directory and open the ‘/practice_3_ambient’ folder. The files and figures generated in this folder are results of DA using enlarged prior parameter ranges for c₃, c₅

and c₇. Compare the results with Figure 5 in Xu et al. (2006).

QUESTIONS:

Did the posterior distributions of c₃, c₅ and c₇ change after their prior parameter ranges were extended? What else can we do to further constrain the parameter uncertainty?

SUGGESTED READING

Xu, T., White, L., Hui, D. and Luo, Y., 2006. Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction. *Global Biogeochemical Cycles*, 20(2).

Huang, X., D Lu, DM Ricciuto, PJ Hanson, AD Richardson, XH Lu, ES Weng, S Nie, LF Jiang, EQ Hou, IF Steinmacher, YQ Luo. 2021. A model-independent data assimilation (MIDA) module and its applications in ecology. *Geoscientific Model Development*, 14: 5217–5238.

UNIT SEVEN

Data Assimilation with
Field Measurements and
Satellite Data



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER TWENTY-FIVE

Model-Data Integration at the SPRUCE Experiment

Daniel Ricciuto

Oak Ridge National laboratory, Oak Ridge, USA

CONTENTS

Introduction /	209
Site Description /	210
Modeling for the SPRUCE Experiment /	211
Model Validation and Uncertainty Quantification /	212
Suggested Reading /	214
Quizzes /	214

This chapter is intended as a brief introduction to carbon-cycle modeling and field measurements at the Spruce and Peatland Response Under Changing Environments (SPRUCE) experiment in a forested wetland in northern Minnesota. The goal is to familiarize the reader with the study in preparation for subsequent training and example applications of data assimilation into models using SPRUCE data.

INTRODUCTION

SPRUCE is a large-scale, decade-long experiment designed to assess the response of a northern peatland bog ecosystem, which contains a large amount of carbon belowground, to changes in atmospheric temperature and carbon dioxide (CO₂) concentrations that approximate possible conditions in the latter half of the 21st century. Peatlands have been identified as vulnerable ecosystems that potentially have large feedbacks to the global carbon cycle, and as a result may affect climate change. Although peatlands comprise a relatively small fraction, about 3%, of global land area, they are estimated to contain at least one-third of all carbon on the land surface. Therefore, understanding the response of these systems to warming, changing moisture conditions, and rising

atmospheric CO₂ is critical to our ability to predict future climate using coupled Earth system models.

One key uncertainty is how carbon may leave the system to the atmosphere, as either CO₂ or as methane gas (CH₄). This is critically important because CH₄ is a much more potent greenhouse gas than CO₂, with more than 25 times greater warming potential over a 100-year timeframe. The saturated peat, biogeochemical environment and the types of vegetation in peatlands are conducive for anaerobic decomposition (breakdown of organic matter in the absence of oxygen) and therefore high levels of CH₄ emissions compared to other ecosystems. While CO₂ fluxes from aerobic decomposition (when oxygen is available) are sensitive to temperature, CH₄ production and emissions may be even more sensitive to warming than CO₂. On the other hand, drying associated with warming conditions may reduce methane production and increase methanotrophy (consumption of CH₄ by microbes in the peat). Observations of CO₂ and CH₄ fluxes from SPRUCE are informing model parameter values and structural representations of key processes in mechanistic models to improve predictive understanding of the system.

SPRUCE researchers are also very interested in how the experimental treatments impact the different types of peatland vegetation in terms

of physiology, phenology (timing and length of growing season), reproduction and mortality. The most extreme level of warming in the experiment is consistent with climate model projections at the end of the century under the strongest greenhouse gas emissions scenarios. It effectively shifts the climate of SPRUCE to that currently experienced in central Missouri, which is hundreds of kilometers to the south.

SITE DESCRIPTION

The SPRUCE experiment is located in a forested wetland called the S1 bog, which is part of the United States Forest Service (USFS) Experimental Forest in North Central Minnesota. The S1 system is a raised-dome ombrotrophic bog, meaning that it is rain-fed, nutrient-poor and has a perched water table that is disconnected from the influence of regional groundwater. The bog has a mean annual air temperature of 3.3°C and mean annual precipitation of 768 mm. The bog experiences cold winters: snow cover generally persists from late autumn until early spring, and ice layers form in the peat that may persist until May or June.

Within the bog, there is microtopography consisting of raised areas (hummocks) and sunken areas (hollows). Hummock areas are generally 15–20 cm higher than hollows and are almost never inundated. In typical years, the water table generally ranges from as low as 20–30 cm below the hollow surface in dry conditions to 10–15 cm above the hollow surface in wet conditions. At the beginning of the study, both hummock and hollow surfaces of the bog were nearly completely covered with *Sphagnum* mosses. Above that, there is a woody shrub layer dominated by two species: *Rhododendron groenlandicum* (Labrador tea) and *Chamaedaphne calyculata* (leatherleaf). There are two main types of trees on the S1 bog, *Picea mariana* (black spruce) and *Larix laricina* (larch). Existing trees were cleared in strip cuts in 1974 for a different experiment, and new trees have been regrowing since then. This area with relatively young, short trees within the strip cut areas is ideal for the SPRUCE experiment because the enclosures do not have to be as large or as costly to expose these trees to whole-ecosystem warming.

SPRUCE uses an open-top enclosure design (Figure 25.1) in which the whole-ecosystem



Figure 25.1. Top-down view of a SPRUCE enclosure.

warming and elevated CO₂ treatments are conducted (Hanson et al., 2017a and 2017b). Air warming is accomplished using propane-fired furnaces in combination with blowers distributed around the enclosure. Peat warming is accomplished using resistance heaters that heat depths between 2 and 3 meters below the surface. A corral system isolates the hydrology within the enclosures, so that the water table within may be lower or higher than the surrounding bog.

SPRUCE has a total of ten enclosures with five different levels of warming ranging from no added heat (+0°C) to +9°C in 2.25°C increments. Half of the enclosures have ambient CO₂ concentrations (about 400 parts per million) at the five warming levels, while the other half have an elevated CO₂ concentration target of +500 parts per million (ppm) over ambient, typically ranging between 800 and 900 parts ppm at the same warming levels. Elevated CO₂ is supplied only during daytime and in the growing season when photosynthesis is occurring. No water vapor is added, causing reduced relative humidity and increased vapor pressure deficit in the warmed enclosures. Extensive measurements within the enclosures include meteorological conditions, peat temperature and water table depth, CO₂ and CH₄ fluxes using a large-collar chamber measurement (Hanson et al., 2016), species-level vegetation biomass and productivity, phenology cameras and porewater chemistry. In addition to the active warming treatments, there are passive enclosure effects including increased longwave radiation input, decreased shortwave radiation and changes in air flow, resulting in additional temperature increases between 1°C and 2°C at all warming levels compared to outside the enclosures. Two additional plots without enclosures are also measured to record the evolution of the ecosystem under ambient conditions. The study's regression-based design allows for the determination of response curves over a range of conditions and facilitates comparison with model outputs.

Lengthy preparation was required for the large-scale experiment. Pretreatment characterization at the S1 bog began in 2009, involving extensive peat and vegetation measurements. Construction for the treatment experiments began in 2012, beginning with roadwork and other infrastructure development. The corral systems were then built, followed by the construction of the enclosures. Whole-ecosystem warming began in August 2015, and

elevated CO₂ treatments began in June 2016. The experiment is anticipated to run through 2025. A number of key science questions were posed at the beginning of the experiment:

- Are peatland ecosystems and organisms vulnerable to atmospheric and climatic change?
- At what rate will ancient carbon be released from accumulated peat in response to deep belowground warming, and what is the relative release of CO₂ compared with CH₄?
- What are the interactions between ecosystem responses to warming and the availability of nutrients and water?
- How does elevated CO₂ modify ecosystem responses to warming and the availability of nutrients and water?

MODELING FOR THE SPRUCE EXPERIMENT

Although the scale of SPRUCE is unprecedented for a peatland ecosystem manipulation experiment, it is difficult to know if the answers to these questions at the study site will be consistent across other peatland systems. The project investigators therefore envision mechanistic modeling as a key method to extrapolate the results of SPRUCE to other systems. This requires a multi-scale modeling framework that can be applied from site to global scales.

The United States Department of Energy (DOE), which provides the primary funding for SPRUCE, has heavily invested in such a modeling framework which we are using. In 2014, the DOE initiated development of a new Earth system model, the Energy Exascale Earth System Model (E3SM). This model branched from the well-known Community Earth System Model (CESM). The land component of E3SM, known as ELM, is a land-surface model that includes cycling of water, carbon, nitrogen and phosphorus. ELM has been used extensively in coupled E3SM (Burrows et al., 2020), uncoupled (land-only simulations driven by observed atmospheric forcings), and at the site-level including eddy covariance and ecosystem manipulation sites like SPRUCE. However, the default version of ELM lacks key processes necessary to simulate peatland carbon, water and nutrient dynamics accurately. A SPRUCE-specific version, ELM-SPRUCE, was recently developed (Shi et al., 2015, 2021). It incorporates wetland hydrology, microtopography, and plant functional types that are not currently

represented in ELM including *Sphagnum* mosses. ELM-SPRUCE also includes a more mechanistic CH₄ model that explicitly represents microbial populations involved in CH₄ production and consumption. ELM-SPRUCE can help to answer the above research questions, test other hypotheses about the impact of environmental change at SPRUCE, and eventually simulate broader peatland regions on a global scale.

In addition to ELM-SPRUCE, other modeling groups are also simulating the experimental treatments at the site. The Terrestrial Ecosystem (TECO) model was introduced in Chapter 2. It has well-developed methods for model-data integration. A version of the TECO model has been developed for application at SPRUCE (Ma et al., 2017). This model, TECO-SPRUCE, is being used to make projections of CO₂ and CH₄ fluxes under the experimental treatments and to estimate the magnitudes of prediction uncertainties that result from uncertainties in forcings and model parameters (see Chapter 26).

In general, having projections from multiple models is an important way to understand the impact of structural uncertainty, which reflects the different ways in which different model frameworks may represent processes. Some models may be more complex than others by including more processes, or they may use different equations or algorithms to represent a specific process. A model intercomparison study focused on SPRUCE is currently under development, and SPRUCE data are being made available to interested modeling groups.

MODEL VALIDATION AND UNCERTAINTY QUANTIFICATION

While having projections from multiple models is useful, quantitatively estimating within-model uncertainty is key to knowing the confidence in our predictions, and for understanding what measurements may be most useful in constraining these predictions further. Model uncertainty quantification (UQ) is a key goal of the SPRUCE modeling work, and it has been used in both the ELM-SPRUCE and TECO-SPRUCE models. An important part of UQ is estimating how uncertain model parameters contribute to uncertainty in predictions such as CO₂ fluxes or stocks. A model such as ELM-SPRUCE is very complex and contains well over 100 uncertain model parameters. An example of an uncertain parameter would be

the sensitivity of heterotrophic respiration to temperature (Q_{10}), for which published values in the literature may range between 50% lower or higher than the assumed model default value. This parameter uncertainty is likely to cause large uncertainties in the predicted response of net CO₂ fluxes to experimental warming. Ultimately, we would like to calibrate such parameters and reduce their uncertainty using measurements; for example, the within-enclosure CO₂ flux information could be used to constrain the Q_{10} value at SPRUCE. However, as the number of uncertain parameters grows, the computational expense of model calibration rises exponentially as it requires a larger and larger number of model simulations (also referred to as ensemble members) to sample the potential parameter space. Therefore, we usually need to identify the most important parameters first using sensitivity analysis.

Sensitivity analyses are usually conducted for a set of model outputs, or quantities of interest (QoIs). For example, a QoI might be the site-averaged net ecosystem exchange over a 10-year period, or the average date of leaf-out in spring. The contribution of each parameter to the variance of a QoI may be estimated in a sensitivity analysis using a model ensemble in which multiple parameters are varied randomly. Fortunately, a smaller number of ensemble members is necessary for sensitivity analysis than for calibration. The objective of the sensitivity analysis is to reduce the number of calibration parameters to a reasonable number, usually around 20 or less. In the ELM-SPRUCE model, we conduct the sensitivity analysis first by identifying minimum and maximum possible values for each parameter. This can be done by surveying the literature, trait databases such as TRY and the Fine Root Ecology Database (FRED), or by making educated guesses about parameter uncertainty (e.g., +/- 25% from default values). An ensemble of model parameter values is then created by randomly sampling parameter values in these ranges. This model ensemble is used to create a surrogate model, or model response surface for each QoI. Many different approaches can be used to create surrogate models, including machine learning; here, we use polynomial functions. Using this approach, we found that about 2500 ensemble members can yield trustworthy sensitivities for about 65 uncertain parameters in ELM (Ricciuto et al., 2018). While running this number of simulations is computationally feasible in ELM

on a mid-size computing cluster, other models may be more or less computationally expensive, allowing for different ensemble sizes.

Model calibration involves finding a set or sets of optimal parameters that best fit observations by minimizing a cost function (Chapter 21). A cost function typically yields a single value that can integrate information from multiple types of observations (e.g. both CO₂ fluxes and biomass measurements) and from observations at multiple times. Observations may be weighted by their uncertainties or using other methods. Model calibration may also involve the estimation of parameter and prediction uncertainties. It may be accomplished using a variety of techniques. A preferred technique is Markov Chain Monte Carlo (MCMC), introduced in Chapter 22. MCMC has the desirable quality that it can calculate the full parameter posterior probability density functions (PPDFs) without making any prior assumptions about the functional forms of the distributions. These PPDFs may also be used to estimate prediction uncertainty for QoIs. However, it is a relatively expensive method that requires a large number of model evaluations (usually at least 10,000) and is not easily parallelized. In models that are fast to evaluate such as TECO-SPRUCED, MCMC may be used directly. However, in expensive models such as ELM-SPRUCED, it is first necessary to construct a surrogate model. An example of surrogate model calibration in ELM is given by Lu et al. (2018a). Similar methods are used as is done for the surrogates used in sensitivity analysis. However, the surrogate models introduce error into the calibration and therefore must be considerably more accurate to ensure a trustworthy calibration result.

Using ELM-SPRUCED, model calibration was performed with pre-treatment observations of vegetation biomass and productivity for four different plant functional types (Shi et al., 2021). The calibrated model parameters differed substantially from the default ELM parameters used in the global parameterization for boreal plant functional types. The calibrated model was then used to predict treatment responses at the site for the most extreme warming scenario of +9°C. The model predicted that black spruce trees would steadily decline in biomass and productivity with warming, while the shrubs and larch would increase slightly. The *Sphagnum* productivity was simulated to decline during dry periods and increase during wet periods. We can now begin to validate the model using

treatment observations. There is some indication that the black spruce trees are actually responding negatively to the warming treatments, especially in comparison to the shrubs and larch trees. However, a recent study showed that *Sphagnum* productivity and biomass are rapidly declining, which was not predicted by ELM-SPRUCED. This may be in part due to a lack of certain processes in ELM-SPRUCED. For example, more productive shrubs may shade out the moss layer, which cannot be represented currently in ELM-SPRUCED because there is no competition for light in that model. It is also possible that the model predictions will improve when we begin to use the treatment data for model calibration. We did find that ELM-SPRUCED predicts the change in net carbon flux with temperature quite accurately, and that both model and observations indicate that warming causes a significant source of CO₂ and CH₄ to the atmosphere (Hanson et al., 2020). If one naively assumes that all peatland systems respond similarly over the entire globe, this would result in a large source of greenhouse gases and a positive feedback which would be large enough to further strengthen global warming. We will test this assumption in the model in the future by running global simulations. Interestingly, however, the observations do not yet indicate a strong effect of elevated CO₂ concentrations on vegetation biomass at SPRUCED. ELM-SPRUCED and TECO-SPRUCED before data assimilation both predict a strong fertilization effect. Reconciling this with observations will take additional empirical and model development work for ELM-SPRUCED and this will be informed by data assimilation using TECO-SPRUCED.

The experimental treatments at SPRUCED combined with the model-data integration framework in ELM-SPRUCED, TECO-SPRUCED and other models provide a useful testbed for improving predictive understanding of peatland ecosystems. The interaction of the site hydrology, biogeochemistry, and multiple vegetation types under varying treatments makes obtaining accurate predictions particularly challenging compared to other study sites, for example eddy covariance tower footprints which often have relatively homogeneous vegetation coverage. Recently it has been observed that responses of the different vegetation types to warming are not uniform; while shrubs are becoming more productive and growing more fine roots (Malhotra et al., 2020), the *Sphagnum* mosses are dying and sharply declining in areal

coverage under strong warming (Norby et al., 2019). The growing season generally becomes longer with warming, but some vegetation types have a stronger phenology response than others (Richardson et al., 2018). Ideally, models of the SPRUCE system must be simple enough to ensure simulations are relatively inexpensive so that we can run parameter ensembles to explore prediction uncertainty. On the other hand, models must contain enough process realism to capture the divergent responses of different vegetation types and to predict the strong increases in CO₂ and CH₄ surface fluxes to the atmosphere. Much research remains to be done to predict how the SPRUCE S1 bog and other peatlands will respond to rapidly changing environmental conditions.

SUGGESTED READING

Paul J. Hanson, Natalie A. Griffiths, Colleen M. Iversen, Richard J. Norby, Stephen D. Sebestyen, Jana R. Phillips, Jeffrey P. Chanton, Randall K. Kolka, Avni

Malhotra, Keith C. Oleheiser, Jeffrey M. Warren, Xiaoying Shi, Xiaojuan Yang, Jiafu Mao, Daniel M. Ricciuto. 2020. Rapid net carbon loss from a whole-ecosystem warmed Peatland. *AGU Advances*, doi:10.1029/2020AV000163

Shi, X., Thornton, P. E., Ricciuto, D. M., Hanson, P. J., Mao, J., Sebestyen, S. D., Griffiths, N. A., and Bisht, G. 2015. Representing northern peatland microtopography and hydrology within the Community Land Model, *Biogeosciences*, 12, 6463–6477, doi:10.5194/bg-12-6463-2015.

QUIZZES

1. Why are peatlands important for land-atmosphere feedbacks?
2. What is the gradient experimental design for this SPRUCE project?
3. What are the benefits of incorporating modeling approaches in such a large experimental study?
4. When modeling results are different from observations, what should we do?

Application of Data Assimilation to a Peatland Methane Study

Shuang Ma

Northern Arizona University, Flagstaff, USA

CONTENTS

Uncertainty in Methane Modeling / 215

Assimilation of Methane Emissions Data into the TECO Model / 216

Suggested Reading / 223

Quizzes / 223

Data assimilation is widely used in terrestrial ecosystem studies. This chapter illustrates the use of data assimilation with a site-level study to project peatland methane (CH_4) emission in response to warming. Wetland CH_4 emissions comprise one third of the global CH_4 source and remain the largest source of uncertainty in the global budget. Wetland CH_4 emission estimated by process-based models (bottom-up) are used as the prior information for atmospheric inversion estimates (top-down). It is thus important to constrain process-based models with *in situ* observations to improve both the bottom-up and top-down estimates. We give a brief background of methane modeling and then show the application of data assimilation in the methane model in seven steps.

UNCERTAINTY IN METHANE MODELING

Methane (CH_4) has 25 times the global warming potential of CO_2 over a 100-year scale (Myhre et al., 2013). It is directly responsible for approximately 20% of global warming since pre-industrial time (Forster et al., 2007). Wetlands are an important natural source of CH_4 emissions to the atmosphere, but constitute a principal source of uncertainty in the global atmospheric CH_4 budget (Saunio et al., 2020). Global wetland CH_4 emission estimated by process-based models have

large disagreement compared with atmospheric inversion model ensembles (Saunio et al., 2020).

There are three major sources of uncertainty in model estimated CH_4 emission. The first is the uncertainty in mechanisms that control biogeochemical processes due to the difficulty to acquire empirical data. For example, it is very difficult to measure the aerobic and anaerobic oxidation of methane. The redox potential effect on methane oxidation awaits more empirical data to be represented in models. The second source of uncertainty is the wide range of possible parameter values for methane-related processes. Flux-based measurements of Q_{10} (temperature sensitivity) of CH_4 release from different warming plots at one single site range from 2.12 to 32.16 (Gill et al., 2017). Manually tuning the parameter values to match the observed CH_4 fluxes could achieve the right answer with diverse combinations of parameter values – the problem of equifinality. The third uncertainty is a poorly mapped wetland extent and seasonal inundation. Current wetland maps are mainly based on inventory data and satellite observations. Inventory maps are limited due to the low spatial and temporal coverage, while satellite-based maps cannot capture the wetland area with dense vegetation cover.

It is critical to understand how wetland CH_4 emissions may respond to climate change, given the much larger warming potential of CH_4 compared

to CO₂ over a 100-year scale (Myhre et al., 2013). Terrestrial biosphere models that include methane processes explicitly describe the CH₄ flux exchange through plant-mediated transport, diffusion, and bubbling (ebullition). These are the three major pathways of wetland CH₄ emission. The relative contributions of these three pathways to methane emissions under climate warming have not been unraveled either using experiments or modeling approaches. In most process-based methane models, these CH₄ emission pathways are calculated based on CH₄ concentration in each peat layer, which is primarily dominated by CH₄ production. If some of the parameters in CH₄ production, plant-mediated transportation, ebullition, and diffusion can be constrained by observational data, we may be able to improve model predictions both by improving accuracy and reducing uncertainty.

ASSIMILATION OF METHANE EMISSIONS DATA INTO THE TECO MODEL

The seven steps of data assimilation were introduced in Chapter 21. As an illustrative example, we will apply these steps to the assimilation of methane emissions data from *in situ* measurement into the TECO model.

Step 1: Define the Objective.

Our objective is to reduce the uncertainty of model estimations of how methane emissions vary in response to warming. Taking the example of a peatland methane study (Ma et al., 2017), data assimilation is used first to constrain parameters with observational data, thereby reducing uncertainty in methane prediction.

Step 2: Prepare Data.

Our data come from the Spruce and Peatland Responses Under Changing Environments (SPRUCE) experiment at a northern peatland site in Minnesota, USA. The SPRUCE project uses experimental warming and elevated CO₂ treatments to assess the responses of northern peatland ecosystems to future environmental conditions (Hanson et al., 2016). Here we will use *in situ* net CH₄ emission data from ambient plots to constrain parameters of a process-based methane model. CH₄ emission measurements were acquired during the snow-free months using a portable open-path analyzer in a static chamber (1.13m² area), near

monthly from 2011 to 2016 (Hanson et al., 2017a and 2017b). The 2010–2014 data are used for data assimilation and 2015–2016 are used for validation. In total, 45 daily CH₄ chamber measurement data points were integrated from ambient plots from 2011 to 2016. A mean and a standard deviation are calculated from all the measurements on the same day in each ambient plot.

Step 3: Choose a Model.

Our example uses a methane-enabled version of the Terrestrial ECOsystem (TECO) model, which incorporates a ten-layer vertical mixing CH₄ module (Ma et al., 2017). The TECO model has been calibrated to the SPRUCE site to study the carbon cycle and soil thermal dynamics (Jiang et al., 2018). A detailed description of TECO can be found in Weng and Luo (2008).

TECO-CH₄ explicitly considers the transient and vertical dynamics of CH₄ production, CH₄ oxidation, and CH₄ transport from belowground to the atmosphere. The structure and processes are adapted from a number of previous studies and models such as CLM4.5 (Riley et al., 2011), LPJ-WHyMe (Wania, Ross, and Prentice, 2010), Walter's model (Walter & Heimann, 2000), and TEM (Zhuang et al., 2004). All the above models assume that soil can be separated into aerobic and anaerobic layers divided by the water table. These models also assume that the majority of CH₄ oxidation occurs in the aerobic layers and rhizosphere, and that most methane production occurs in the anaerobic layers. Within each soil layer, CH₄ concentration is calculated as the mass balance of CH₄ production (gain), CH₄ oxidation (loss), CH₄ diffusion from adjacent layers, CH₄ ebullition (loss), and plant-mediated transport (loss). The flow diagram of TECO-CH₄ is shown in Figure 26.1 and further described below.

CH₄ production is determined by carbon availability represented by heterotrophic respiration, and by soil environmental conditions such as water table height and soil temperature. As in most methane models, CH₄ production only occurs when soil temperature is above 0°C and below 45°C. Given that CH₄ oxidation is largely controlled by CH₄ concentration, it is assumed to follow the Michaelis-Menten kinetics.

The CH₄ diffusion across soil layers follows Fick's law, which relates the diffusive

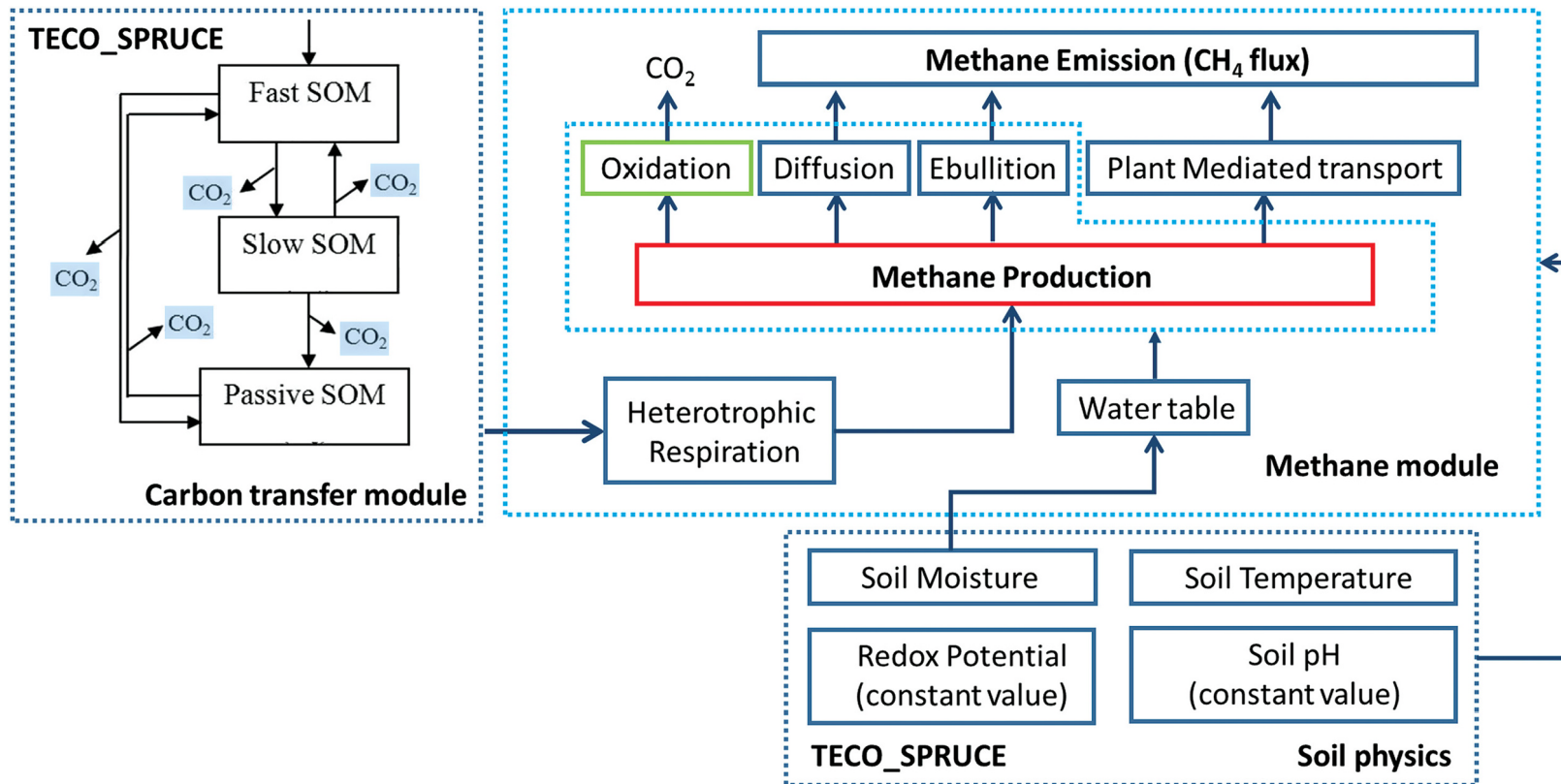


Figure 26.1. Flow diagram of CH_4 module and linkage to soil C model in TECO- CH_4 . Reproduced from Ma et al. (2017).

flux to the gradient of the concentration, and Henry's Law, which resolves the diffusive flux at the liquid-atmosphere boundary. The net exchange between the surface soil layer and atmosphere is accounted as the diffusive part of CH₄ emission (or uptake). The methane flux at the bottom boundary is set to zero and the atmospheric CH₄ concentration at the soil surface (or water surface if the water table is at or above the soil surface) is set to standard atmospheric CH₄ concentration.

Air-filled aerenchyma tissues of plants act as a chimney to quickly emit CH₄ from the rhizosphere directly into the atmosphere. A portion of CH₄ is oxidized within the plant tissue during the transport. TECO-CH₄ uses a parameter (T_{veg}) to represent the ability to transport CH₄ through tissues at a plant community level. The growth of plants also affects the amount of gas transported through the influence of Leaf Area Index (LAI). Ebullition entails the formation of bubbles when the CH₄ concentration exceeds a certain threshold and directly emits into the atmosphere if the water table is above the soil surface, bypassing the aerobic zones that lead to CH₄ consumption. The bubbles are otherwise added to the soil layer just above the water table and then diffuse through the upper layers if the water table is below the soil surface.

Once a model is chosen, the next step is to choose parameters for optimization. The performance of data assimilation is affected by the variety and amount of observational data as well as the parameters that are targeted for optimization. One common way to choose parameters is through a sensitivity test. In this study, we choose nine key parameters used in TECO-CH₄ for an initial sensitivity test (Table 26.1). Four of these parameters are revealed to be particularly important for the modelled variability in CH₄ emission, i.e., the emissions are sensitive to those parameters. We thus pick these parameters for data assimilation. The prior ranges of these parameters are estimated from published experimental measurements or empirical values used in CH₄ biogeochemistry models (Table 26.1).

Step 4: Cost Function.

As the data assimilation algorithm for this study, we will choose the adaptive Metropolis-Hastings Monte Carlo Markov Chain (MCMC).

The approach was introduced in Chapter 22. Figure 26.2 shows the logic flow of data assimilation written into the source code of the TECO data assimilation framework. At each step in the chain, a new set of parameter values is chosen at random, the model generates results with these parameter values, and the disagreement to the observations is quantified using the cost function:

$$J_{new} = \sum_{i=1}^n \sum_{t=1}^{30} \frac{[Z_i(t) - X_i(t)]^2}{2\sigma_i^2(t)} \quad (26.1)$$

The cost function aggregates a total model-data mismatch value (J_{new}) from all the 30 time points (since we have 30 near-monthly net CH₄ emission observation from 2011 to 2014). In this study n equals 1 as we have only one set of observation data (net CH₄ emission rate). We save modelled CH₄ emission as $X(t)$ when the corresponding observed CH₄ emission is available ($Z(t)$) at time t . The standard deviation ($\sigma(t)$) is considered as the confidence level of the observation. In this study, it reflects errors from instruments, measurement, and spatial-temporal heterogeneity. It is used in the cost function to adjust the weight of model-data mismatch for individual data points. In data assimilation, both mean and standard deviation from observations are very important and should always be carefully considered in practice.

Step 5: Optimization.

If J_{new} (Step 4) passes the acceptance criteria, the proposed parameter values are saved. Acceptance depends on the value of the cost function relative to that (J_{last}) from the previous iteration of the MCMC. An initial value for J_{last} is set at 9,000,000. This is just a large initial value, to ensure that first proposed value is accepted and the iteration begins. At each step, if $J_{new} < J_{last}$, the proposed parameter values are accepted, J_{last} value is updated with J_{new} and used in the next round. After a warm-up period, parameter values start to converge (accepted parameter values during this period are discarded from the posterior distribution). Figure 26.3 shows the trajectory of updated model-data mismatch (J_{new}) on the MCMC. As the MCMC seeks for global minimum mismatch, the acceptance

TABLE 26.1

Major parameters in CH₄ production, oxidation, diffusion, ebullition, and plant mediated transportation in TECO-CH₄. Bold signifies parameters used for initial sensitivity test. Parameters with a range indicate the model is sensitive to their values and are used for data assimilation

Process	Parameters	Values	Range	Unit	Description	References
CH ₄ production	r_{me}	0.65	(0.0,0.7)	—	Potential ratio of anaerobically mineralized C released as CH ₄	Zhuang et al. (2004), Segers (1998), Zhu et al. (2014)
	Q _{10-pro}	7.2	(0.0,10)	—	Q ₁₀ for CH ₄ production	Walter and Heimann (2000)
	T _{opt-pro}	20.0	—	°C	Optimum temperature for CH ₄ production	Wilson et al. (2016)
CH ₄ oxidation	K _{CH₄}	5.0	—	μmol L ⁻¹	Michaelis-Menten coefficients	Walter and Heimann (2000), Zhang et al. (2002),
	O _{max}	15.0	(3.0,45.0)	μmol L ⁻¹ h ⁻¹	Maximum oxidation rate	Zhuang et al. (2004)
	Q _{10-oxi}	2.0	—	—	Q ₁₀ for CH ₄ oxidation	Walter and Heimann (2000), Meng et al. (2012)
	T _{opt-oxi}	10.0	—	°C	Optimum temperature for CH ₄ production	Zhuang et al. (2004)
CH ₄ diffusion	f _{tort}	0.66	—	—	Tortuosity coefficient	Walter and Heimann (2000)
	D _{air}	0.2	—	cm ² s ⁻¹	Molecular diffusion coefficient of CH ₄ in air	Walter and Heimann (2000)
	D _{water}	0.00002	—	cm ² s ⁻¹	Molecular diffusion coefficient of CH ₄ in water	Walter and Heimann (2000)
CH ₄ ebullition	(CH₄)_{thre}	750	—	μmol L ⁻¹	CH ₄ concentration threshold above which ebullition occurs	Walter and Heimann (2000), Zhu et al. (2014)
Plant-mediated transportation	T _{veg}	0.7	(0.01,15.0)	—	factor of transport ability at plant community level	Walter (1998), Zhuang et al. (2004)

Reproduced from Ma et al. (2017).

- Read in Observation data, mean ($Z(t)$) and standard deviation (σ)
- do $i=1,50000$
 - Generating new parameter sets
 - Run model to get simulated CH_4 emission, $X(t)$
 - Cost function (compare $X(t)$ to $Z(t)$)
 - Accept/reject the new parameter sets
 - Save all the accepted parameter sets
- enddo
- Randomly save 500 sets of model output generated by the accepted parameter sets

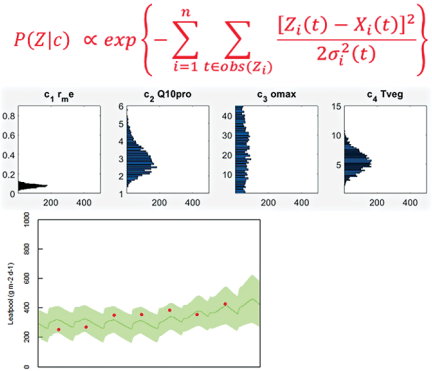


Figure 26.2. Logic flow of the TECO- CH_4 data assimilation framework.

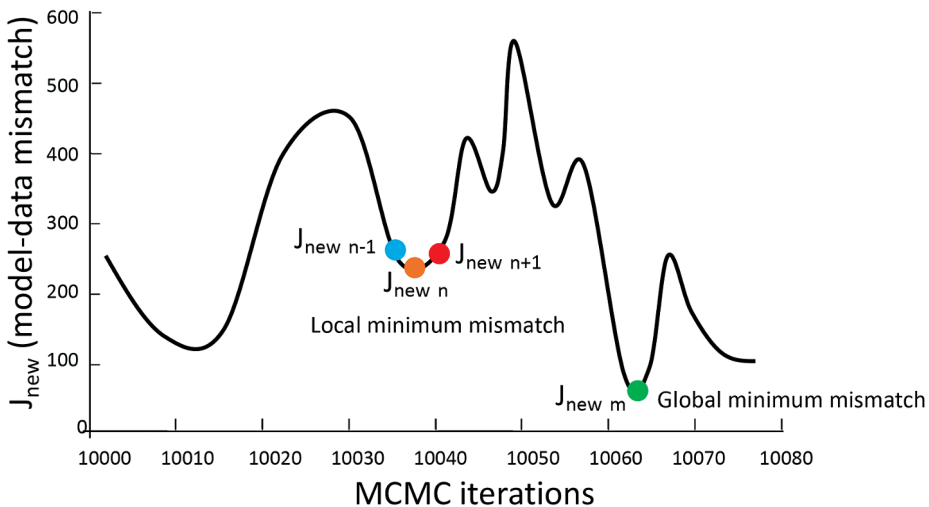


Figure 26.3. A diagram showing trajectory of updated model-data mismatch (J_{new}) on MCMC chain. Blue is J_{new} at $(n - 1)$ th iteration, orange dot is one of the local minimum mismatch point ($J_{\text{new } n}$), red is the next accepted J_{new} at $(n + 1)$ th iteration, green is global minimum mismatch ($J_{\text{new } m}$).

criteria also allow J_{new} to be accepted with a very small probability when $J_{\text{new}} > J_{\text{last}}$, so that the chain gets a chance to leave the local minimum mismatch point and reach the global minimum mismatch. See Chapter 22 for a further discussion of acceptance criteria in the MCMC.

Step 6: Estimating Parameters.

The posterior parameter distributions achieved by MCMC reveal that both of the CH_4

production related parameters (the potential ratio of anaerobically mineralized C released as CH_4 , and temperature sensitivity of CH_4 production) are well-constrained (Figure 26.4). By applying a linearized Q_{10} function to measured CH_4 emission fluxes, Gill et al. (2017) estimated the mean value of CH_4 flux Q_{10} to be 5.63 (2.92–10.52 with 95% confidence interval) at the same study site during the 2015 growing season. Our constrained Q_{10} range

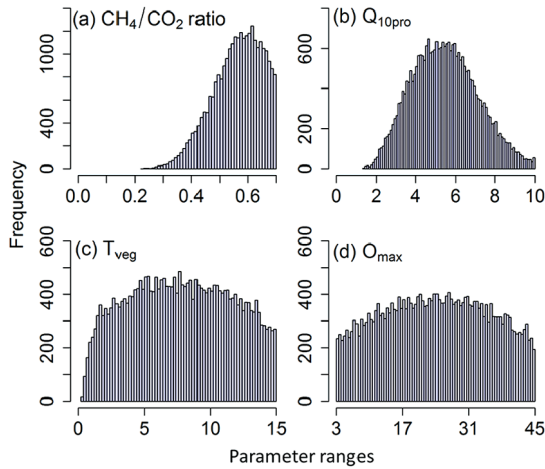


Figure 26.4. Posterior distributions of parameters of 50,000 samples from M-H simulation. (a) Potential ratio of anaerobically mineralized carbon released as CH_4 ; (b) Q_{10} for CH_4 production; (c) maximum oxidation rate; (d) factor of transport ability at plant community level. Reproduced from Ma et al. (2017).

is 2.34–6.33 with 95% confidence interval, which overlaps with but has a narrower range than this estimate by Gill et al. (2017). The other two parameters – maximum oxidation rate and factor of plant transport ability at community level – are not well constrained by the data. A longer/denser record of observation data and extra datasets such as peat CH_4 concentration are likely to be helpful for constraining these parameters.

Step 7: Generating Methane Predictions

To quantify model uncertainty due to parameter values, we can randomly draw sets of parameters from the posterior distribution and run the model with each, resulting in a distribution of outputs for CH_4 flux. Here, we will perform 500 simulations with different, randomly drawn, parameter sets, with forcing consisting of stochastically generated environmental variables (2017–2024) based on historical meteorology data. Results including a baseline historical simulation (2011–2016) are shown in Figure 26.5. It is apparent that the data constrained TECO- CH_4 simulations match both the magnitude and seasonal variations of CH_4 emission reasonably well for both the 2011–2014 and the 2015–2016 period.

The historical part of the simulation reveals the prediction accuracy of the model, and its sensitivity to environmental forcing. The model tracks the interannual variability of the measurements,

notably including the spike emission in 2016. By comparing the seasonal variation of CH_4 emission to environmental drivers, we find that wetland CH_4 emission is dominated by surface inundation and soil temperature. Soil temperature is the restricting factor below 10°C , but the water table level controls peak growing season CH_4 emission when soil temperature is well in the active range for methane processes. CH_4 emission is more sensitive to soil temperature during wet periods when the whole soil is inundated.

Data-constrained parameter probability distributions are then used to predict CH_4 response to warming. An increase of $+2.25^\circ\text{C}$, $+4.5^\circ\text{C}$, $+6.75^\circ\text{C}$ and $+9^\circ\text{C}$ in both air and soil temperature is added to drive the TECO- CH_4 model (a set of warming scenarios corresponding roughly to the experimental warming treatments at the SPRUCE site). We find exponential increase of CH_4 emission in response to warming, with four times increase under $+9^\circ\text{C}$ warming (Figure 26.6, panel a). The uncertainty in plant-mediated transport and ebullition increases most under warming and contributes to the overall change in CH_4 emission uncertainty (panels d–f).

In summary, this chapter shows how data assimilation is applied to reduce the uncertainty of modeled methane emission in a northern wetland ecosystem. The data assimilation approach used in this case study enabled parameter estimation and uncertainty quantification for forecasting methane fluxes in response to climate change.

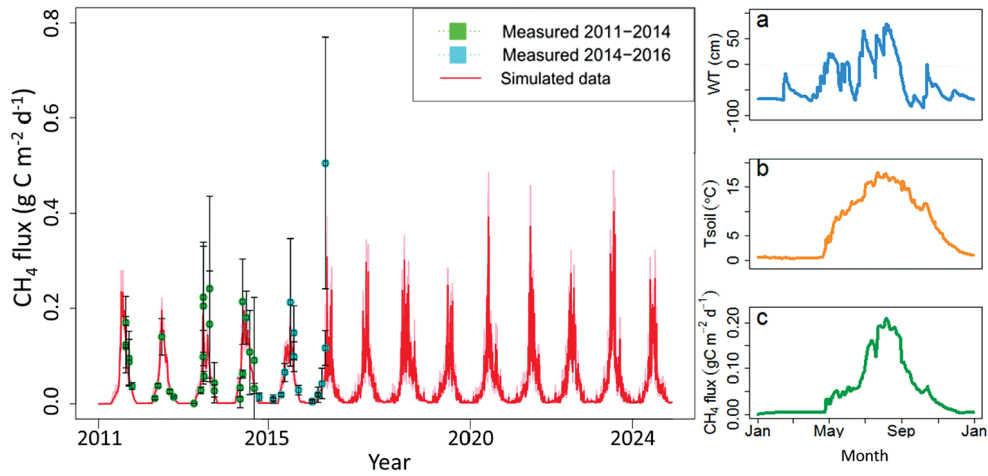


Figure 26.5. Simulation of CH_4 emission dynamics based on actual (2011–2016) and stochastically generated (2017–2024) weather forcing data. Green dots refer to observations from 2011 to 2014 which are used for data assimilation. Blue dots indicate observations from 2015 to 2016 which are used for model validation, and error bars indicate the standard deviation of each observation. Red line is simulated mean methane emission. The shading area corresponds to 1 standard deviation based on 500 model simulations with parameters randomly drawn from the posterior distribution. (a–c) The 2011 daily variation of water table, surface soil temperature, and methane emission. Reproduced from Ma et al. (2017).

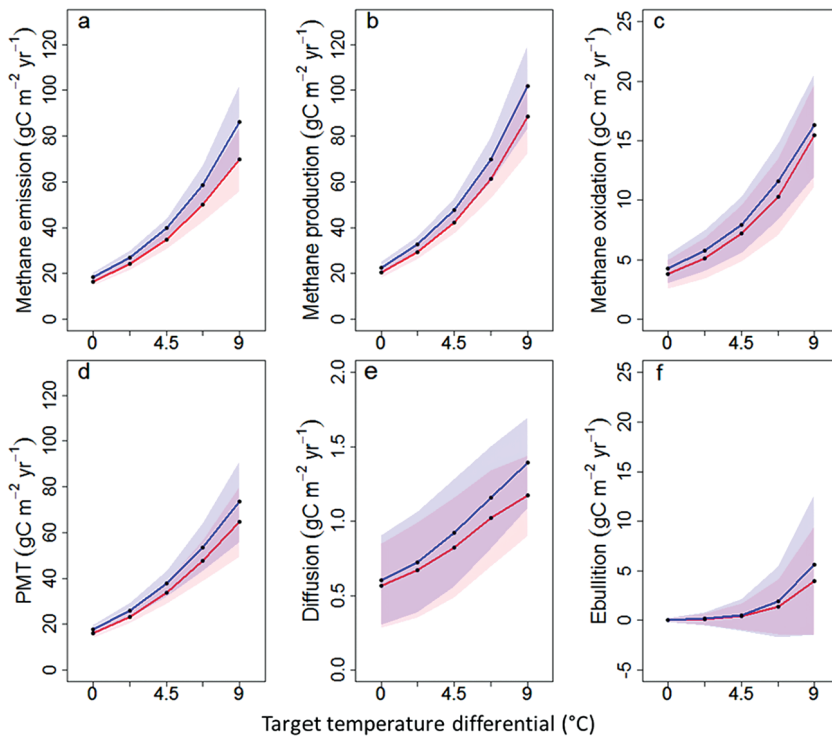


Figure 26.6. Responses of annual CH_4 emission to warming and elevated CO_2 (eCO_2). Red lines indicate CH_4 fluxes under warming treatments and 380 ppm CO_2 , blue lines indicate CH_4 fluxes under warming treatments and 880 ppm CO_2 . X-axes indicate the warming treatments of +0°C, +2.25°C, +4.5°C, +6.75°C and +9°C above ambient level. Shading area correspond to mean \pm one standard deviation based on 500 randomly chosen model simulations with parameters drawn from the posterior distribution. Reproduced from Ma et al., 2017.

SUGGESTED READING

Ma, S., Jiang, J., Huang, Y., Shi, Z., Wilson, R. M., Ricciuto, D., ... Luo, Y. (2017). Data-constrained projections of methane fluxes in a northern Minnesota peatland in response to elevated CO₂ and warming. *Journal of Geophysical Research: Biogeosciences*, 122, 2841–2861.

QUIZZES

1. Give two examples of observation data streams that could be used to constrain a methane model.

2. What are the main sources of uncertainty in wetland methane emission in terrestrial ecosystem models?

- A. Model structure
- B. Parameter values
- C. Wetland extent/inundation map
- D. All the above

3. Could you still perform data assimilation if your observation data had gaps?

4. Is the CH₄ emission data able to constrain all the parameters in this study?



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER TWENTY-SEVEN

Global Carbon Cycle Data Assimilation Using Earth Observation The CARDAMOM Approach

Mathew Williams

University of Edinburgh, Edinburgh, UK

CONTENTS

Introduction / 225
Challenges for Modeling / 225
Model Complexity / 226
Model Error / 227
Data-Model Integration / 227
CARDAMOM and DALEC – An Example Framework for C Cycle Diagnostics / 228
The Data Assimilation Linked Ecosystem Carbon (DALEC) Model / 229
The Carbon Data Model Framework (CARDAMOM) / 230
Innovations in the CARDAMOM Approach / 232
An Example of CARDAMOM / 232
Key Challenges and Opportunities for Data Assimilation / 234
Suggested Reading / 235
Quizzes / 235

The goal of this chapter is to explore the potential for data to support diagnostics and forecasting of the terrestrial carbon cycle via model-data fusion. This understanding will be built by explaining and exploring an existing framework, CARDAMOM, which is linked to an intermediate complexity model, DALEC. The key learnings will include the concept of ecological and dynamical constraints, the potential to generate emergent maps of key parameters, the role of observational error, and the key avenues for future research using earth observations.

INTRODUCTION

Challenges for Modeling

We begin with the premise that all models are wrong, but some are useful (Chapter 2). Models are

wrong because none fully describes the simulated system, being instead a simplified and incomplete representation. Models are constructed based on a series of hypotheses about the target system, and these hypotheses and their connections determine the model's structure. The structure represents the interacting components of a system (its state variables), and describes the processes that determine system evolution (changes to state variables). We recognize that the underlying hypotheses may be incorrect, over-simplified and incomplete, to varying degrees.

A core requirement for modeling is data, for calibration, validation, and forcing. Data may quantify exogenous forcing factors such as weather, which affect the rates of processes, such as photosynthesis, or may input stochastic adjustment such as disturbance, which can disrupt state variables. Data also support calibration of process parameters and

the setting of initial conditions for state variables. For instance, measurements of leaf photosynthesis under varied conditions can be used to calibrate the rate parameters for electron transport and carboxylation. The data required for drivers and for calibration will be incomplete as not every process is measurable. This incompleteness leads to poorly determined parameters and missing forcing data. Available data will have errors, systematic or random. It is the interaction of hypothesis/structural error (e.g., missing processes) and data error/gaps that causes models to be wrong.

Despite these challenges, models are useful for testing theoretical understanding and providing practical support for management and decision making. System models, the focus of this book, are particularly useful for understanding feedbacks between processes and state variables. These feedbacks occur over a variety of scales of time and space. For example, stomatal conductance responds to atmospheric conditions, that change on the scale of minutes to hours, and also responds to soil moisture, which changes on the scale of days to weeks. Soil moisture responds to external, stochastic factors such as precipitation, but also to the activity and penetration of plant roots, which might vary over months and years, and water demand from the total leaf area and its stomatal opening. Only process based models can allow these complex hypothesized linkages to be made and explored. Models provide a means to diagnose and understand what controls changes in the system, identifying key timescales, feedbacks, and interactions. Models are capable of interpolation in space and time, and therefore of making forecasts for the components of the system that are represented.

Model Complexity

Forest carbon (C) models are structured on a variety of different hypotheses and levels of complexity. For instance, some simulate the dynamics of individual trees, others of cohorts of different ages, and some of pools of C in live and dead organic matter. These different representations of the forest system trade off realism versus simplicity. Modeling individual trees is more realistic, including the potential to simulate competition among them and to model adjustments to stand microclimate, that are known from observation and experiment to feed back to growth processes and therefore system

dynamics. But this realism requires more hypotheses, for instance on competition, and therefore results in more model complexity. Complexity generates more potential connections to observations, with increased requirements for drivers and for calibration data. If these demands for data cannot be met then additional complexity may result in at best a poor characterisation of model error and at worst a heavily biased model. Complex models tend to have more parameters, and this extended demand for parameterization becomes increasingly challenging to meet. A key challenge in model construction is to determine the appropriate and justifiable complexity.

Process rates in models are determined by influences (either *internal* from associated state variables or *external* from drivers such as weather), functional forms, and parameters. Internal influences generate interactive control, whereby the magnitude of a state variable affects processes that influence that or other state variables. Thus, the magnitude of leaf area influences the magnitude of photosynthesis and evapotranspiration in many models. In more complex models the leaf area of individual stems may affect the light conditions and therefore photosynthesis of other stems, influencing competition. External influences in carbon cycle models can include weather, management, or disturbance. For photosynthesis, downwelling solar radiation is a key control, provided as a driving variable for the model at the appropriate temporal resolution and specified units. Parameters determine how the magnitude of state variables and driving variables are connected to the magnitude of the process modeled. For instance, photosynthesis ($\text{gC m}^{-2} \text{d}^{-1}$) might be estimated in a simple linear model by multiplying the light energy absorbed by a canopy ($\text{MJ m}^{-2} \text{d}^{-1}$) by a calibrated light use efficiency parameter (LUE, gC MJ^{-1}). Functional forms determine how inputs and parameters are connected algebraically, for example, determining whether the response of the process is linear, saturating or exponential.

Models strive to be realistic, general and accurate. But there are complex trade-offs required in seeking these goals, particularly the need for models to be tractable, and so simple. There is ongoing debate about whether realistic (i.e. including all known processes) or general (i.e. globally applicable) models are necessarily more accurate. It is more practical for globally-applied models to be simpler, as a simple structure means leaner

requirements for parameters and drivers. Mapping parameter variation globally is less demanding for simple models because they have fewer parameters, and these are usually easier and/or possible to obtain from the literature. For example, it is simpler to parameterize a LUE model for photosynthesis (as above), with its single parameter for calibration for each biome or land cover type, than to parameterize typical photosynthesis models in land surface schemes that might have >10 parameters for the same process. The potential advantage of the more complex model is that it is realistic, combining all the dominant controls on a process, rather than just a limited selection as in the simple model. Realism introduces more detailed processes and extra state variables, and makes more connections between them. Additional complexity adds parameters and often requires more complex functional forms. However, data sparsity can mean that parameters and functional forms are not well determined. In this case the complex model may be less accurate, particularly in making forecasts. With large numbers of parameters there is a risk of over-fitting, i.e. the complex model can be calibrated to match noise rather than signal in the data, given its many adjustable parameters. A complex model fitted to available data may make poorer forecasts than a simple model if the calibration data are biased or its errors are poorly quantified.

Model Error

While models are imperfect, they can be useful if their error can be quantified and understood. These errors can be determined through calibration and validation against observational data with robust error statistics. Validation allows testing for over-fitting. Observational error allows the modeler to weight the importance of data for calibration, and avoids over-fitting. Observational error can be incorporated into the model forecasts by propagating the error through the calibration process into parameter uncertainty. Once parameter uncertainty is quantified and included in model analyses and forecasts, models can become useful tools for generating understanding, constraining prediction, and supporting management and control. Alternate model structures, based on varied hypotheses, can be compared to understand the error associated with the model structure. Model structural error can be compared to error from the parameterization process.

Data provide objective and independent measures of the system of interest and the basis for evaluating and developing the hypotheses that create models. While models can provide useful theoretical tools for developing ecological thinking, for practical purposes related to diagnosing and forecasting global change effects it is vital that models are calibrated and validated using observations of key state variables and their controlling factors. The value of observational data is enhanced by clear description of their confidence intervals, which allows modelers to weight their importance robustly. Observational data, particularly time series, have exceptional value for evaluating understanding of dynamical systems. But the complexity and expense of collecting these data means that data replication is difficult and so uncertainty quantification is a challenge.

Data-Model Integration

Combining models with data has a long history in ecological science. Detailed studies of the photosynthetic process provided insights into the functional forms of the reactions and the critical parameters that led to robust models. Photosynthetic measurements at canopy scale were used to parameterize simple response functions by minimizing the sum of square differences between observation and model. Soil respiration models were derived empirically from large samples of respiration data under varying soil conditions. These data helped produce response models for key processes, but further work was required to produce system models with feedbacks. Combining ecophysiological process models (e.g., photosynthesis or respiration) with simulation of state variables that provide inputs to and respond to these processes has proved more challenging. Photosynthesis is a function of CO₂ concentration within the leaf, leaf area, leaf enzyme content, temperature, and irradiance. Leaf area is an ecological variable that is itself determined by allocation from C fixed by photosynthesis and phenological factors such as temperature, photoperiod, interaction with labile or nonstructural C stores and so on. Photosynthesis varies on time scales of seconds to hours while leaf area varies on time scales of weeks, so the interaction between ecophysiology and phenology is complex. To forecast future photosynthesis requires coupling a model of photosynthesis and a model of phenology to determine

interactions and codevelopment. Times series data are required for both these processes to support coupled model calibration.

Eddy covariance (EC) data have revolutionized the modeling of ecosystem C cycling by providing long time series of net CO₂ fluxes, the outcome of photosynthesis, phenology and respiration. As time series have extended, EC data have allowed evaluation of model simulated diurnal and seasonal cycles in ecophysiology, and in some cases of succession and disturbance effects on C cycling. However, characterizing the uncertainty on eddy covariance measurements of net CO₂ exchange has proved very challenging and is an ongoing area of study. Converting high frequency measurements of wind velocity and CO₂ concentration into net CO₂ fluxes relies itself on a model that makes assumptions about atmospheric processes on a range of temporal and spatial scales that are highly dynamic. It is rare to have colocated eddy covariance systems to assess instrumental error. Understanding errors arising from the eddy covariance assumptions is still developing, particularly biases that may arise due to averaging time scales and complexity of terrain. To maximize the information content of these hard-won data for model improvement requires a renewed focus on error characterization, particularly in tropical systems where EC data are rare.

We have seen that for creating robust and useful ecosystem models the fusion of model and data must provide information on multiple processes operating over a range of timescales, from physiological to phenological. Data are particularly sparse for processes operating at longer time scales, such as the evolution of the large pools of C in wood and soil. Information for understanding these relatively slow changing state variables is limited. The signal for change in these pools is difficult to extract from eddy covariance data, which are dominated by the large gross fluxes of photosynthesis and respiration, rather than the slower rates of C changes to soil and wood pools. But it is the change in the large pools that ultimately determines whether ecosystems are C sources or sinks. So, there are open questions on how processes that govern these large pools of C can be calibrated effectively. Forest inventory data provide useful information. But, like EC, high quality inventory data, particularly for soils, are rare and sparse, concentrated in a few locations globally. Soil and forest inventories rarely coincide. For global application, models need to be calibrated and

validated with relevant data that describe the complex heterogeneity of our planet and its dynamics over annual to decadal time scales.

Earth observation (EO) data are now providing vast and expanding data sources for the terrestrial carbon cycle that offer exciting new opportunities for model calibration and validation. Leaf area index, a key auxiliary state variable related to photosynthesis, has had global and continuous products since ~2000 from optical satellites operated by NASA (MODIS) and ESA (European Space Agency Copernicus data sets), with new ESA Sentinel products promising even further refinement. However, while LAI products are now delivered at finer resolutions in space and time, the products have poorly quantified biases. Optical satellites also provide products related to burned area and to deforestation. Meanwhile radar and lidar satellites are providing intermittent regional maps of above-ground biomass. The errors on these products are also poorly developed and a subject of research. It is critical that EO products are calibrated and validated against detailed field data at relevant scales. But the strengths of these EO data are compelling, providing insights into ecological variables across the globe and their changes over time. If EO data can be used to calibrate and validate models there is the potential to test the generality of models at a new level of detail and across biomes and landscapes. Thus, a major research effort has been conducted to build approaches for model calibration and validation that use EO data. The rest of this chapter describes one of these efforts.

CARDAMOM AND DALEC – AN EXAMPLE FRAMEWORK FOR C CYCLE DIAGNOSTICS

CARbon DAta MOdel fraMework (CARDAMOM) is an approach constructed to quantify the interaction between C model structure and data, with a focus on earth observations and local to global application. CARDAMOM is designed to produce a probabilistic model calibration based on local observations, sensitive to their number, type, and error. This section aims to demonstrate the potential of model-data fusion to generate understanding of the terrestrial carbon cycle through evaluating the hypotheses contained in models and to diagnose C cycling and its uncertainty. In this example, CARDAMOM calibrates the DALEC ecosystem C model. The text first describes the DALEC model, then the CARDAMOM framework,

and finally provides an example application, and some analysis of its results.

The Data Assimilation Linked Ecosystem Carbon (DALEC) Model

DALEC is a pool-based, mass balance model of the terrestrial C cycle, of intermediate complexity. The model runs on daily to monthly time steps, and can be applied across a broad range of spatial scales (from 1 ha to ~100 × 100 km). The version of DALEC described here has 17 parameters and six pools (Figure 27.1). DALEC’s six state variables include live and dead pools of organic C. Live pools represent the vegetation as foliage, wood, and fine root C stocks, and a labile pool of C that flushes leaves at the start of the growing season. Dead pools represent litter and soil organic matter C stocks. These pools evolve according to inputs to and losses of C from each, which are determined by a range of processes. Inputs of C to the system are determined by photosynthesis, which is a function of weather variables (drivers), leaf area index and leaf traits. Leaf area index is determined from the foliar C pool and the leaf trait of leaf C mass per area (LCMA). The foliar N parameter

describes the potential of a unit leaf area to fix C, and is another important leaf trait. DALEC uses the aggregated canopy model (ACM) to simulate photosynthesis. ACM is a response surface model which relates climate, soil, and atmospheric drivers to internal variables to predict gross primary production (GPP). ACM is a simplified version of the process-based Soil-Plant-Atmosphere (SPA) model. The advantage of ACM over SPA is that it operates at daily time steps, and so is orders of magnitude faster than sub-daily SPA, while still capturing the sensitivity of photosynthesis to key ecological and physical variables.

Once GPP is determined, DALEC apportions a fraction of this to autotrophic respiration (R_a), leaving the remainder for net primary production, NPP. A simplified model of R_a , requiring a single parameter, is selected due to the lack of a robust process model – this is a key area for future research. NPP is then allocated to all plant live pools. A phenological model determines how allocation to the labile pool and foliage varies over annual cycles, and how labile C is used to flush leaves, requiring parameters to determine the start of leaf flush and of leaf senescence, and their duration. The remaining C is then allocated at fixed

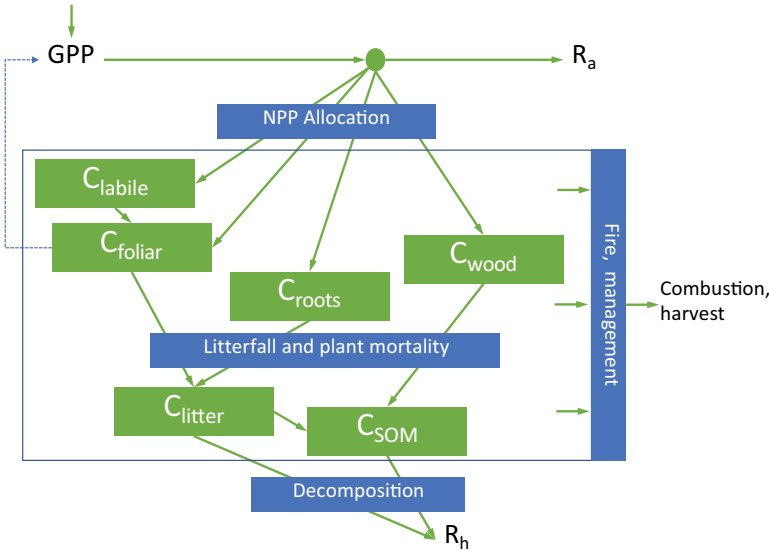


Figure 27.1. The DALEC model structure. Green boxes are pools of C in live and dead pools. SOM is soil organic matter. The labile C pool represents stored C which supports leaf flush at the start of the growing season. Green lines show C fluxes, identified by the text in the blue boxes. GPP is gross primary production. NPP is net primary production. R is respiration, either heterotrophic (h) or autotrophic (a). Fire and management can remove C from any of the pools by combustion or harvest, with prescribed loss fractions. Climate influences rates of GPP, R_h , and in some versions of DALEC influences the fluxes into and out of the labile pool. The mass of foliar C determines canopy leaf area index, which is a determinant of GPP – this influence is shown by the dashed blue arrow and generates feedback within the system.

fractions to fine roots and wood. Wood and fine root C pools have a parameter that determines their lifespans, the inverse of turnover rate. The generation of plant litter is a first order process determined by the mass of C in each foliage, root and woody pool and its turnover rate.

Plant litter from foliage and fine roots enters the litter pools, and litter from wood enters the soil organic matter (SOM) pool. Both these pools have turnover rate parameters that determine the mineralisation of these pools to CO₂. This turnover is also sensitive to air temperature, according to an exponential function with a calibrated parameter. Litter C also has a decomposition rate parameter that converts its C to SOM. Heterotrophic respiration is the sum of mineralisation from both litter and SOM pools.

Disturbances such as fire can be imposed on the C stores in DALEC, so that if the drivers indicate such an event, a fraction of C in live and dead pools is removed from the ecosystem. Foliage, wood, and litter tend to have higher fractional losses from fire than roots and SOM. Harvest effects can also be imposed, again with pool-specific parameters that determine what fraction is removed and what is added to litter and SOM. In this application these disturbance parameters are derived from the literature and are not calibrated.

The Carbon Data Model Framework (CARDAMOM)

DALEC here has 23 unknowns. As each parameter and the initial condition of each pool must be known for the model to be run, CARDAMOM is used to calibrate these model parameters using available information at the location simulated (Figure 27.2). Calibrating a 23-dimensional hypervolume is highly challenging computationally. One approach to calibration involves differentiating the model code to determine the sensitivity of its outputs (predictions) to its parameters. This process allows the model to be inverted and parameters found that produce outputs consistent with independent measurements of those outputs. But in CARDAMOM we use a forward modeling approach. This avoids the challenging process of code differentiation. Instead, Markov Chain Monte Carlo (MCMC) methods are deployed.

MCMC was introduced in Chapter 22. The MCMC approach involves running a model many times, with varying parameters. The approach finds those parameter sets that generate model outputs (e.g., LAI time series) that match independent observations (e.g., from EO), weighted by their errors, and keep these sets. By running a model millions of times, the MCMC approach searches parameter space to find parameter sets consistent with observations and observational

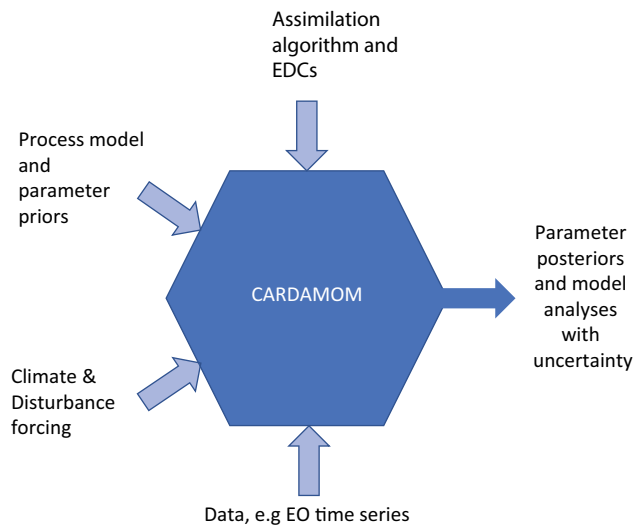


Figure 27.2. A schematic of the Carbon Data Model Framework, CARDAMOM, showing the inputs to the framework and the outputs. EDCs are ecological and dynamic constraints. The process model in this example is DALEC. The framework can be applied at a single site or over multiple pixels using earth observations (EO) as inputs.

error. The approach to select or reject parameter sets is Bayesian. That is, the approach assesses the likelihood that model and data set are consistent. This likelihood is stored with each parameter set, and the search process ends when likelihoods have converged. By running separate chains of MCMC runs, starting from different initial parameter guesses, the approach can check that different chains arrive at consistent likelihoods. If chains do not converge, the approach is judged to have failed. Failure to converge can occur if there are problems with the data, their error specification, and due to model structural errors.

Note that the output of this Bayesian approach is a set of accepted parameter sets. Parameter values are correlated because of the model structure, which connects parameterized processes via pool interactions. We can use a sample of accepted parameter sets to generate statistics about each parameter's posterior distribution. For instance, it is typical to report the 90% confidence interval on each parameter posterior. It is also informative to inspect the covariance of parameters produced by MCMC approaches to understand process interactions within the model. An interesting result from the CARDAMOM approach when used at global scales was that the estimates of leaf lifespan and leaf mass per area were correlated. This result is consistent with expectations from the leaf economic spectrum but was emergent from the structure of DALEC and the climate and time series of LAI assimilated.

Prior parameter ranges are needed for Bayesian calibration, because the calibration process must be informed of the search range of each parameter. Setting the parameter priors for MCMC calibration approaches is a challenging task. If the prior is made too wide, then the search effort is enormous and the MCMC approach may not find a calibration that is accepted, as likelihoods are too low and do not converge. If the prior is made too narrow, calibration may exclude values that are actually realistic. This can lead to edge-hitting, where the posterior parameters are found to cluster at the upper or lower bound of the prior range. It is important to inspect the posterior distribution and compare to the prior to spot edge-hitting. In this case the prior range may be expanded and the calibration repeated. Another challenge in setting the prior is to inform the MCMC of the prior distribution across the range. The simplest approach is to set a uniform prior with equal likelihood of selecting any value in the range during the MCMC

approach. But in some cases, it may make sense to set a gaussian (i.e., normal, or bell-shaped) prior, which increases the likelihood of selecting some values within this range. A gaussian prior may be chosen if one has independent knowledge about the value of the parameter.

Setting observational errors is another important task for Bayesian approaches, because these determine the likelihoods of parameter sets. Observational data, for instance earth observation products, should be provided with an estimation of uncertainty. However, in some cases errors are not available, and where these errors are provided they may be over-confident. Thus, caution suggests that relatively large uncertainties should be applied. But if observational uncertainties are too large then they do not provide constraint. In the absence of reliable observational uncertainties, we walk a fine line balancing between the need to maximize the information content of data without overfitting to error-filled observations. It is also advisable to apply geometric errors rather than arithmetic errors to avoid zero crossing when data with values close to zero are assimilated. Thus, the uncertainty on LAI might be set at $\times/\div 1.5$, rather than ± 0.5 .

To assist the process of searching dimensions of variability and to ensure that nonsensical solutions are avoided, CARDAMOM uses the concept of ecological and dynamics constraints (EDCs). EDCs ensure that model simulations are realistic and ecologically viable. EDCs reduce the uncertainty in the model parameters by rejecting estimations that do not satisfy various conditions applied to C allocation and turnover rates as well as trajectories of C stocks. For example, EDCs ensure that turnover rates of wood are always slower than for fine roots and foliage. They ensure that root:shoot biomass ratios are within broad but realistic bounds. EDCs can be used to favor parameters that result in pools close to steady state values. Ensuring a close-to-steady-state system means that pools and fluxes are broadly consistent with hypothesized processes and climate forcing. This assumption is broadly valid for low resolution studies with coarse grid cells (e.g., at $1 \times 1^\circ$ grids). At finer resolutions (e.g., ha) increasing heterogeneity will include aggrading and recently disturbed systems with accumulating biomass, an ongoing challenge for CARDAMOM.

CARDAMOM is a codebase that integrates the DALEC model, the driving data sets, the observational data sets and the Bayesian MC algorithm (Figure 27.2). CARDAMOM can be applied at a

single site, or at multiple sites, including pixel-by-pixel across a gridded land surface, even up to global scale. However, computational demands are high, so running CARDAMOM globally requires low resolution ($1 \times 1^\circ$ pixels) at present. The outputs of CARDAMOM are accepted parameter sets for each pixel/location, consistent with local forcing and observations. From these parameter sets the full C cycle at each location can be reproduced probabilistically, by using these and the driving data to run multiple instances of DALEC. Typically, 500 randomly sampled parameter sets are stored from each of three chains in the MC process. From the resulting 1500 stored parameter sets a detailed distribution of uncertainty in processes, traits, carbon fluxes and stores, and their covariances, can be determined at pixel scale.

Innovations in the CARDAMOM Approach

The outputs of CARDAMOM have clear differences from typical C cycle model simulations. CARDAMOM produces large ensembles of simulations allowing uncertainty to be characterized, whereas typical models produce only one simulation, and so lack uncertainty estimates. CARDAMOM avoids a strict imposition of steady state conditions which typical models employ. The typical model is run under steady state drivers for hundreds of years or more, until all C pools reach a steady state. Then, from this steady state, adjustments to drivers are introduced, such as climate change, and the adjustment to stocks and fluxes are followed. In CARDAMOM there is no spin-up. For each location or pixel, EDCs ensure that the C cycle is in quasi-steady state if there is no clear information from assimilated data on whether the system is a source or sink. Thus, the ensemble of accepted parameter sets will result in some that are sinks, and some that are sources, spanning a balanced net ecosystem exchange (NEE) of CO_2 . Thus, the ensemble will register uncertainty about source/sink characteristics if this information is not clear from assimilated data.

Typical ecosystem C models use the concept of plant functional types (PFT) to assign parameters. Thus, there are parameter sets for evergreen broad-leaf forest, for C_3 grasslands, tundra, etc. that are the same in all such locations globally. Competition among PFTs is simulated by the model in each pixel to determine the relative coverage. In CARDAMOM parameter sets are independent for each pixel, determined only from information available in that

pixel, its forcing, and EDCs. Thus, parameters differ spatially, adjusting along climate gradients and varying between continents depending on available information. Where information is limited, then parameter ensembles will be less constrained and will more closely match the prior information on parameter bounds. The information content of data in each CARDAMOM pixel can be determined by quantifying how much each parameter posterior is reduced in scale from the parameter prior. Likewise, comparing this prior-posterior metric among parameters provides insights into which processes have been constrained by available data, and which have not. For instance, it is typical for parameters related to foliage (e.g., leaf lifespan, canopy efficiency) to be more strongly constrained than those related to root turnover or litter mineralisation. This difference is because CARDAMOM usually has available time series of satellite LAI data which provide useful information on foliar processes directly. Information on roots or litter is sparse from earth observation.

An Example of CARDAMOM

An example of CARDAMOM application at sub-continental scale is revealing. Here we have completed an analysis for Brazil during 2001–2017. The input data include Copernicus LAI time series, a biomass map, and a soil C map. Drivers include meteorological data, a forest loss product, and a burned area product (Figure 27.2). CARDAMOM produces parameter estimates, whose median values can be mapped to show the ecological spatial variation across Brazil (Figure 27.3). It is clear from inspection that there are emergent patterns resulting from the pixel-by-pixel analysis for leaf traits. One can see distinct differences in leaf N, leaf C mass per area and leaf lifespan. The Amazon basin has patterns clearly different from Cerrado (bordering the Amazon to the SE), Caatinga (NE Brazil), and Atlantic Forest (E coast of Brazil), for example. Within the Amazon there is evidence of some parameter differences perhaps related to wetland areas. Boundaries between these key biomes are clearer or more blurred in some areas.

These biome patterns are also clear in parameters related to biomass stocks (Figure 27.4). Fractional allocation of NPP to leaves, wood and roots shows clear emergent trade-offs across Brazil. In the Amazon, allocation to wood is the dominant fraction, reflecting the productive and competitive nature of tropical moist forests. Similar

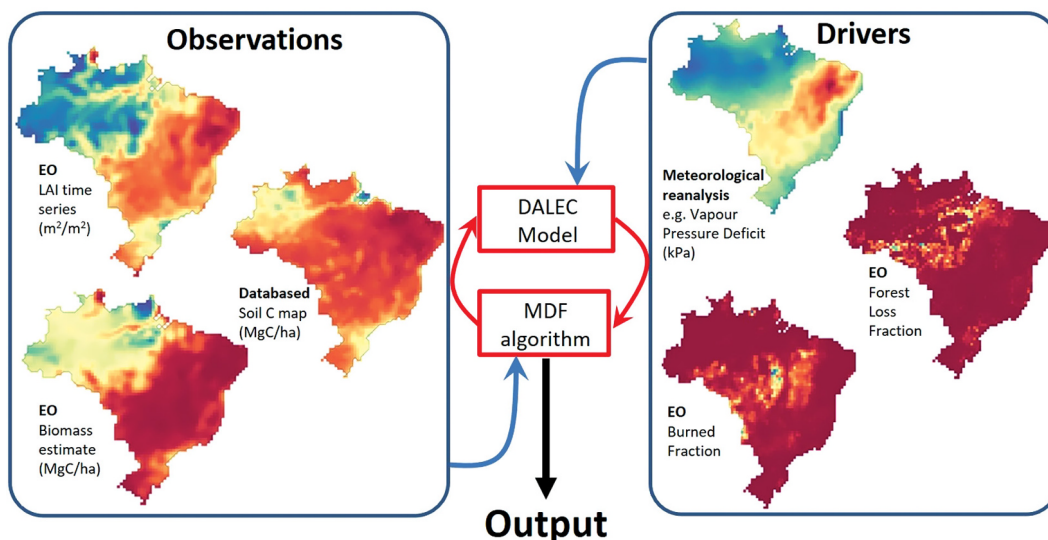


Figure 27.3. An example of the inputs used in generating a Brazilian model-data fusion (MDF) analysis of C cycling at 1° resolution monthly over the period 2001–17. The left-hand panel shows the observational data used to constrain the C cycle, including time series of leaf area index (LAI) from satellite observations, a map of soil C from interpolated field data, and a biomass map created from multiple satellite data. The right-hand panel shows observations used to drive process rates, such a vapor pressure deficit (an input to the GPP model in DALEC) and observations used to force land use change and fire impacts. Earth observations (EO) products are used to determine deforestation losses of C, and the burned fraction for combustion losses. The MDF algorithm runs very large ensembles in each grid cell using the drivers, and selects parameter sets that generate pools of C consistent with the observations and their errors. The output is a selection of these accepted parameter sets for each pixel in the analysis. Analysis provided by T.L. Smallman.

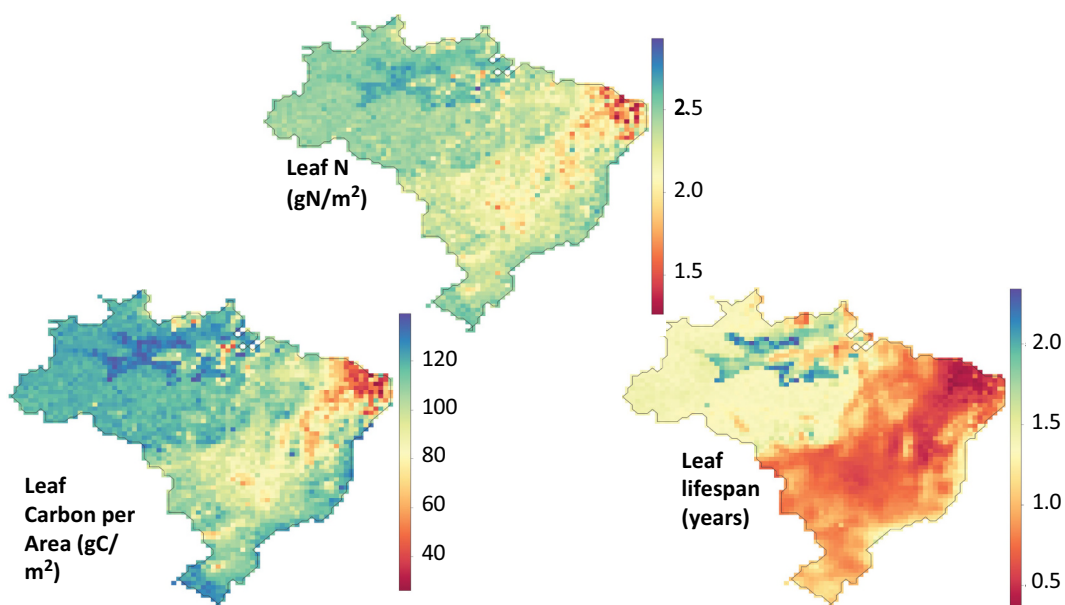


Figure 27.4. Retrieved median estimates of leaf traits for the Brazilian analysis using CARDAMOM. These traits are parameters within DALEC. Leaf N determines the maximum rate of photosynthesis. Leaf C per area determines the relationship between foliar C mass and leaf area index (LAI), and the investment requirement in C to construct leaf area. Leaf lifespan determines the turnover time of foliage, and therefore the investment of C required each year to maintain the canopy at the observed LAI. For each pixel a probability distribution for each parameter is determined. Analysis provided by T.L. Smallman.

patterns can be observed in the narrow band of Atlantic Forest along the coast. In the dry forests of the Cerrado and even drier Caatinga the dominant allocation fraction is to roots, reflecting the water limited nature of these systems, and their disturbance by fire. Within the Amazon there is evidence of regional variability, with the northeastern area having relatively higher allocation to wood and lower allocation to leaves than other areas.

A key novelty of the CARDAMOM approach is that it generates emergent patterns in C processing, between and within biomes – for instance across the equatorial forest biome. Field studies are now providing better spatially resolved, bottom-up information on leaf and plant traits to guide model calibration. The top-down results from CARDAMOM can be evaluated against these independent data to check for consistency across biomes and continents (Figure 27.5).

The typical model evaluation approach is to test using data independent from the calibration data. For instance, CARDAMOM estimates of leaf mass per area across the globe, emergent from earth observations, climate drivers and DALEC model structure, could be compared to interpolations of field observations. Areas of agreement and mismatches could be highlighted. The magnitude of mismatches could

be compared using different versions of DALEC (e.g. with alternate phenology schemes), different climate forcing, and different earth observations of LAI to explore their validity. But there is an alternate philosophy for model-data fusion approaches, which is to assimilate the field observations of leaf traits (i.e., not leave any for independent testing). We can evaluate the resulting analysis for consistency of model and data, and for its likelihood. The probabilistic approach used in CARDAMOM provides a clear quantification of model value. The information content of new data sets like a bottom-up traits map can be evaluated by comparing analyses with and without the new data. Thus, what is the impact on analytical likelihood of the new assimilate? Does this have a spatial or temporal impact?

KEY CHALLENGES AND OPPORTUNITIES FOR DATA ASSIMILATION

The earth's land surface is both dynamic and highly heterogeneous with variation in C stocks occurring at varied length scales across the globe according to biological processes and exogeneous factors linked to disturbance and management. Disturbance and management operate on a range

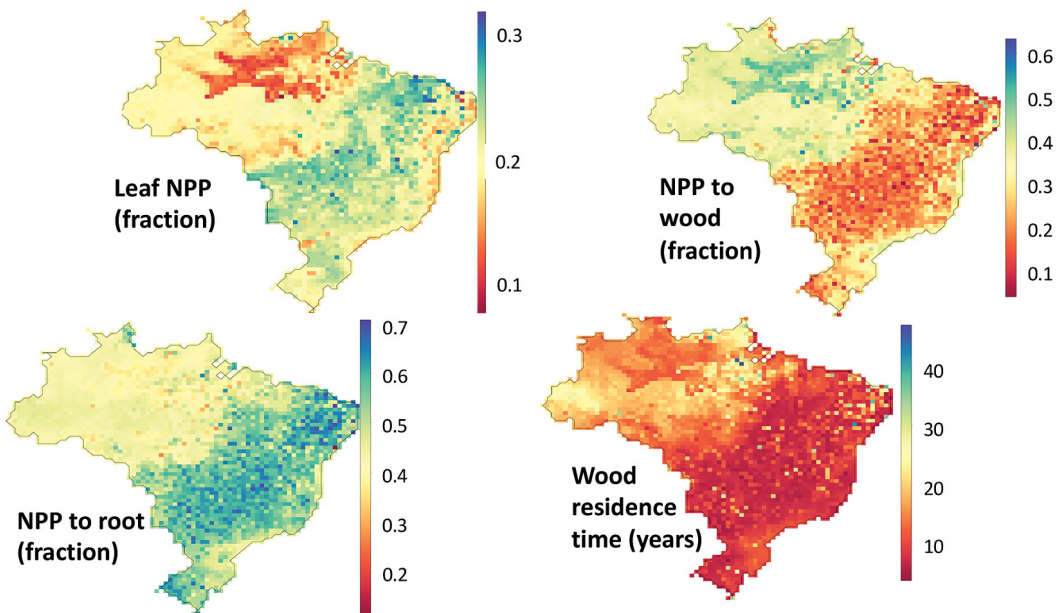


Figure 27.5. Retrieved median estimates of plant traits for the Brazilian analysis using CARDAMOM. Three of the panels show the fraction of net primary production (NPP) allocated to leaves, wood and fine roots (i.e., these fractions sum to 1). The final panel shows the wood residence time, indicating the longevity of C storage in live woody biomass. The patterns are emergent, as each pixel is analyzed independently to produce likely estimates of these parameters. Analysis provided by T.L. Smallman.

of scales, but typically are more important at finer spatial resolutions. Thus, as analyses sharpen their resolution from $1 \times 1^\circ$ to 1 ha, for instance, disturbance and management become more important and ecosystem properties become more dynamic and potentially further from steady state. At $1 \times 1^\circ$ the annual rate of forest loss is relatively small compared to stocks, for typical forested landscapes. But in these same landscapes a 1 ha forest plot can lose most of its stocks in one year, due to fire or clearance, and rates of aggradation can also be fast.

There is a need to resolve forest C processes at fine resolutions (e.g., 1 ha) to support monitoring, reporting and verification processes linked to global treaties for climate. For this reason, there are numerous satellite missions under construction or in orbit to provide data on forest structure at these resolutions. Frameworks such as CARDAMOM can assimilate these satellite observations to produce C flux estimates at similar resolutions. There are technical challenges relating to computing power requirements and data management. There are also algorithmic challenges to calibrate models for dynamic forest patches where disturbance and recovery from disturbance are common states. High temporal resolution data from satellites provide a means to monitor for rates of change and thereby to calibrate internal processing of C. Thus, repeated biomass stock estimates have been shown to provide insights into carbon allocation parameters, and even into key soil parameters, as these are downstream of wood dynamics. Further, the spatial distribution of biomass stocks also provides information on the local landscape dynamics. A steady state landscape will have a normal distribution of biomass stocks. A degrading landscape will be dominated by lower biomass with a long tail of higher biomass from remnant forest patches. Using this information will require new algorithms not currently used in CARDAMOM, to link information from neighboring pixels in a landscape analysis.

The atmospheric C community has a long history of using inversion approaches to identify source and sink regions from linking atmospheric transport models and atmospheric CO₂ concentration data in carbon cycle data assimilation system (CCDAS). There is a clear opportunity to link the ecological data assimilation in frameworks such as CARDAMOM to atmospheric approaches. In fact, links to atmospheric data would be valuable particularly because identifying whether a landscape is a source or sink is challenging. Future developments

include the potential to link to atmospheric inversions for independent tests of net exchanges.

SUGGESTED READING

For more details on the CARDAMOM method read:

- Bloom, A.A. & M. Williams (2015). Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological “common sense” in a model-data-fusion framework. *Biogeosciences* 12: 1299–1315.
- Bloom, A.B., J.-F. Exbrayat, I. R. van der Velde, L. Feng, M. Williams (2016) The decadal state of the terrestrial carbon cycle: global retrievals of terrestrial carbon allocation, pools and residence times. *Proceedings of the National Academy of Sciences* 113: 1285–1290.
- Smallman, T.L., J. -F. Exbrayat, M. Mencuccini, A. A. Bloom and M. Williams (2017) Assimilation of repeated woody biomass observations constrains decadal ecosystem carbon cycle uncertainty in aggrading forests. *Journal of Geophysical Research Biogeosciences* 122: 528–545.

Compare the CARDAMOM approach with other model-data fusion approaches in the papers presented below. What are the advantages of these other approaches?

- Peylin, P., Bacour, C., MacBean, N., Leonard, S., Rayner, P., Kuppel, S., Koffi, E., Kane, A., Maignan, F., Chevallier, F., Ciais, P., and Prunet, P. (2016) A new stepwise carbon cycle data assimilation system using multiple data streams to constrain the simulated land surface carbon cycle. *Geophysical Model Development* 9: 3321–3346.
- Kaminski, T.; Scholze, M. L. U.; Vossbeck, M.; Knorr, W. L. U.; Buchwitz, M. and Reuter, M. (2017) Constraining a terrestrial biosphere model with remotely sensed atmospheric carbon dioxide. *Remote Sensing of Environment* 203: 109–124

QUIZZES

1. Models aim to be general, realistic, and accurate. Why is it so hard to meet all three goals at once?
2. What are the arguments for and against calibrating and validating global C cycle models at eddy covariance sites alone?
3. How does the CARDAMOM approach to generating C model parameters differ from the typical plant functional type approach? What are the relative advantages of each method?
4. What are the challenges to applying CARDAMOM at very high resolutions (e.g., 1 ha) across the globe? How might these challenges be addressed?



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER TWENTY-EIGHT

Practice 7

DATA ASSIMILATION AT THE SPRUCE SITE

Shuang Ma

Northern Arizona University, Flagstaff, USA

CONTENT

Practice Design / 237

This practice uses the SPRUCE field experiment as a case study to explore how a model's parameter values influence its output via sensitivity analysis, and how parameters may be estimated via assimilation of data from field measurements. Exercise 1 addresses how certain model output variables are sensitive to changes in certain parameter values. Exercise 2 illustrates how observational data change posterior parameter distributions in comparison to the priors. In this practice, we use data from the SPRUCE research site in Northern

Minnesota, USA, as a context to show possible studies you could conduct by using data assimilation. The exercises may be performed using the CarboTrain software. Instructions for installing and working with CarboTrain are available in Appendix 3.

PRACTICE DESIGN

Data sets that are used in this practice are listed in Table 28.1.

TABLE 28.1
The SPRUCE site data used in this practice

Purpose	Data name	Year	Period	Time step
Environmental variables (input) to drive the TECO model, spin up and forward run	Soil temperature at 0, 5, 10, 20, 30, 40, 50, 100, 200cm depth	2011–2018	Whole year	Hourly
	Air temperature at 2m			
	Relative Humidity at 2m			
	Wind speed at 10m			
	Precipitation			
	Photosynthetically Active Radiation (PAR) at 2m			
Water balance calibration	Soil moisture at 0, 20cm	2014–2018	Whole year	Hourly
	Water table depth	2014–2018	Whole year	Hourly
Data streams used in data-model fusion	Leaf, wood, root biomass	2014–2018	End of growing season	Once a year
	Soil C content	2012	August 13–15	One time
	NEE, GPP, ER fluxes	2015–2018	Growing season	1–2 times a month

NOTE: NEE = Net Ecosystem Exchange, GPP = Gross Primary Production, Reco = Ecosystem Respiration.

EXERCISE 1: Parameter sensitivity of model output

In this exercise we will examine how the parameter values of a model affect its outputs. Launch CarboTrain on your computer (see Appendix 3) and follow these steps:

- a) Select Unit 7 → Exercise 1 → Set output directory (e.g. `mydir/EX1/default`) → Click *Run Exercise*.
- b) Output files appear in subdirectory simulation of the output directory you specified in the previous step. CarboTrain plotted `daily_cpool.png` and `daily_fluxes.png` with the data in the same folder for a quick look. The full model outputs are also provided as `.csv` files. The model output can be found in the file: `'your_output_dir/simulation/Simu_dailyflux14001.csv'`. The observation data can be found here: `'/CarboTrain/Source_code/TECO_2.3/input/SPRUCE_cflux.txt'` and `'/CarboTrain/Source_code/TECO_2.3/input/SPRUCE_cpool.txt'`. Annotations for output data columns is here: `your_CarboTrain_dir/CarboTrain/Source_code/TECO_2.3/annotations_in_TECO_modeloutput_file_format.jpg`
- c) Select Unit 7 → Exercise 1 → Set output directory (e.g., `mydir/EX1/increase_vcmax`). Click *Set Initial parameters* and increase the parameter 'Vcmax' value by 30%. Click *Run Exercise*.
- d) Repeat step (c), set a new output directory (e.g., `mydir/EX1/decrease_vcmax`). Click *Set Initial parameters* and decrease the parameter 'Vcmax' value by 30%.
- e) Repeat steps (c) and (d) for increases and decreases of the following parameters. Remember to create a new name for output directory each time after you

changed the parameter value. You should end up with ten different output files:

- i. SLA, the specific leaf area;
 - ii. Tau_leaf, leaf carbon turnover time;
 - iii. Tau_root, root carbon turnover time;
 - iv. Tau_slowSOM, recalcitrant soil C pool turnover time.
- f) Plot the output data from your ten model simulations and compare the differences. It will be helpful for ease of comparison to plot the results for one parameter with $\pm 30\%$ changes (e.g., Vcmax) on the same figure.

After Step (b), you can inspect the default run results, such as daily leaf, wood, and root pool changes, GPP, NEE, Reco, and other carbon fluxes from 2011 to 2016. When you decrease and increase the Vcmax value by 30% in Steps (c)–(d), you will see substantial changes in GPP and autotrophic respiration. Similarly, when the value of specific leaf area is changed, we can see effects on GPP and autotrophic respiration. However, in both cases, we don't see much change in the carbon pools. From these comparisons, you may realize that Vcmax and SLA have similar effects on model outputs.

By contrast, when you change the turnover time of the leaf pool, you may see that the steady state size of the leaf pool has considerably changed whereas GPP or NEE do not change much. At steady state, the leaf pool is about 350 gC m⁻² when we decrease the turnover time by 30%, increasing to about 450 gC m⁻² if we increase the turnover time by 30%. Similarly, tuning the turnover time of root C changes the steady state size of the root carbon pool. Thus, we may learn that turnover time can affect the steady state of carbon biomass.

Exercise 1 shows us that certain output variables are sensitive to changes of certain parameter values. We have also seen that adjustments in different combinations of parameter values can

sometimes generate the same changes in output variables. In other words, we might sometimes get right answers but for a different reason. An example is that the value of leaf C turnover time, SLA, and Vcmax can all be responsible for the magnitude and pattern of leaf carbon pool sizes and GPP. This is so-called equifinality, an issue for both manual tuning and Bayesian-guided parameter estimation (Luo et al. 2009). However, manually tuning parameter values

can be tricky and subjective. Data assimilation searches for the whole prior ranges of parameter values with millions of iterations to generate posterior probability distributions.

QUESTION:

Which output variables are most sensitive to change in each of the five parameters? Can you explain why?

EXERCISE 2: Data assimilation with TECO

This exercise will help you learn to perform data assimilation with the comprehensive ecosystem model, TECO. We will use CarboTrain as a toolbox for the exercise.

- 1) Launch CarboTrain and select Unit 7 → Exercise 2 → Set output directory (e.g., `mydir/EX2/default`) → Click Run Exercise. The model runs iteratively 20,000 times with different parameter values, the accepted parameter values are saved in the file `DA/paraest.txt`. It takes several hours to finish the full data assimilation process. If you want to get fast but not accurate results, you may let it run for 500 times, reducing the run time to 5–10 mins. To do this, monitor the value for 'isimu' in your terminal window; when it gets to 500, terminate the run by pressing Ctrl+C. Then CarboTrain will randomly select 100 sets of accepted parameters saved in the file `DA/paraest.txt`, and save model forward run output in the file `forecasting/Simu_daily_flux***.txt`. '***' could be any of the 3-digit numbers ranging between 1–100. A quick plot of the results is also provided in the file `pool_and_flux.png`.

- 2) This exercise is a teaser for you to walk through the whole process of data assimilation. In a formal data assimilation study, tens of thousands to millions of iterations are normally needed to get to a stabilized chain. The number of iterations needed depends on the model structures, the parameter sampling and selection algorithms (e.g., different variants of the Monte Carlo Markov Chain algorithm; see Chapter 22), and the quality and quantity of observations.

Figure 28.1 shows parameter posterior distributions from a full data assimilation run informed by observational data from the SPRUCE field experiment. We can see that SLA and leaf C turnover time are well-constrained by the observational data. In Exercise 2, we see that parameters to which the model is sensitive are more likely to be constrained by data. Instead of one value, the data assimilation study generates probabilistic distributions of parameter values, allowing the range of uncertainty to be displayed and analyzed. With data assimilation, you can quantify uncertainties and partition the uncertainties into different sources of error and process variabilities.

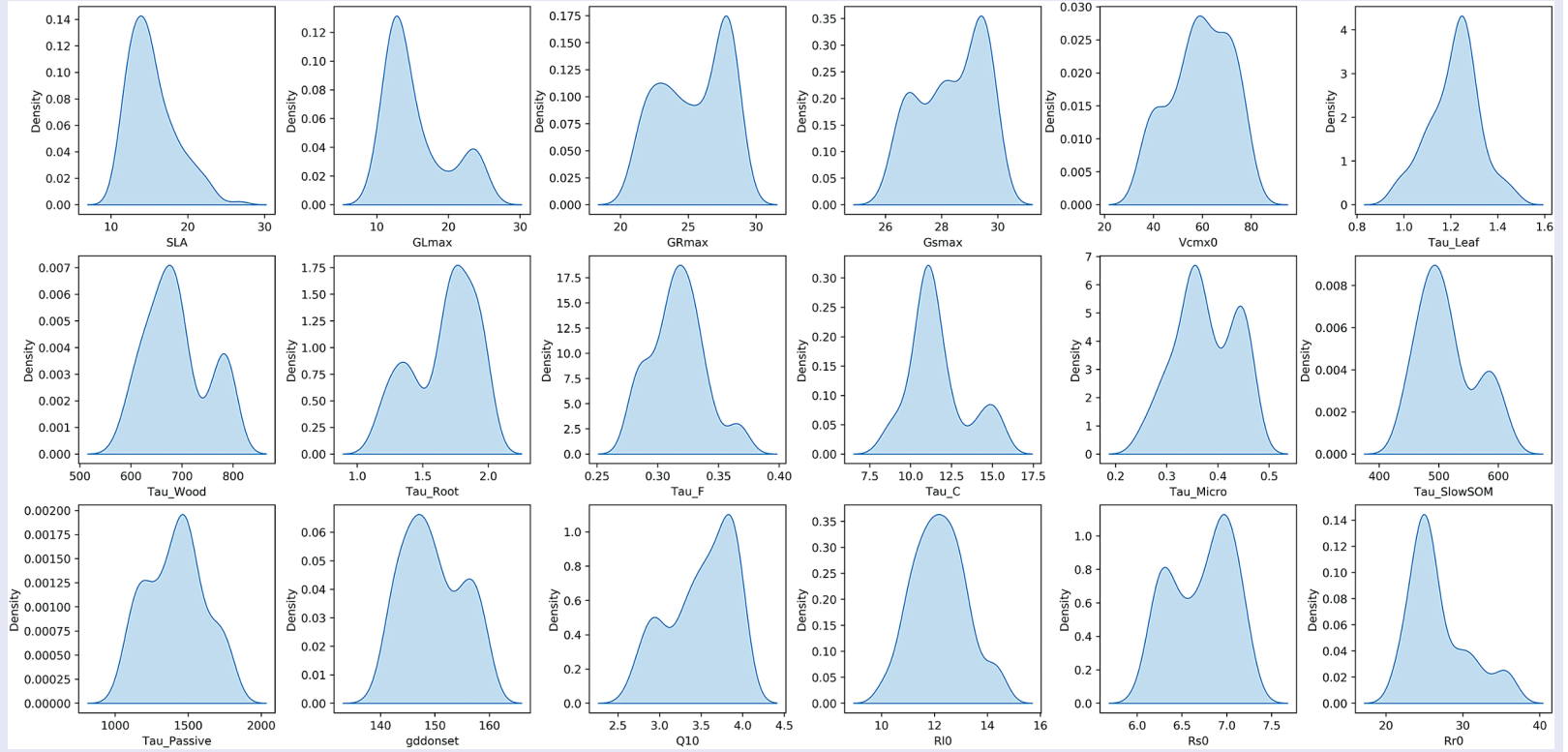


Figure 28.1. Parameter posterior distribution from data assimilation based on SPRUCE experimental data using the TECO model.

QUESTIONS:

1. What is a constrained posterior distribution?
2. Which parameters are most likely to be constrained by the observation data sets, why? (hint: think of equifinality and parameter sensitivity)



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

UNIT EIGHT

Value of Data to Constrain Models and Their Predictions



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER TWENTY-NINE

Information Contents of Different Types of Data Sets to Constrain Parameters and Predictions

Enqing Hou

Northern Arizona University, Flagstaff, USA

CONTENTS

Introduction / 245

An Overview of the Information Contents of Model and Data / 246

A Method to Quantify the Information Contents of Model and Data / 246

Short- and Long-term Information Contents of Model and Data / 248

The Information Contents of Data Depend on the Amount and Type of Data / 248

Model Equifinality / 251

Prediction of Land Carbon Dynamics After Data Assimilation / 252

Summary / 253

Suggested Reading / 253

Quizzes / 253

There is an urgent need to reduce model uncertainty in predicting land carbon dynamics. The model uncertainty can potentially be reduced by assimilating the increasing amount of observational and experimental data into the model. This chapter first gives an overview of the information contents of models and data and then introduces a method to quantify the information contents of data and models. Next, the chapter introduces the use of data assimilation to extract information from the data and models, and shows how the information contained within a model and in data can vary depending on the timescale of focus: short- versus long-term. The extracted information is useful to constrain model parameters and predictions, as well as to guide future data collection.

INTRODUCTION

Model uncertainty in predicting future land carbon balance is huge. Reducing this uncertainty is a major challenge in land carbon cycle studies, yet is urgently needed in order to support the development of

strategies to mitigate climate change. Meanwhile, the availability of relevant observational and experimental data is increasing, driven by research advances and societal needs for data to support the management of natural resources in the context of global change (Luo et al. 2011). To meet these societal needs, we must address critical questions such as how best to use data to constrain models and how to select appropriate datasets to optimize information collection. These questions can potentially be addressed with the aid of data assimilation or data-model fusion techniques, which use data to constrain initial conditions and parameters of models, thereby constraining their predictions of the land carbon cycle (Luo et al. 2011).

In this chapter, we first give an overview of the information contents of models and data. Here, the information of a model means the structure and parameters of a model, which reflect knowledge of processes of the system the model strives to represent. We then introduce a method to quantify information content, both for models and for data. Next, we discuss how the information

contained within a model and in data, pertinent to the prediction of land carbon dynamics, can vary depending on the timescale of focus: short versus long-term. After that, we relate the information contents of land carbon datasets to model parameters, and define the issue of equifinality of model parameters. In the final part of the chapter, we address the value of data for constraining model predictions on land carbon dynamics.

AN OVERVIEW OF THE INFORMATION CONTENTS OF MODEL AND DATA

A well-known saying in the modeling community is “all models are wrong, but some are useful” (Box, 1979). That means we cannot construct a “perfect” model, but we can make reasonable and useful predictions. If a model can predict land carbon dynamics reasonably well during hindcast, based on an evaluation of the model output relative to available observations, we assume it may inform future land carbon dynamics with a high confidence. Whether a model can predict the carbon dynamics well during hindcast depends on both model structure and parameter values, which constitute the prior knowledge of a system (Weng and Luo 2011).

The structure of a model (the equations and their linkages to forcing, state variables, and parameters) contains information relevant to its behavior and performance as a predictive tool. A model expresses the logical consequences of a set of hypotheses, formalized through the model structure, and generates predictions that can be compared with observations in a quest to falsify these hypotheses (Canham 2003). In other words, the model is an integration of theory and knowledge, which can be used to test hypotheses and make predictions (see also Chapter 2). For example, ecosystem respiration is usually modeled as an exponential or a unimodal function of temperature in land carbon models, which are built upon empirical relationships between ecosystem respiration and temperature reported in the literature. These alternative formulations predict ecosystem respiration similarly at low temperatures but differently at high temperatures; either way, it is information on the prediction of ecosystem respiration (no matter whether correct or not). The two response functions represent competing hypotheses that can be tested and falsified by comparing their predictions to measurements.

Model parameter values also contain information relevant to the predictions made. Once we

have a model, whether the model can predict well or not will depend on parameter values. Model parameters can be tuned rigorously and automatically with data assimilation. Prior values or ranges (or uncertainty) of model parameters may be set based on empirical studies of processes or expert judgment informed by empirical measurements, which themselves are a kind of information on prediction. Assimilating relevant datasets, such as land carbon pool and/or flux measurements, into a model can further constrain initial conditions and parameter values of a land carbon model in a rigorous way, resulting in a data-informed process-oriented modeling approach. This allows researchers to do more realistic predictions on land carbon dynamics.

A good example to show how the information contents of model and data constrain predictions of land carbon dynamics is the study of Weng and Luo (2011). They compared model predictions on carbon dynamics in a temperate forest using the Terrestrial ECOSystem (TECO) model with and without data assimilation of eight datasets on carbon pools or fluxes. They found that probability density functions (PDFs) of the carbon pool sizes in initial predictions by the model were bell-shaped for slow-cycling carbon pools including woody biomass, structural litter, slow-cycling soil organic matter (SOM), and passive SOM, but skewed to their low carbon content ends for fast-cycling carbon pools including foliage, fine roots, metabolic litter and fast-cycling SOM (Figure 29.1). Departure of the PDFs from a uniform distribution (null knowledge) suggests that model structure, together with the prior ranges of parameters, contains information on land carbon pools, particularly on slow-cycling carbon pools. After assimilating the eight datasets into the TECO model, they found that simulated sizes of all eight carbon pools except metabolic litter pool were bell-shaped and well constrained (Figure 29.1). These results suggest that data can provide a substantial amount of additional information on land carbon dynamics.

A METHOD TO QUANTIFY THE INFORMATION CONTENTS OF MODEL AND DATA

Weng and Luo (2011) introduced the use of Shannon information index (Shannon 1948) to measure the relative information of land carbon models and datasets. In information theory, entropy (H) is used to measure the “information”,

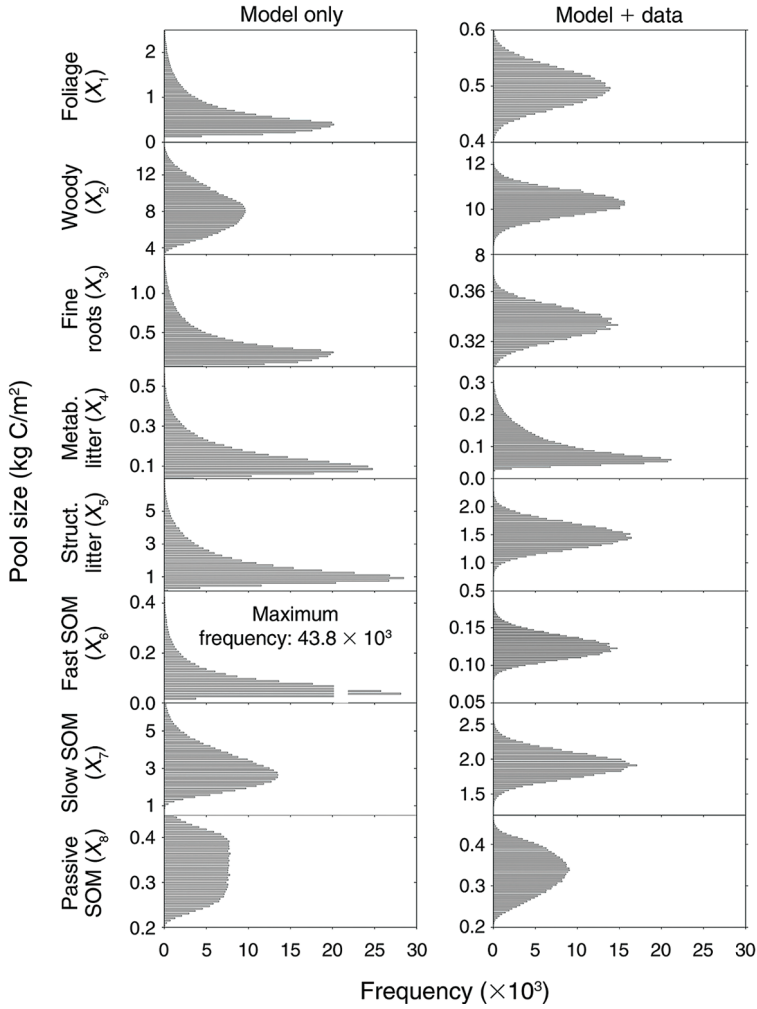


Figure 29.1. Simulated carbon content with parameters sampled in prior distributions (model only) and posterior distributions (model + data), respectively. Struct. indicates structural, metab. indicates metabolic, and SOM indicates soil organic matter. Derived from Weng and Luo (2011).

“surprise”, or “uncertainty” of a random variable (X); a higher entropy indicates more information, surprise, or uncertainty. In carbon modeling, entropy can be used to measure the uncertainty of an output variable (e.g., predicted leaf carbon pool) of a model; a higher entropy indicates a larger uncertainty of the variable. Entropy is maximum when there is null knowledge about an output variable, i.e., a uniform distribution of values. The information content of a model can be calculated as the difference in the entropy of a uniform distribution (null knowledge) and the posterior PDF of an output variable, which carries the trace of model information contained in its structure and parameters. The information content of a dataset

can be calculated as the difference in the entropy of posterior PDF of an output variable between the model alone and the model when informed by a dataset, for instance through parameter calibration to fit the model to the dataset.

Specifically, the entropy H of a discrete random variable X in $\{x_1, \dots, x_n\}$ is calculated as:

$$H(X) = -\sum_{i=1}^n p(x_i) \log_b p(x_i) \quad (29.1)$$

where $p(x_i)$ is probability of event x_i and n is the number of events. The negative sign ensures that the result is always positive or zero, because $p(x_i)$

is smaller than or equal to 1.0. If the base b equals 2, the unit is bit. Equation (29.1) can be applied where X is an output variable from a model, or a variable in a dataset, to characterize the information it contains. The set of X may comprise, for example, n points in a time series, or n cells of a spatial grid. The information (I) of a model and/or dataset(s) is expressed in a relative term, that is, relative to information without either a model or any dataset.

Null knowledge on dynamics of a carbon pool is defined by a uniform distribution of the pool size within its range. The entropy of null knowledge, H_0 , is calculated as:

$$H_0 = \log_2 n \quad (29.2)$$

Relative information of null knowledge (I_0) would be 0.

The entropy of the PDF of a carbon pool, when simulated by a model alone, $H(X_m)$, is calculated as:

$$H(X_m) = - \sum_{i=1}^n p(x_{m,i}) \log_2 p(x_{m,i}) \quad (29.3)$$

where X_m is a carbon pool simulated by the model alone, $x_{m,i}$ is the mean value of X_m in a bin, and n is the number of bins with equal widths in the range of the simulated carbon pool. Relative information of the model (I_m) is calculated as:

$$I_m = H_0 - H(X_m) \quad (29.4)$$

Similarly, the entropy of the PDF of a carbon pool simulated by a model informed by dataset(s), $H(X_{md})$, is calculated as:

$$H(X_{md}) = - \sum_{i=1}^n p(x_{md,i}) \log_2 p(x_{md,i}) \quad (29.5)$$

where X_{md} is a carbon pool simulated by the model with the dataset(s), $x_{md,i}$ is the mean value of X_{md} in a bin, and n is the number of bins with equal widths in the range of the simulated C pool. Relative information of the dataset(s), I_d , is calculated as:

$$I_d = H(X_m) - H(X_{md}) \quad (29.6)$$

The above method was developed by Weng and Luo (2011) to quantify model and data information on a carbon pool based on the PDF of simulated C pool sizes. This method has been used by

Keenan et al. (2013) to quantify model-data mismatch in a carbon variable based on the PDF of mismatches between modeled values and observed values. Moreover, the method may be used to quantify model and data information on a model parameter based on the posterior PDF of model parameter values, as introduced in Chapter 32.

SHORT- AND LONG-TERM INFORMATION CONTENTS OF MODEL AND DATA

Projecting changes in the land carbon cycle over decades to a century is crucial for developing strategies to mitigate climate change. However, most data assimilation studies have focused on land carbon cycling in the short term (i.e., within a decade). Weng and Luo (2011) explored whether and how the information contents of the TECO model and data, in terms of their relative contributions to predicted land carbon dynamics, change with projection time. They found that when predicting slow-cycling carbon pools, for example, woody and soil C pools, the importance of the model increased with time while the importance of data decreased with time (Figure 29.2). In contrast, when predicting fast-cycling carbon pools, for example, foliage and fine roots, the relative importance of the model and data did not change much with time (Figure 29.2). Moreover, eight datasets examined by the authors contain more information on the simulated carbon pools of foliage and fine roots than the model (Figure 29.2). The model, however, contained more information on the simulated litter carbon pool, fast-cycling SOM, and passive SOM than the datasets (Figure 29.2). These results suggest that both data and model are important in simulating land carbon cycling; data is particularly important in constraining predictions in the short term; the model is increasingly important in the long term.

THE INFORMATION CONTENTS OF DATA DEPEND ON THE AMOUNT AND TYPE OF DATA

Since data can inform models, we may expect an increase in information with the increase in data amount. Indeed, the information content of data generally increases with the amount of data. For a specific type of data (e.g., the measurement of net ecosystem exchange), information content may initially increase but later saturate as additional data are added. Not only the amount but

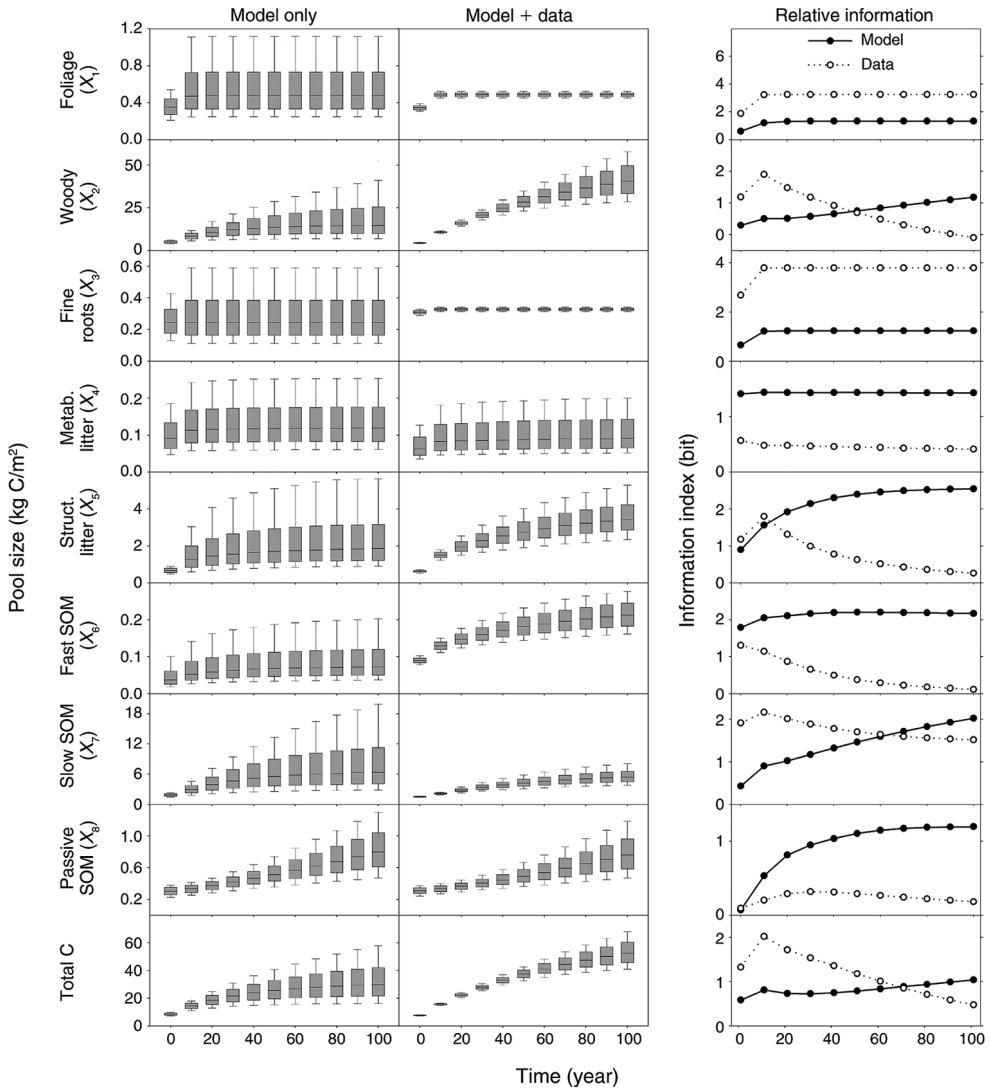


Figure 29.2. Simulated C contents and the relative information of model and data on simulated C contents over 100 years. Box plots show carbon content distributions in the 5% (bottom bar), 25% (bottom hinge of the box), 50% (line across the box), 75% (upper hinge of the box), and 95% (upper bar) intervals. Derived from Weng and Luo (2011).

also the type of data is important to constrain models. For example, Richardson et al. (2010) assimilated six datasets on carbon dynamics in a spruce-dominated forest into a simple forest carbon cycle model, DALEC. They found that the joint use of woody biomass increments, and flux tower-based measurements of net ecosystem exchange, constrained parameter values of DALEC markedly better than the use of net ecosystem exchange measurements alone. The degree to which parameter uncertainty was reduced depended on both the dataset used and its relationship to a given parameter. For example, either soil respiration or

plant biomass increment data were needed to constrain estimates of the fraction of gross primary production respired. Litterfall data were required to narrow estimates of turnover rate of foliage. Soil respiration data were necessary to constrain the turnover rate of soil organic matter.

Data streams usually overlap in the information they contain, and data collection is limited by funding and logistical factors. Therefore, it is critical to know which data streams are more useful to constrain models than other data streams. To address this, Keenan et al. (2013) constrained a simple forest carbon cycle model with 17 different

datasets from the Harvard Forest in Massachusetts, USA. They iteratively ranked each dataset according to its ability to reduce model uncertainty. They found that some types of data were more important than others in constraining the model (Figure 29.3). For example, the measurement of net ecosystem exchange was most useful to constrain the model, followed by soil carbon turnover, soil respiration, and litterfall (Figure 29.3). A combination of measurements of fast and slow carbon flows was necessary to achieve optimal model performance. Models generally become better constrained with increasing types of measurements, but some measurements are relatively redundant in the presence of some other measurements. This is the case where measurements overlap in terms of the information they encode

relating a desired output of the model to its parameters and/or structure. Keenan et al. (2013) showed that when all their 17 datasets were used, 26 of the 40 model parameters could be identified or constrained. Parameter uncertainty was reduced by 60% on average, but most of this reduction was achieved with the use of relatively few data streams (Figure 29.3). For example, when the six most important data streams were used, 14 parameters were constrained. Adding the remaining 11 data streams only constrained 12 more parameters. Fourteen parameters were not constrained even when all the available types of measurements were used. The soil carbon pool did not provide additional information to constrain the model in the presence of soil respiration and soil carbon turnover rate. Litter turnover did

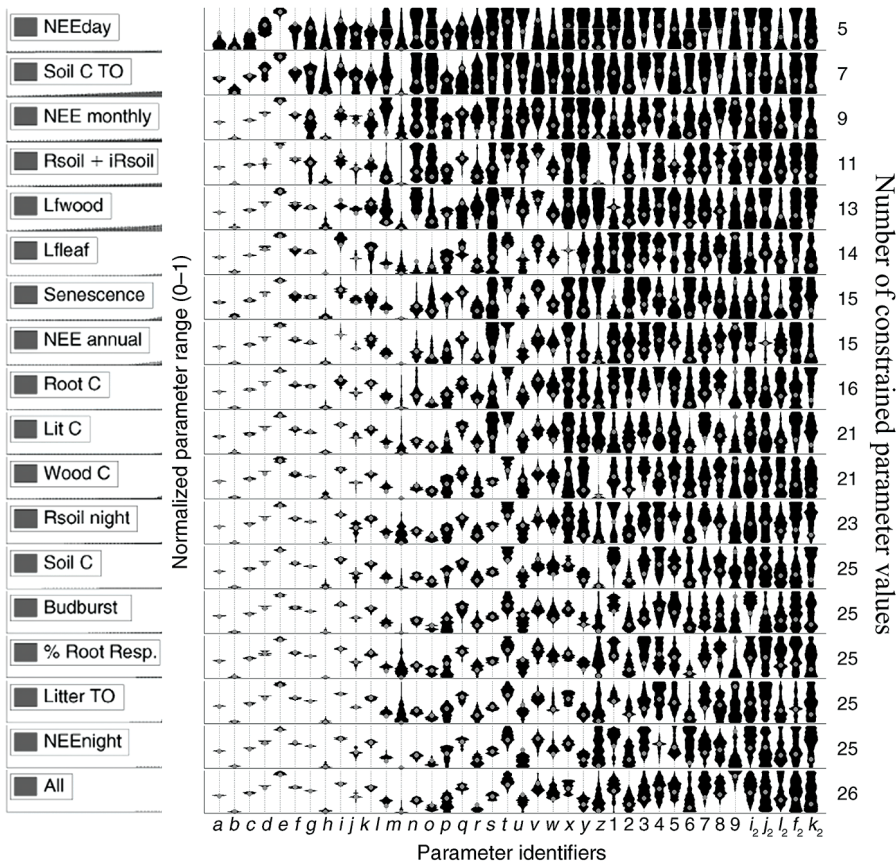


Figure 29.3. The posterior parameter distributions for the best data combination at each stage in a hierarchical optimization process. Variables on the left-hand side are the variables used to constrain the model at each stage (i.e., indicated by each row). Numbers on the right-hand side are the number of parameters well constrained at each stage. Parameters are deemed to be well constrained if their posterior distribution occupies at most half the range of the prior distribution. Solid gray circles represent the optimum parameter value; black areas represent mirrored posterior probability distributions for each parameter. Derived from Keenan et al. (2013).

not add information in the presence of litterfall. Results like these can guide future data collection and optimize the use of funding and the deployment of labor.

MODEL EQUIFINALITY

Model equifinality is defined as the situation where different sets of model parameter values can yield similar model predictions (Luo et al. 2009). Assimilating data into a model can reduce

equifinality, but it is hard to eliminate completely. Equifinality of parameter sets may be indicated by bivariate correlation between posterior parameter values. However, lack of such bivariate correlation does not necessarily indicate a well-constrained model. Instead, whether a model is well constrained or not may be better indicated by bivariate covariance between posterior parameter values (Figure 29.4). This is because covariance scales correlation by the standard deviation of parameters, of which the latter is lowered in a well-constrained model.

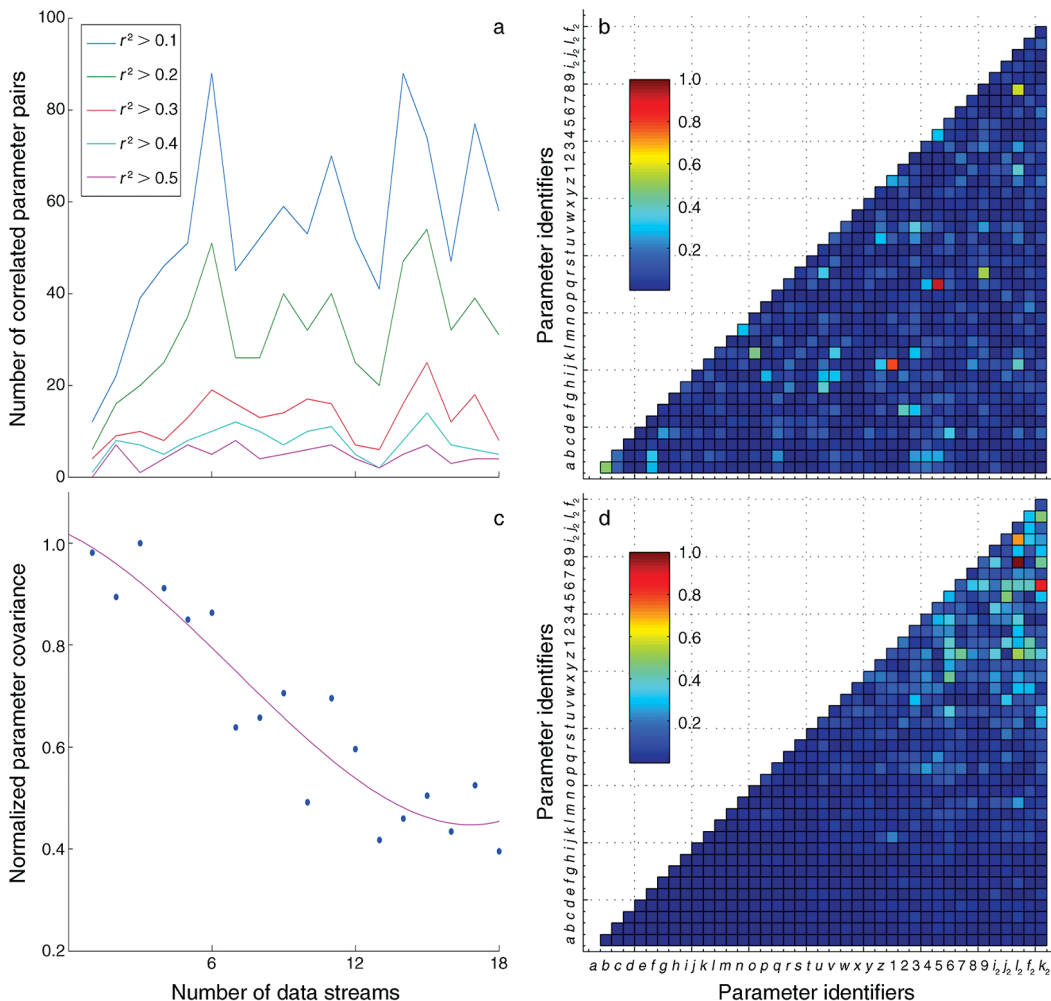


Figure 29.4. Characteristics of correlations or covariances among parameters after data assimilation. (a) The number of posterior parameter distributions that show significant ($P < 0.01$) correlations for different levels of correlation and different numbers of constraining data sets. (b) The correlation matrix of model parameters for the model constrained by all available data sets. The color scale represents the r^2 correlation between each pair of parameters. (c) The posterior parameter covariance (dots) for different numbers of constraining data sets, normalized to the maximum total covariance observed. The line represents a polynomial fit to the data. (d) The covariance matrix for the model parameters for the model constrained by all available data sets. The color scale represents the covariance normalized to the maximum observed covariance value. Derived from Keenan et al. (2013).

Correlation between posterior parameter values may occur when high values of one parameter (or a term scaled by it) can be compensated for by low values of another parameter or term in the same response function or process. For example, soil heterotrophic respiration may be calculated as a multiplication of basal respiration rate and temperature sensitivity of respiration (i.e., Q_{10}). While both base respiration rate and temperature sensitivity of respiration may be informed by measurements of soil respiration, trade-off between the two parameters allows the model to predict similar values of soil respiration with many, anticorrelated, combinations of the two parameters. In other words, the posterior distribution of the two parameters may be negatively correlated. In this case, a direct measurement of one of the parameters may constrain the other model parameter and produce a more accurate model.

Correlation between posterior parameter values may also occur when counteracting processes in a model cannot be distinguished by the data. For example, forest floor litter is usually separated into metabolic litter and structural litter. However, turnover rates of the two carbon pools may not be well informed given only litterfall and/or forest floor biomass measurements. This is because a fast turnover rate of one of the two litter pools may be compensated by a slow turnover rate of the other litter pool. In this case, measured pool size of either metabolic litter or structural litter may be helpful to constrain the turnover rates of both litter pools. If no such measurements are available, a more parsimonious model with the two litter pools binned together may be used to perform data assimilation and model predictions.

Note that the absence of a correlation between posterior parameter values does not necessarily indicate the identifiability of model parameters. A correlation may be absent because of high-dimensional (i.e., > two dimensional) relationships among posterior values of model output variables, especially when most model parameters are poorly informed by data. For example, leaf biomass is positively correlated with litter biomass, which in turn is negatively correlated with litter turnover rate, but leaf biomass may not correlate with litter turnover rate. If litter biomass but not leaf biomass or litter turnover rate is informed

by a dataset (e.g., litter biomass measurement), the relationship between leaf biomass and litter turnover rate may become positive. Indeed, Keenan et al. (2013) found that the number of correlated parameter pairs increased with the number of datasets used for data assimilation, up to six (Figure 29.4). This could be because when more datasets are used to constrain the model, the dimensionality of model parameters may decrease (Luo et al. 2009), resulting in more apparent correlations between some parameters. In the study of Keenan et al. (2013), using more datasets, in addition to these six, did not significantly change parameter correlations (Figure 29.4), as the correlations seemed to be mainly determined by the structure of the model.

PREDICTION OF LAND CARBON DYNAMICS AFTER DATA ASSIMILATION

Model predictions can be improved by data assimilation. Data assimilation can reduce uncertainties in predicted land carbon dynamics during both hindcasts and forecasts. Similar to model parameters, uncertainties in predicted land carbon dynamics generally decrease with the use of more data; meanwhile, the uncertainties can be reduced more using some datasets than using some other datasets. Assimilating data such as ten years' measurements of land carbon pools into a model can substantially reduce uncertainties in predicted land carbon pools, especially for fast-cycling carbon pools (e.g., foliage and fine roots). Dynamics of slow-cycling carbon pools (e.g., passive soil carbon pool) may not be well informed by short-term (i.e., ≤ 10 years) measurements of pool size, because a slow degrading pool would not be expected to exhibit much change over such a relatively short period of measurements. The ^{14}C signature of soil carbon pools, which can indicate the age of carbon in soil, could provide a solution for informing the dynamics of slow-cycling carbon pools. Measurements of soil ^{14}C signature and models that have incorporated such measurements are still uncommon, however.

Data assimilation can reduce not only the uncertainty but also the bias of model predictions during forward runs. The impact of data assimilation on model predictions during forward runs

may be evaluated by splitting data streams into two time periods. Data streams during one time period are used for model calibration and data assimilation, with the other time period used for data validation. The bias of model prediction may be assessed using root mean square error, which measures the difference between data and model predictions. When forecasting future land carbon dynamics, the maximum likelihood estimates of land carbon pools can be altered by data assimilation. The estimates after data assimilation are supposed to be more reliable than those before data assimilation. If forecasted in a near term (e.g., days to years), the forecasts may be validated by future measurements as they become available.

SUMMARY

Data assimilation is useful to extract information from model and data. The information contents of models and data can be quantified using the Shannon information index. A good forward model is fundamental to long-term predictions of land carbon dynamics. Data is particularly important in constraining predictions of short-term land carbon dynamics. While the amount of information generally increases with the amount of data, some types of data are more useful than others in informing models. When some model parameters cannot be constrained by available measurements, it suggests the need for new types of measurements or a simpler model. Overall, data

assimilation exercises can aid model development and guide data collection to constrain model predictions of land carbon dynamics.

SUGGESTED READING

- Keenan, T. F., E. A. Davidson, J. W. Munger and A. D. Richardson (2013). Rate my data: quantifying the value of ecological data for the development of models of the terrestrial carbon cycle. *Ecological Applications* 23(1): 273–286.
- Weng, E. and Y. Luo (2011). Relative information contributions of model vs. data to short-and long-term forecasts of forest carbon dynamics. *Ecological Applications* 21(5): 1490–1505.

QUIZZES

1. What kind of metric can be used to quantify the information content?
2. How does the contribution of a model to projected carbon pool sizes change with projection time?
3. Do more land carbon measurements mean more information on model predictions on land carbon dynamics?
4. How can correlations and covariances between posterior values of parameters be used to explore model equifinality?
5. How can data assimilation help model predictions of land carbon dynamics?



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER THIRTY

Using Data Assimilation to Identify Mechanisms Controlling Lake Carbon Dynamics

Oleksandra (Sasha) Hararuk

University of Central Florida, Orlando, USA

CONTENTS

Models and Data-Model Fusion /	255
Processes That May Control Epilimnetic C Dynamics /	256
Model Calibration and Selection /	259
Processes That Control Epilimnetic C Dynamics /	260
Suggested Reading /	262
Quizzes /	262

This chapter demonstrates how mathematical models and data-model fusion can be used to identify the processes controlling carbon (C) dynamics. A suite of process-based models was built to describe the C dynamics in a temperate lake. The models were built so as to create a factorial modeling experiment aiming to identify the processes that contributed most to explaining the observed variation in the observed data. All models were calibrated using the Bayesian Markov Chain Monte Carlo algorithm and their fit indices were used to identify the processes that contributed most to model improvement. Although the example used in this lecture focuses on C dynamics in a temperate lake, the presented hypothesis-testing algorithm is transferable to other response variables and ecosystems.

MODELS AND DATA-MODEL FUSION

Ecosystem models are useful tools for exploring how sensitive a variable of interest is to changes in environmental conditions as well as projecting its state into the future. A process-based model is a mathematical expression of our theoretical knowledge that was initially obtained through observational and manipulative experiments. Models are

typically developed in three stages. The first stage is model formulation: we express the theoretical knowledge about the dynamics of the response variables as a system of mathematical equations. The second stage is model calibration: the parameters in the mathematical equations are tuned so as to improve representation of the observations. The initial values of the parameters, that are tuned up or down in the calibration stage, are typically obtained from the literature or prior experimental data. The last stage is model validation: model performance is evaluated against a set of observations that were not used for model calibration.

We may run into issues during the last stage of model development: our model may not perform well when evaluated against the observations. Once this becomes an issue, one may be tempted to attribute the poor model performance to inaccurate representation of processes controlling the dynamics of the response variable. To remedy this, we might be tempted to add more processes into the model. However, adding new processes increases model complexity and may lead to overfitting, i.e., capturing the noise in the observations and not the signal. In addition, an increase in the number of model parameters (which is an inevitable consequence of increasing model complexity)

may inflate the uncertainty in the model predictions, thus reducing its utility.

To reduce the risk of overfitting and uncertainty inflation, it is important to implement the second stage of the model development process – model calibration – by using a data-model fusion (or data assimilation) approach. A data-model fusion algorithm, such as 3DVAR, 4DVAR (Lewis et al., 2006) Ensemble Kalman Filter, or Bayesian inversion (Evensen, 2009), finds such parameter values that either minimize the least-squares function, or maximize the likelihood function. In other words, these algorithms find such values for model parameters that result in the closest match between the model output and observations. Because these algorithms for model calibration are not restricted by the model structure, they have been adopted by atmospheric scientists, hydrologists, as well as biogeoscientists as a formal model calibration approach.

A model calibrated using a data-model fusion approach performs to the best of its ability, therefore if a calibrated model still performs poorly, the poor performance can be attributed to the fact that the chosen model structure is not optimal for the given system. Such properties of data-model fusion provide grounds for accepting or rejecting a structural model hypothesis based on its performance, thus opening an avenue for learning about the processes that best describe a system (Figure 30.1). So, if we have a number of competing hypotheses about the processes controlling the dynamics of our response variable, we can build a

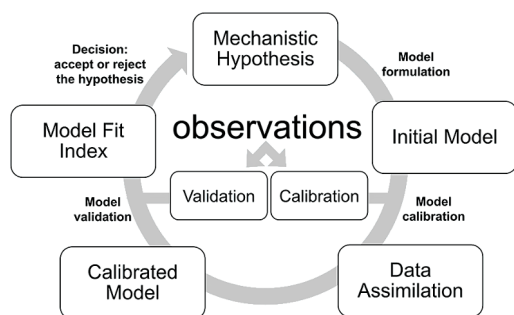


Figure 30.1. Conceptual representation of model-based hypothesis testing. Mechanistic hypotheses about the processes within an ecosystem are expressed as a process-based model; model parameters are calibrated using a data assimilation technique and a portion of observations reserved for calibration; afterwards the calibrated model is evaluated on the portion of observations reserved for validation, and the fit index is used to accept or reject the mechanistic hypothesis that was represented in the model.

suite of mathematical models that represent these processes, calibrate them against the observations, and evaluate which process representations (and the hypothesis they entail) consistently improve model performance. Such a multiple-hypothesis testing framework aligns with principles of “strong inference” (Platt, 1964), which have led to rapid progress in science, particularly in the field of molecular biology.

In this chapter, we study an example of implementing a multiple hypothesis testing approach to learn about the mechanisms controlling the dynamics of epilimnetic carbon (C) in Long Lake, a temperate lake located at the University of Notre Dame Environmental Research Center.

PROCESSES THAT MAY CONTROL EPILIMNETIC C DYNAMICS

To start, let us define the general model structure and explore how it can be modified to test various structural model hypotheses. There are two forms of C in the epilimnion, or the portion of the lake above the thermocline: dissolved organic C (DOC) and dissolved CO₂. Similar to the dynamics of the C pools in terrestrial ecosystems (see Chapter 1 “Theoretical Foundation of the Land Carbon Cycle and Matrix Approach”), dynamics of these two aquatic C pools can be expressed as a system of first order linear differential equations:

$$\frac{dX(t)}{dt} = I(t) - A(t)X(t) \quad (30.1)$$

where $X(t)$ is a vector of epilimnetic DOC and CO₂ pools at the time t ; $I(t)$ is the vector of inputs to the DOC and CO₂ pools, which includes stream, precipitation, and overland flow; and $A(t)$ is the C loss and transfer matrix. The $A(t)$ matrix combines the effects of biotic and abiotic factors on DOC mineralization; hydrologic C export, or loss of DOC and CO₂ with outflow; atmospheric exchanges; and fluxes to or from the hypolimnion, or the portion of the lake below the thermocline.

The model structural assumptions (or hypotheses) that are tested here affect the size of $X(t)$, and the elements of $I(t)$ and $A(t)$ (Figure 30.2). The evaluated model structural hypotheses are listed below. These hypotheses and subhypotheses are combined so as to create a factorial experimental design, yielding 40 model variants in total. Once the model variants are calibrated against the

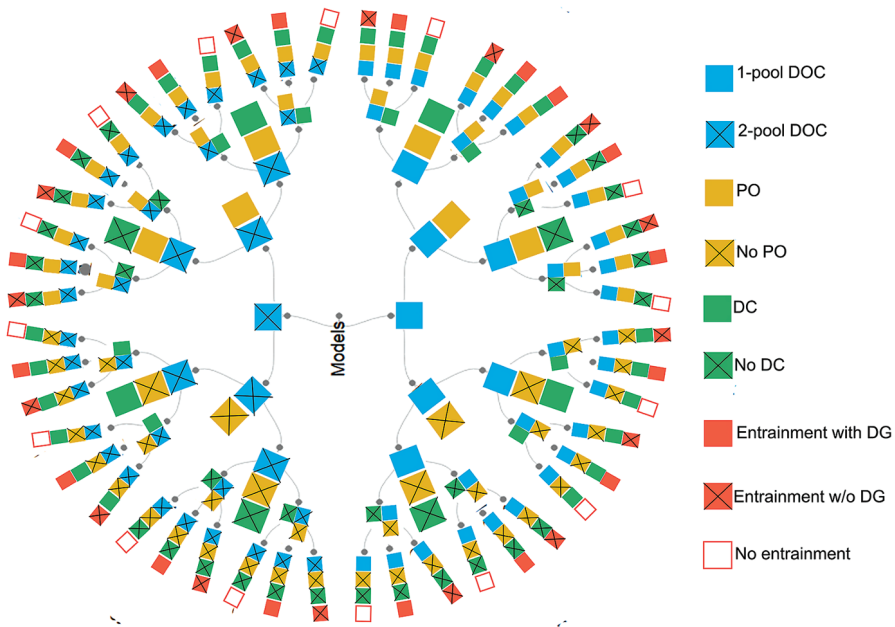


Figure 30.2. Diagram depicting a subset of the ensemble of model formulations (represented as terminal nodes). Each model formulation is calibrated against the time series of DOC and CO_2 , and the processes that improve the model fit are identified via modeling the fit indices as a function of presence or absence of a process. PO: model variant with photooxidation; DC: model variant with representation of density currents; DG: model variant with the representation of the depth gradient in hypolimnetic DOC and CO_2 .

HYPOTHESIS 1

Accounting for varying DOC lability improves model performance. Even though DOC may be treated as a labile form of organic matter in terrestrial ecosystems, incubation studies have indicated a range of lability of DOC in aquatic ecosystems. The model variant that represents heterogeneity in lability of DOC includes two DOC pools: a labile and recalcitrant one. To incorporate two DOC pools, a third row is added to the $X(t)$ vector as well as $I(t)$ vector and $A(t)$ matrix from Equation 30.1:

$$\begin{bmatrix} \frac{d\text{DOC}_e^l(t)}{dt} \\ \frac{d\text{DOC}_e^r(t)}{dt} \\ \frac{d\text{DIC}_e(t)}{dt} \end{bmatrix} = \begin{bmatrix} I_{\text{DOC}_l}(t) \\ I_{\text{DOC}_r}(t) \\ I_{\text{DIC}}(t) \end{bmatrix} - \begin{bmatrix} Q_{\text{out}}(t) + k_l & 0 & 0 \\ 0 & Q_{\text{out}}(t) + k_r & 0 \\ -k_l & -k_r & Q_{\text{out}}(t) + \text{evasion}(t) \end{bmatrix} \begin{bmatrix} \text{DOC}_e^l(t) \\ \text{DOC}_e^r(t) \\ \text{DIC}_e(t) \end{bmatrix} \quad (30.2)$$

where subscripts or superscripts l and r denote process affecting labile or recalcitrant DOC; subscript “e” denotes epilimnetic C pools; $Q_{\text{out}}(t)$ is total water discharge out of lake via groundwater and outlet at the time t; k_l is the mineralization rate of the labile DOC; k_r is the mineralization rate of the recalcitrant DOC, and $\text{evasion}(t)$ is the outgassing of DIC to the atmosphere calculated as $\text{evasion}(t) = \frac{k\text{CO}_2(t)}{z_{\text{thermo}}}$. $k\text{CO}_2(t)$ is gas transfer rate for CO_2 in m day^{-1} and z_{thermo} is the average daily thermocline depth in m.

HYPOTHESIS 2

Representing gravity-driven density currents in the lake improves model performance. The Long Lake is fed by a stream, which also delivers DOC and CO₂ into the lake. Depending on the concentration of the solutes in the stream water and its temperature, its density may be higher or lower compared to that of the epilimnion. If stream water density is lower or equal to that in the epilimnetic water, the solutes will be delivered to the epilimnion. However, if stream water density is higher compared to that in the epilimnion, the solutes will sink to the hypolimnion. In the latter case, incoming DOC and CO₂ will essentially be locked in the lake, whereas in the former case CO₂ will escape to the atmosphere (given that the water is supersaturated with CO₂), and DOC will be transformed into CO₂, which, in turn will escape to the atmosphere. To test this hypothesis elements of $I(t)$ are modified by a function $f(\Delta\rho)$, which allocates less stream C to the epilimnion as the water density gradient $\Delta\rho$ increases:

$$\begin{bmatrix} \frac{dDOC_e(t)}{dt} \\ \frac{dDIC_e(t)}{dt} \end{bmatrix} = e^{-k_{dens}\Delta\rho(t)} \begin{bmatrix} I_{DOC}(t) \\ I_{DIC}(t) \end{bmatrix} - \begin{bmatrix} Q_{out}(t) + k & 0 \\ -k & Q_{out}(t) + evasion \end{bmatrix} \begin{bmatrix} DOC_e(t) \\ DIC_e(t) \end{bmatrix} \quad (30.3)$$

where $\Delta\rho(t)$ is the difference between stream and epilimnion densities at time t ; k_{dens} is sensitivity of epilimnion DOC and DIC load to differences between stream and epilimnetic water densities; and k is the DOC decay rate.

HYPOTHESIS 3

Representing photooxidation of DOC improves model performance. Water incubation studies have shown that DOC is depleted at a higher rate in the presence of ultraviolet (UV) light compared to that in the dark incubations. Indeed, UV light may aid in oxidation of organic molecules and break covalent bonds in large organic molecules, thus increasing their susceptibility to microbial uptake and mineralization. Here, the effect of photooxidation is tested through modifying the elements of $A(t)$ by an empirical function of photosynthetically active radiation, $f(PAR)$:

$$\begin{bmatrix} \frac{dDOC_e(t)}{dt} \\ \frac{dDIC_e(t)}{dt} \end{bmatrix} = \begin{bmatrix} I_{DOC}(t) \\ I_{DIC}(t) \end{bmatrix} - \begin{bmatrix} Q_{out}(t) + k + f(PAR) & 0 \\ -k - f(PAR) & Q_{out}(t) + evasion \end{bmatrix} \begin{bmatrix} DOC_e(t) \\ DIC_e(t) \end{bmatrix} \quad (30.4)$$

where $f(PAR) = k_{base_photo} e^{\left(\frac{k_{photo} PAR(t)}{70000}\right)}$. $PAR(t)$ is photosynthetically active radiation at the time t in mmol photons $m^{-2}day^{-1}$; k_{base_photo} is base photodegradation rate, and k_{photo} is sensitivity of DOC degradation to PAR .

HYPOTHESIS 4

Representing entrainment of hypolimnetic water and a depth gradient in hypolimnetic CO₂ and DOC concentrations improves model performance. When the thermocline deepens, some hypolimnetic water moves into the epilimnion. The hypolimnetic water

is an additional source of CO₂ and DOC that is not accounted for by the default elements of the $I(t)$ vector in Equation 30.1. In addition, entrainment can be a significant source of nutrients to the epilimnion and enhance nutrient-limited productivity and DOC decay rates. The distribution of DOC and CO₂ in the hypolimnion may not be uniform across depths, which can be treated as two additional sub-hypotheses of Hypothesis 4. For instance, CO₂ concentrations tend to be higher at greater depths in hypoxic lakes. If DOC and CO₂ concentrations vary across depths, the entrainment fluxes of C into the epilimnion may be biased, and explicitly testing for the nonuniformity of concentrations of these solutes will correct the bias. To test the entrainment and nonuniformity of C concentrations hypotheses, elements of $I(t)$ were modified so as to include entrainment of the hypolimnetic water that resulted from variations in thermocline depth. For detailed equations as well as R scripts implementing this structural hypothesis please see the suggested reading.

observations, their performance is evaluated and the processes that consistently improve model performance are identified.

MODEL CALIBRATION AND SELECTION

To ensure the model variant performs to the best of its ability it needs to be calibrated using a formal data-model fusion approach. The parameters in the models are calibrated using a Bayesian Markov Chain Monte Carlo (MCMC) technique, which is a variant of the Bayesian inversion. Mathematically, Bayesian inversion can be summarized as:

$$p(c|Z) = v_c \times p(Z|c) \times p(c) \quad (30.5)$$

where $p(c|Z)$ is the posterior probability density of parameters c ; $p(Z|c)$ is the likelihood function of parameters c ; $p(c)$ is the prior probability density of parameters c ; and v_c is a normalization constant. If the errors are assumed to be normally distributed, the likelihood function can be expressed as:

$$p(Z|c) = v_L \times \exp \left\{ - \sum_{j=1}^2 \sum_{i=1}^N \frac{(Z_{i,j} - X_{i,j})^2}{2\sigma_{i,j}^2} \right\} \quad (30.6)$$

where $Z_{i,j}$ is the j th observation type at i th time point (there were two observation types: DOC and CO₂ pools); $X_{i,j}$ is the model output for j th observation type at i th time point; N is the total number of j th observation type; $\sigma_{i,j}^2$ is the variance of j th observation type at i th time points; and v_L is a constant. There are two observation types used in this study: (1) two-year time series of epilimnetic DOC pool; and (2) two-year time series of epilimnetic CO₂ pool. These observations are split into two groups: the measurements collected in the first year are used to calibrate the model parameters, and the measurements collected in the second year are set aside for use in evaluating the performance of the models with calibrated parameters. Such design helps control for overfitting, ensuring that the best-performing models capture the signal rather than the noise in the observations.

The prior distributions for the parameters ($p(c)$) are assumed to be uniform. The posterior parameter distributions $p(c|Z)$ are created by sequentially proposing parameters that are randomly sampled from their prior distributions and accepting or rejecting them using the Metropolis criterion (Spall, 2005). The parameters in the posterior distributions may be correlated, which is often the reason for reduction in the acceptance rate of the proposed parameters. To account for the correlations among the parameters and increase the parameter acceptance rate, the proposal distribution is set to uniform only for the first 10,000 iterations (the number of iterations given here yielded optimal parameter acceptance rate, however the reader can experiment with the initial number of iterations). Afterwards, the accepted parameters are used to estimate the parameter covariance matrix, and the proposal distribution for the next 490,000 simulations is switched to the multivariate normal:

$$c^{new} = N(c^{k-1}, C_k) \quad (30.7)$$

where C_k is defined as

$$C_k = \begin{cases} C_0 & k = 10,000 \\ s_d \text{ cov}(c^0, \dots, c^{k-1}) & k > 10,000 \end{cases} \quad (30.8)$$

where $s_i = 2.38/\sqrt{n}$; n is the number of parameters in a model variant; and C_0 is the parameter covariance matrix constructed from the first 10,000 iterations. The Equations 30.7 and 30.8 summarize the parameter sampling technique in the Adaptive Metropolis algorithm (Haario et al., 2001), which is aimed at optimizing parameter acceptance rate. It is important to monitor the parameter acceptance rates and keep them between 23% and 44%, because (1) high acceptance rates lead to misrepresentation of the tails of the posterior parameter distributions, and (2) low parameter acceptance rates may lead to a computationally infeasible increase in total number of simulation, which will be necessary for obtaining sufficient a number of samples and construction of posterior parameter distributions (Gelman et al., 1996). The first half of the accepted parameters is considered to be a “burn-in” period and should be discarded to avoid using model parameters from a nonstationary posterior distribution to calculate maximum likelihood parameter estimates. The second half of the accepted parameters can be used to generate the marginal posterior parameter distributions and calculate the maximum likelihood parameter estimates. Lastly, the model output produced with the maximum likelihood parameter estimates is used to calculate the model fit index, which in turn will help evaluate which model structural assumptions consistently improve model performance.

The model fit index used in this chapter is the Akaike information criterion (AIC, Akaike, 1974), which is calculated as:

$$AIC = 2n - 2 \ln(\hat{L}) \quad (30.9)$$

where n is the number of parameters in a model variant and \hat{L} is the likelihood function ($p(Z|c)$ from Equation 30.5) calculated using the portion of observations reserved for model validation. With 102 model formulations, identifying processes that consistently improve model performance is not trivial. However, if we treat AIC as a function of presence or absence of a particular process in the model (i.e., use binary predictors to model AIC), we can test the significance of the coefficients associated with the presence of a particular process. If the coefficients are significantly smaller than 0, representing a process in the model improves its performance. Such analysis essentially performs a linear regression on the AICs. However, a linear regression model has to satisfy a suite of assumptions, such

as linearity of the relationship and homogeneity of residual variance. Therefore, a nonparametric method for identifying the processes that improve model performance may be preferable.

Here, we will use a machine learning technique, recursive partitioning by conditional inference (Hothorn et al., 2006), to identify which processes consistently reduce the AIC values, thus indicating a consistent improvement in model performance. Unlike many recursive partitioning methods, conditional inference trees do not overfit the models and are not biased towards covariates with many possible splits. This method is available in the R “partykit” package (Hothorn et al., 2015), which is used here to analyze the AIC values.

PROCESSES THAT CONTROL EPIIMNETIC C DYNAMICS

Multiple hypothesis testing revealed that representing physical controls of epilimnetic C dynamics consistently improved model performance, whereas refining the effect of biological controls did not affect it significantly (Figure 30.3). The conditional inference machine learning algorithm identified a cluster of twenty models that had significantly lower AIC values compared to other model variants. All these twenty models represented entrainment of hypolimnetic water, a depth gradient in hypolimnetic CO₂ content, and density-dependent input partitioning of DOC and CO₂ load between epi- and hypolimnion. Notably, representing heterogeneity in DOC lability neither improved nor worsened model performance, which was also true for photooxidation of DOC. The lack of sensitivity of model performance to these structural assumptions likely indicates that the observations do not contain information that could be used to evaluate these hypotheses. In summary, multiple hypothesis testing led to accepting hypotheses #2 and #4.

The question that remains is: how well do the best-performing models represent the observed variation in epilimnetic C dynamics? Since the performance of the top 20 models in simulating temporal variation in lake C pools did not vary substantially, we include the illustration of performance of only one model (Figure 30.4). The chosen model variant does not represent photooxidation and heterogeneity in DOC lability. The model explained a substantial portion of the temporal variability in the epilimnetic CO₂ and DOC pools in the validation data set (81% and 64%,

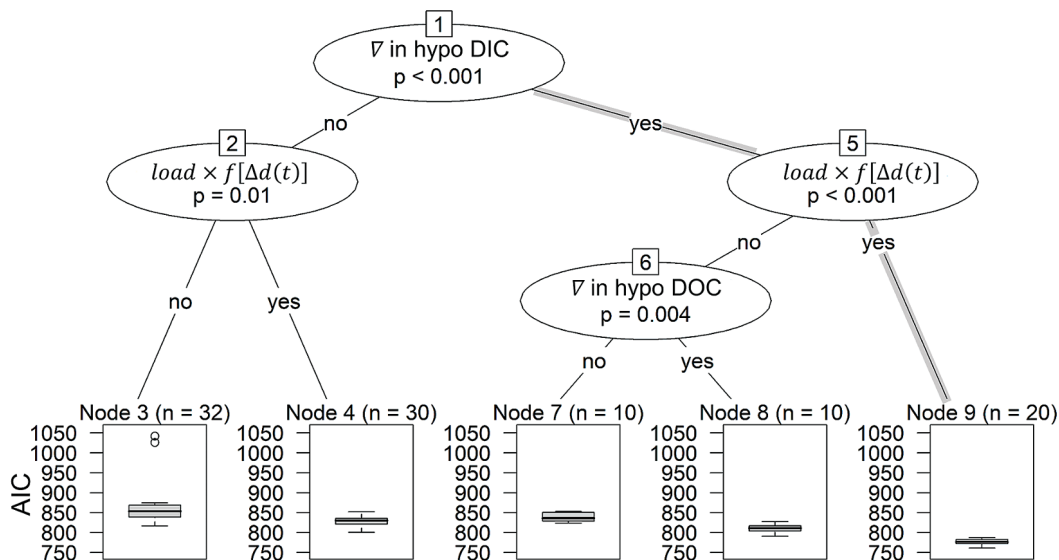


Figure 30.3. Model structural assumptions that significantly reduced AIC. “Yes” and “no” indicate presence or absence of a process or assumption in a model. Thick gray lines highlight a path to a cluster of best-performing models (Node 9). ∇ in hypoDIC denotes the assumption that hypolimnetic CO_2 concentrations were nonuniform across depth and by default includes the representation of entrainment in model variants; $\text{load} \times f[\Delta d(t)]$ indicates that stream loads of DOC and CO_2 are partitioned between the epilimnion and the hypolimnion as a function of the water density difference between the stream and the epilimnion.

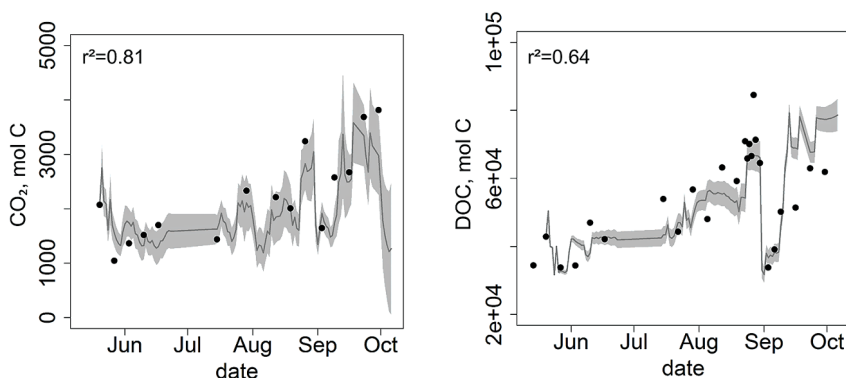


Figure 30.4. Performance of the best models in representing epilimnetic CO_2 and DOC that were not used for model calibration. Gray shaded region represents 2 standard deviations from the model mean, black dots are observed data points. Best-performing models explained 81% of temporal variation in epilimnetic CO_2 pool and 64% - in DOC pool. Although there were 20 model variants that were “best-performing” (Figure 30.3), their representation of the two observed C pools did not vary substantially, therefore performance of only one model is illustrated.

respectively; Figure 30.4) and most observations fell within two standard deviations of the mean predicted value. Thus, model-data fusion combined with multiple hypothesis testing approach helped build a well-performing model for simulating epilimnetic C dynamics.

Although the example in this chapter focuses on lake biogeochemistry, the general framework of model-based multiple hypothesis testing could

potentially be applied in terrestrial biogeochemistry or many other fields as it is not limited to a particular model structure. Given that the available observations are informative for the hypotheses being tested, much knowledge can be gained about mechanistic controls of a variable of interest. The results of the analysis described here may also indicate that the observations are noninformative for evaluating whether a particular mechanism is

important, which may manifest as lack of response of the fit index to the additional processes. Such negative results may provide guidance for future experimental design so as to maximize information content in the collected observations.

SUGGESTED READING

Hararuk O, Zwart JA, Jones SE, et al. (2018) Model-Data Fusion to Test Hypothesized Drivers of Lake Carbon Cycling Reveals Importance of Physical

Controls. *Journal of Geophysical Research Biogeosciences* 123(3):1130–1142.

QUIZZES

1. What are the three stages of model development?
2. What is the objective of data-model fusion?
3. How does data-model fusion facilitate strong inference?
4. What are the criteria to accept or reject a hypothesis?

CHAPTER THIRTY-ONE

Data-Constrained Uncertainty Analysis in Global Soil Carbon Models

Zheng Shi

University of Oklahoma, Norman, USA

CONTENTS

Introduction /	263
Alternative Model Structures /	264
Datasets and Data-Model Fusion /	267
Posterior Distribution of Model Parameters /	268
Uncertainties in Soil Carbon Projections Under Representative Concentration Pathway 8.5 /	269
Sensitivity to Initial Conditions and Model Parameters /	270
Suggested Reading /	271
Quizzes /	271

Large model uncertainty in projected future soil carbon (C) dynamics has been well documented. However, our understanding of the sources of this uncertainty is limited. This chapter is designed to illustrate the projection uncertainties induced by model structures and parameter values. Three representative soil carbon models are compared in terms of their predictions of soil carbon change in a future climate warming scenario. The parameter values in each model were derived by fitting modeled soil carbon to observations. We see that uncertainty often increases with complexity of a model. The larger uncertainty in the complex models suggests that we need to strike a balance between model complexity and the need to include diverse model structures in order to forecast soil C dynamics with high confidence and low uncertainty.

INTRODUCTION

The structure and parameter values of a model together determine its predictions, and both contribute to model uncertainty. Differences in structure and parameters among global land carbon models give rise to projection uncertainty, which is typically high across models in future scenario studies. Past studies have linked different model structures to differences

in soil C projections (Wieder et al., 2013), and parameterization (i.e., the choice of parameter values) has been shown to cause large uncertainty in projected change in soil C (Hararuk et al., 2015). In other words, parameterization is likely to interact with model structure to impact a model's projections.

Soil C decomposition models, in particular, are becoming more diverse. For example, recent innovations include microbial models that simulate decomposition processes with explicit microbial traits, and models with dynamics that explicitly account for interactions between soil layers at different depths (Koven et al., 2013; Wieder et al., 2015). Exploring uncertainty generated by model structures and parameterization is critical for global land C modeling, but until now, relatively little effort has been dedicated to addressing it, in part due to enormous computational cost.

Here we demonstrate how a Bayesian framework in which the posterior parameter distributions are sampled via a Markov Chain Monte Carlo (MCMC) technique, can overcome the computational hurdle and generate a large ensemble of potential parameter sets within reasonable physical and biological boundaries. Applied to the soil carbon system, the approach relies on the simplifying assumption that the constituent SOM pools are at steady state. This

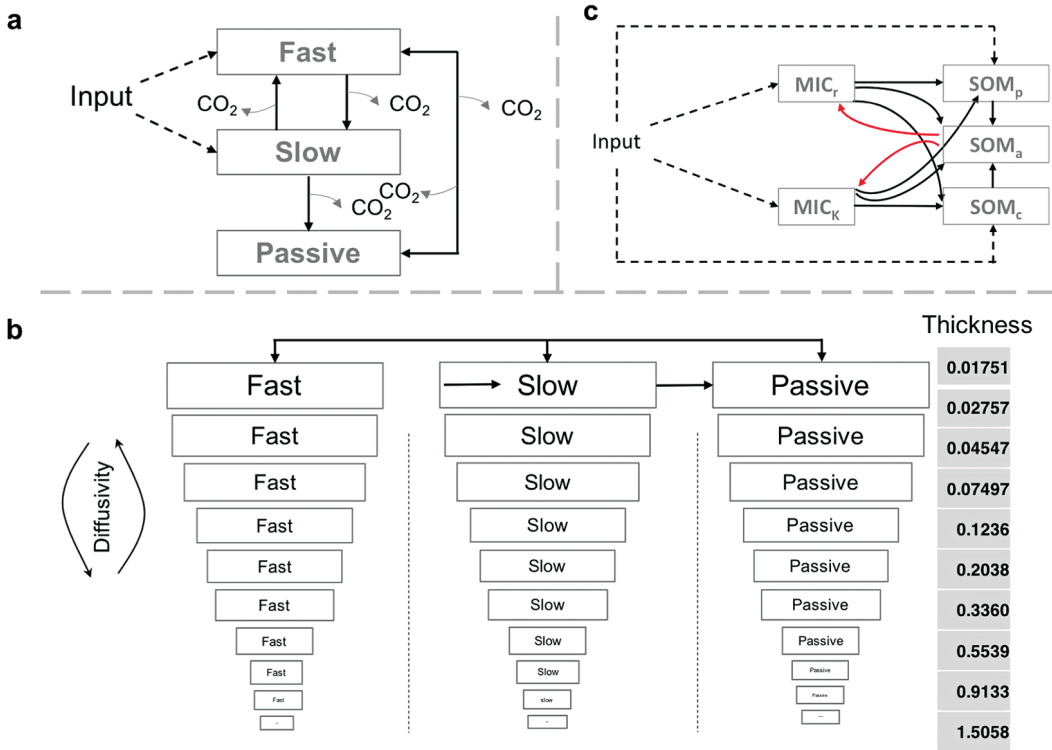


Figure 31.1. Soil carbon decomposition models with distinct structures. (a) Conventional Century-type model of CLM 4.0; (b) Vertically-resolved model in CLM 4.5bgc; (c) Microbial-Mineral Carbon Stabilization (MIMICS), which represents microbial biomass explicitly. MIC: microbial biomass; SOM: soil organic matter.

ensemble generates similar current global soil carbon content compared with the Harmonized World Soil Database (HWSD) and Northern Circumpolar Soil Carbon Database (NCSCD). We choose three representative soil C decomposition models with distinct structures, suitable for application across a global grid (Figure 31.1).

The three models include the conventional soil C decomposition model embedded in the Community Land Model version 4.0 (CLM 4.0), a soil C decomposition model with explicit soil depth embedded in CLM 4.5 (Koven et al., 2013) and a microbial model (the Microbial-Mineral Carbon Stabilization: MIMICS; Wieder et al., 2015). A projection based on the RCP (Representative Concentration Pathway) 8.5 emissions scenario was carried out on a global grid for all the three models with their posterior parameter ensembles.

ALTERNATIVE MODEL STRUCTURES

The conventional model (Figure 31.1a) represents soil C decomposition using three C pools and can be written in a matrix form (Chapter 5) as:

$$\frac{d\mathbf{X}}{dt} = \mathbf{I} + \mathbf{A}\xi(t)\mathbf{K}\mathbf{X}(t) \quad (31.1)$$

$\mathbf{I} = [i_1 \ i_2 \ i_3]$ is the litter input to the three soil carbon pools (labile, slow and passive soil C).

$$\mathbf{A} = \begin{pmatrix} -1 & f_{12} & f_{13} \\ f_{21} & -1 & 0 \\ f_{31} & f_{32} & -1 \end{pmatrix}$$

is a transfer matrix

among soil C pools, where f_{21} is derived using the equation $f_{21} = 1 - a - f_{31}$ where $a = a_1 - a_2 \times (1 - \text{sand}\%)$ where $\text{sand}\%$ is sand proportion. The respiration coefficients can be derived by $1 - f_{i,j}$ for each carbon pool. $\mathbf{K} = [k_1, k_2, k_3]$ is baseline turnover rate (yr^{-1}) of soil C pools. These baseline rates are modified by the environmental modifier, ξ , which is a product of temperature scalar, soil moisture scalar, soil nitrogen scalar and oxygen scalar. $\mathbf{X} = [x_1, x_2, x_3]$ is the time-varying soil C content in each of the three pools, the state variables of the model. To be consistent with the other two models, we only calculate temperature scalar using the equation $Q_{10}^{((T_{\text{soil}} - 25)/10)}$ where Q_{10} is temperature sensitivity of decomposition and T_{soil} is soil temperature. The

$$V_{m+1,m} = \begin{bmatrix} 0 \\ v_{m+1,m} \\ v \\ v_{m+1,m} \\ v_{m+1,m} \\ v_{m+1,m} \\ v_{m+1,m} \\ v_{m+1,m} \end{bmatrix}$$

is the received fraction of carbon transferred from the upper soil layer, m , to layer $m+1$. Also, $v_{m,m} = v_{m-1,m} + v_{m+1,m}$ for a given m . \mathbf{V} can be approximated using the model parameter diffusivity (D_1 for non-permafrost diffusivity and D_2 for permafrost diffusivity). Note that compared to the environmental modifier in the conventional model, there is an additional scalar, the depth scalar (r_z) which is computed with $r_z = \exp\left(-\frac{z}{z_\tau}\right)$ where z_τ is the e-folding depth. To be consistent with the other two models, we only calculate temperature scalar using the equation $Q_{10}^{(T_{soil} - 25)/10}$ where Q_{10} is temperature sensitivity of decomposition and T_{soil} is soil temperature. The rest of the environmental scalars are direct outputs from running CLM 4.5. Therefore, there are 13 global parameters (f 's, k 's, a_1 , a_2 , Q_{10} , D_1 , D_2 , z_τ) in the vertically-resolved model.

The Microbial-Mineral Carbon Stabilization (MIMICS) model was described by Wieder et al. (2015). There are two soil microbial C pools (MIC_r and MIC_k) and three soil C pools, available soil C (SOM_a), physically-protected C (SOM_p) and chemically recalcitrant C (SOM_c) (Figure 31.1c). Michaelis-Menten equations are adopted to describe soil C uptake by soil microbes. The dynamics of the soil C can be represented by the following equations:

$$SOM'_p = R_{l-p} + R_{mic-p} - SOM_p \times D \quad (31.3)$$

$$SOM'_c = R_{l-c} + R_{mic-c} - U_{c-k} - U_{c-r} \quad (31.4)$$

$$SOM'_a = R_{mic-a} + U_{c-k} + U_{c-r} + SOM_p \times D - U_{a-k} - U_{a-r} \quad (31.5)$$

where R_l is the input to soil C from litter ($R_{l-p} = f_m \times \text{total_input}$ and $R_{l-c} = f_s \times \text{total_input}$) and R_{mic} is the input to soil C from microbial decay, D is the turnover rate for SOM_p , U_{c-k} is the uptake of SOM_c by K-selected microbes, U_{c-r} is the uptake of SOM_c

by r-selected microbes, U_{a-k} is the uptake of SOM_a by K-selected microbes and U_{a-r} is the uptake of SOM_a by r-selected microbes. The uptake process takes the form of Michaelis-Menten equation:

$$U = MIC V_{max} \frac{SOM}{K_o K_m + SOM}$$

where MIC is the microbial biomass, V_{max} is the maximum reaction rate, SOM is the soil C content, K_o is the modifier for oxidation of SOM , K_m is the half saturation constant. For more model details, see Wieder et al. (2015).

Our implementation is based on the original MIMICS model, with slight modifications for simplification. In total, there are 22 global parameters. Since the range for most of the parameters is not well characterized in the literature, we prescribe the minimum of each parameter as the default value divided by three and the maximum as the default value multiplied by three. In addition to the soil C dynamics, the two microbial C pools are represented by the equations:

$$\frac{dMIC_r}{dt} = R_{l-r} + U_{a-r} \times MGEI - MIC_r \times \tau_r \quad (31.6)$$

$$\frac{dMIC_k}{dt} = R_{l-k} + U_{a-k} \times MGEI - MIC_k \times \tau_k \quad (31.7)$$

where R_{l-r} and R_{l-k} are the input to r- and K-selected soil microbes, respectively.

$$R_{l-r} = (U_{m-r} + U_{s-r}) / (U_{m-r} + U_{m-k} + U_{s-r} + U_{s-k}) \times (\text{total_input} - R_{l-p} - R_{l-c}) \quad (31.8)$$

$$R_{l-k} = (U_{m-k} + U_{s-k}) / (U_{m-r} + U_{m-k} + U_{s-r} + U_{s-k}) \times (\text{total_input} - R_{l-p} - R_{l-c}) \quad (31.9)$$

U_{m-r} and U_{m-k} are the uptakes of metabolic litter by r- and K-selected microbes, respectively; and U_{s-r} and U_{s-k} are the uptakes of structural litter by r- and k-selected microbes, respectively; all the U's are calculated with default parameters in MIMICS and litter from CLM 4.5 with the sole purpose of normalizing the input to r- and K-selected microbes. U_{a-r} and U_{a-k} are the uptake of available soil C by r- and K-selected microbes, respectively. MGE1 is the microbial growth efficiency for uptake of SOM_a. τ_r and τ_k are the turnover rates of r- and K-selected microbes, respectively.

Carbon use efficiency or microbial growth efficiency (MGE) is a key parameter in microbial models. However, we do not consider it as a parameter in our study since we are using microbial biomass data as an input to the MIMICS model.

DATASETS AND DATA-MODEL FUSION

The observations we will use are re-gridded top-soil organic carbon (0~30 cm) and subsoil organic carbon (30~100 cm) from the Harmonized World Soil Database (HWSD; <https://daac.ornl.gov/SOILS/guides/HWSD.html>). The native resolution (30 arc sec) in HWSD is re-gridded to match the default grid used by the CLM 4.5 model.

Due to the possible under-estimation of permafrost soil C in HWSD, we replace it with the Northern Circumpolar Soil Carbon Database (NCSCD; <http://bolin.su.se/data/ncscd/netcdf.php>) in permafrost areas. The NCSCD was developed to quantify the Northern Circumpolar permafrost soil C stocks down to three meters. There are four soil layers in this database, 0–30 cm, 0–100 cm, 100–200 cm and 200–300 cm. We regrid the NCSCD from 1-degree resolution to that of the CLM 4.5 grid.

The microbial biomass C is prescribed as the steady state in the MIMICS model, based on the Global Microbial Biomass dataset (https://daac.ornl.gov/SOILS/guides/Global_Microbial_Ciomass_C_N_Ph.html). The steady state assumption is justified, given the fast turnover of microbes, typically less than one year, which leads to fast equilibration of the microbial biomass pool. A global parameter, f_r (fraction of r-selected microbial biomass) is multiplied by total microbial biomass from the input database to calculate the r-selected microbial biomass. The remainder is the K-selected microbial biomass. We regrid the original 0.5° resolution data to the CLM grid, matching the other datasets above.

To fit the observations, we use Bayes' theorem to estimate parameter values and associated uncertainties:

$$p(\theta|Z) = \frac{p(Z|\theta) \times p(\theta)}{p(Z)} \quad (31.10)$$

where, $p(\theta|Z)$ is the posterior distribution of the parameters θ given the observations Z . $p(Z|\theta)$ is the likelihood function for a parameter set calculated with the assumption that each parameter is independent from all other parameters, and has log-normal distribution with the difference between model and data being a zero mean:

$$P(Z|\theta) \propto \exp\left\{-\sum \frac{[Z_i - \emptyset_i \times X]^2}{2\sigma_i^2}\right\} \quad (31.11)$$

Here Z_i is the logarithm of i^{th} soil C observation in the observational database, X are the logarithms of the carbon pools from the model, and \emptyset is the mapping vector that maps the simulated carbon pools to observations. X is derived by assuming the current soil status is at steady state. We will be conservative in assigning errors to the soil C with $\sigma = 0.5 Z_i$. For the conventional model, we assimilate the data by aggregating all the soil layers together which is 0–100 cm in non-permafrost regions and 0–300 cm in permafrost regions; for CLM 4.5, we assimilate data for 0–100 cm in HWSD for non-permafrost soils, 0–100, 100–200 and 200–300 cm in NCSCD for permafrost soils, independently. In contrast to soils elsewhere, permafrost regions contain a huge amount of carbon stock in deeper soil, which justifies the use of deeper soil carbon data for these regions. For MIMICS, we assimilate the soil C data down to 100 cm only due to the explicit 1 m depth parameterization of this model.

We will assume that the parameters are distributed uniformly within their prior ranges. Since the range for most of the parameters in MIMICS is unknown, we assume the range of the distribution to be $[\theta_0/3, 3\theta_0]$, where θ_0 is the default value.

To obtain posterior probability distributions of parameters we will employ the Metropolis-Hastings (M-H) algorithm, which is a Markov Chain Monte Carlo (MCMC) technique (Hastings, 1970). A detailed description of the M-H algorithm can be found in Xu et al. (2006).

In brief, the M-H algorithm consists of iterations of two steps: a proposing step and a moving

step. In the proposing step, a new parameter set θ^{new} is proposed based on the previously accepted parameter set θ^{old} and a proposal distribution, which is uniform in our study:

$$\theta^{new} = \theta^{old} + r \times (\theta_{max} - \theta_{min}) / D \quad (31.12)$$

where θ_{max} and θ_{min} are the maximum and minimum values of parameters, r is a random variable between -0.5 and 0.5 , and D is used to control the proposing step size and is set to 5 as in Xu et al. (2006)⁴². In each moving step, θ^{new} is tested against the Metropolis criterion to examine if the new parameter set should be accepted or rejected. We treat the first 2500 accepted samples as a burn-in period, discarding those samples, then use the rest to generate posterior parameter distributions. In total, there are 50000 accepted samples to construct the posterior distribution.

POSTERIOR DISTRIBUTION OF MODEL PARAMETERS

Outcomes of the probability inversion are shown in Figure 31.2. A narrow posterior distribution indicates that a parameter is well constrained. We can see that the inversion was effective in terms of constraining the targeted parameters in the

three soil C decomposition models. Specifically, coefficients (i.e., t_1 and t_2) for calculating f_{21} (fraction of C in fast soil C transferring to slow soil C), decay rate of passive soil C (k_3) and temperature sensitivity (Q_{10}) were well constrained in the conventional Century-type model (Figure 31.2a); observed soil C vertical profiles further helped constrain the decay rate of slow soil C (k_2) in the vertically-resolved model, particularly for the vertical diffusivity parameters (D_1 and D_2) and e-folding depth (z_i) (Figure 31.2b). Interestingly, the posterior mean of D_2 was found to be larger than D_1 , as diffusivity in permafrost soil was found to be faster than non-permafrost soil, which is mainly due to higher cryoturbation. The Q_{10} mean is 1.25 in the conventional model and 1.06 in the vertically-resolved model, both of which are less than the default value (2), but close to empirical values. The transfer coefficients (f_{ij}) were not well constrained in either of the two models.

In MIMICS, parameters related to uptake rate (V_s and V_i) and desorption rate of physically-protected soil C (D_a and D_b), and proportion of litter input and the two microbial C pools (f_m , f_s and f_r) were well constrained (Figure 31.2c). None of the modifiers (e.g., V_{mrc} , V_{mkc} , K_{mrc} and K_{mkc}) for calculating uptake rate and half saturation constant were well constrained. Additional

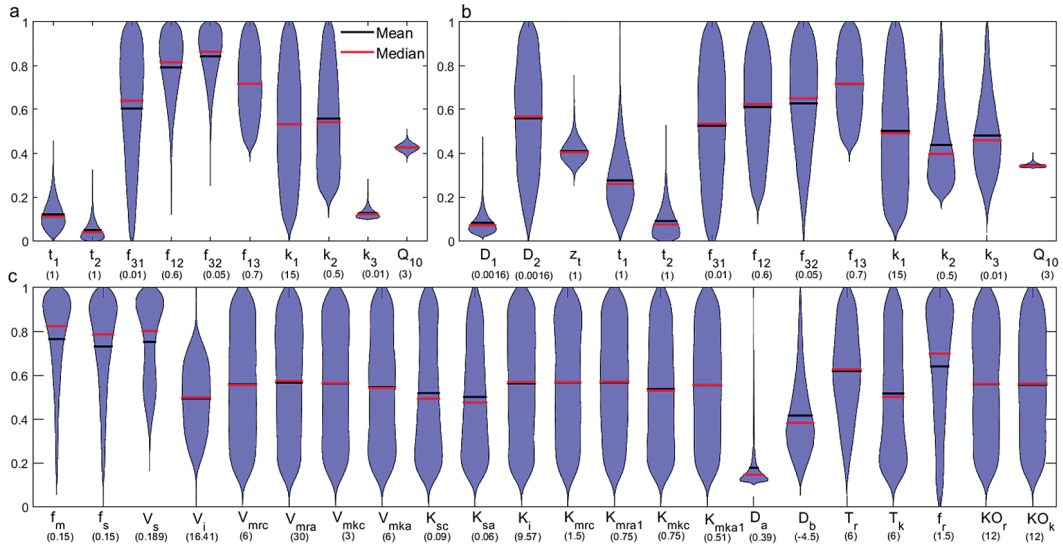


Figure 31.2. Violin plot of the accepted model parameter values in the three decomposition models. (a) Conventional model; (b) Vertically-resolved model; (c) Microbial model (MIMICS). The narrower the distribution of a parameter, the better it is constrained. Note that the values on the y-axis are normalized to the range of $[0, 1]$ by the corresponding values in the brackets under each parameter. In the other words, the real values can be derived by multiplying by the values in the brackets.

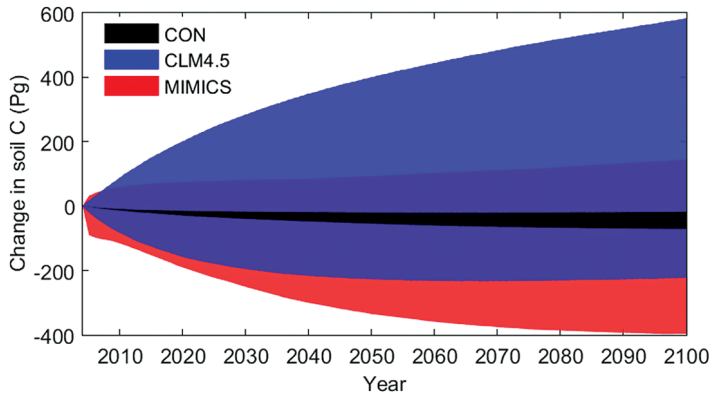


Figure 31.3. Changes in global total soil carbon under RCP 8.5.1000 parameter sets were randomly sampled from the posterior distribution of parameters to generate the temporal trajectories in each model. The shaded area represents the 95% confidence interval generated by 1000 parameter sets. The time step is shown in years, from 2005 to 2100. CON = conventional Century-type model; CLM4.5 = vertically-resolved model.

datasets are especially needed to tease apart multiple processes and further reduce the uncertainty. Results from this study highlight that data constraints may limit the ability of data assimilation to reduce uncertainty in more complicated model structures. Isotopic data in C processes show great potential to constrain these processes. Indeed, ^{14}C soil profiles have been used to constrain transfer coefficients and turnover rates at multiple sites; isotopic labeling to trace C pathways is another powerful tool to provide additional constraints to relevant processes, such as distinguishing root respiration from total soil respiration, sources of input, or proportion of different soil C pools. Other additional data, such as soil respiration and soil C incubation datasets, are equally valuable constraints. We therefore advocate using isotopic data and other datasets as complementary sources to better constrain model parameters and hence projections.

UNCERTAINTIES IN SOIL CARBON PROJECTIONS UNDER RCP8.5

To illustrate the impact of parameter uncertainty on long-term soil C projection, we performed forward runs over the 21st century with 1000 sets of parameter values drawn from the posterior distribution for each model. The soil C input and environmental modifiers (except the temperature scalar) used to drive the models were derived from running original CLM 4.5 under RCP 8.5 scenario, with climate forcing from the Community Earth

System Model (CESM) for 2005–2100. We output monthly litter production from the CLM 4.5 run and use it to provide the soil C input to drive our soil C models.

We randomly sampled 1000 parameter sets out of the accepted posterior values. With each of the sampled parameter sets, we forced the vertically-resolved model with C input and environmental modifiers obtained from CLM 4.5 under RCP 8.5 from 2005~2100. For the conventional and microbial model, we derived total inputs and mean environmental modifiers of all the ten soil layers to force the two models. For the two non-microbial models, we used a monthly time step. For the microbial model, daily time step was used due to its nonlinear nature causing instability for longer time steps.

The three models projected substantially different changes and trajectories in global total soil C over the 21st century (Figure 31.3). The conventional model projected consistent soil C loss with the least uncertainty (95% confidence interval: $-71 \sim -17$ Pg). Adding vertical resolution or microbial dynamics to the conventional model increased the projection uncertainty (95% confidence interval: $-222 \sim 583$ Pg C and $-397 \sim 144$ Pg C, respectively) as well as the sign of the soil C-climate feedback, depending on parameters. The uncertainties in the vertically-resolved model or MIMICS are more than ten times larger than that in the conventional model. This interesting result shows that using more parameters and more explicit dynamics may lead to a larger

prediction uncertainty due to feedbacks in the model dynamics.

SENSITIVITY TO INITIAL CONDITIONS AND MODEL PARAMETERS

Our results show that the initial conditions (S_i) tightly correlated with the projected soil C in the two nonmicrobial models at global scale, but did not correlate well with the changes in soil C in any of the models (Figure 31.4). The microbial model's initial conditions were not strongly correlated with projected soil C at the global scale. The findings suggest that, in general, uncertainty in the initial conditions propagates through the simulation to the projection of future soil C, and this propagation is especially evident in the two nonmicrobial models.

Besides initial conditions, model parameters are also able to affect predicted soil C or C changes directly, or indirectly through influencing initial conditions following a spin-up (Figure 31.4). In the conventional model, predicted soil C did not significantly correlate with any individual model parameter; changes in soil C were positively associated with k_2 (decay rate of slow soil C), but negatively with k_3 (turnover rate of passive soil C) and

Q_{10} (temperature sensitivity of soil C turnover). In the vertically-resolved model, predicted soil C positively associated with D_1 (diffusivity in non-permafrost soils) and k_3 , but negatively with k_2 ; like predicted soil C, changes in soil C were positively associated with D_1 and k_3 , but negative with k_2 . Predicted soil C change weakly associated with V_s (regression coefficient for calculating maximum reaction rate) and D_a (coefficient for calculating desorption rate from physically protected soil C to available soil C); projected soil C content had no significantly linear relationships with any of the model parameters.

In summary, this chapter demonstrates the importance of model structure and parameterization in determining the predicted soil C response to climate change. To increase confidence in soil C projection, diverse model structures are necessary. The vertically-resolved CLM4.5 model and the microbial MIMICS model showed much greater uncertainty in projected soil C under RCP 8.5, while the conventional model consistently predicted strong positive C-climate feedback. CLM 4.5 and MIMICS outperformed the conventional model in terms of estimation in the spatial distribution of soil C. However, the larger uncertainty in the projections of soil C by the latter two models

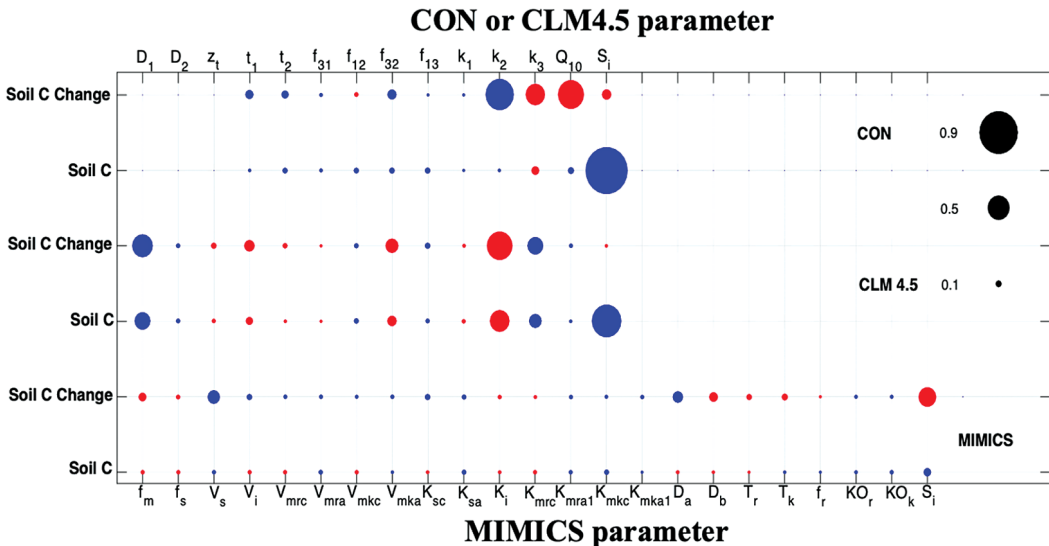


Figure 31.4. The relationships between soil C, model parameters and initial conditions. The linear correlations between soil C or C changes, and initial conditions (S_i) and model parameters under RCP 8.5 in the three models (see Figure for key). Blue circles represent positive correlations and red circles represent negative correlations. The size of a circle corresponds to the correlation coefficient (legend) between model parameters or initial conditions (horizontal axis) and soil C or C changes (vertical axis).

also suggests that we need to strike a balance when adding process detail, as increased model complexity tends to amplify uncertainty.

SUGGESTED READING

Shi, Z., Crowell, S., Luo, Y., & Moore, B. (2018). Model structures amplify uncertainty in predicted soil carbon responses to climate change. *Nature Communications*, 9(1): 1–11.

QUIZZES

1. What are the major differences between conventional models and microbial models?
2. Describe the matrix equations for the conventional model and depth-resolved model.
3. What are the dominant parameters for modeled soil carbon and soil carbon changes?
4. How might we further constrain model predictions in more complex models?



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER THIRTY-TWO

Practice 8

INFORMATION CONTENTS OF LAND CARBON POOL AND FLUX MEASUREMENTS TO CONSTRAIN A LAND CARBON MODEL

Enqing Hou

Northern Arizona University, Flagstaff, USA

CONTENTS

Introduction / 273

Summary / 283

Land carbon models are useful tools to predict the impacts of global change on terrestrial ecosystems. However, uncertainties in the model predictions are large, and need to be reduced to increase our confidence in model predictions. Both carbon pools and carbon fluxes are usually measured in land carbon studies. This practice helps you understand the value of pool and flux measurements to constrain a land carbon model through data assimilation. To this end, four exercises are designed, which are model run without data assimilation and model runs with assimilation of carbon pool and/or carbon flux measurements. Through these exercises, you will learn how to select appropriate measurements to constrain model predictions of land carbon dynamics.

INTRODUCTION

Land carbon models are useful tools to investigate global change impacts and explore options to mitigate climate change. However, uncertainties in model predictions of land carbon dynamics are large. There is an urgent need to reduce the uncertainties and increase our confidence in model predictions. Model uncertainties have three major sources, which are model structure, model parameterization, and external forcing (see Chapter 21).

Data assimilation directly addresses uncertainties stemming from model parameterization by optimizing model parameters for a good model-data fit. It also helps shed light on uncertainty stemming from model structure.

Both carbon pools and carbon fluxes are usually measured in land carbon studies. This practice investigates the relative benefits of assimilating pool and flux measurements into a model in terms of improving our predictive understanding of land carbon dynamics. The practice encompasses four exercises: model runs without data assimilation; with assimilation of pool measurements; with assimilation of flux measurements; and with assimilation of both pool and flux measurements. Pool measurements include sizes of leaf, wood, and root biomass carbon. Flux measurements include net ecosystem exchange (NEE), gross primary production (GPP), and ecosystem respiration (Reco).

The model runs are performed in the training software CarboTrain by selecting unit 8 and Exercises 1–5. Note that a full data assimilation run in this practice will take about eight hours to execute, depending on the specifications of your personal computer. Therefore, Exercises 1–4 are based on prepared model outputs. Only Exercise 5 involves a full data assimilation run.

EXERCISE 1: Model run without data assimilation

A model run without data assimilation is useful as a starting point. It can give information of the model and the data assimilation system, and acts as a baseline to compare to model runs with assimilation of pool and/or flux measurements. As described in Chapter 29, the structure and prior parameter uncertainty of a model constitute the prior knowledge about a system. Prior parameter uncertainty is characterized by the range and prior probability of a parameter. The range of possible values for a parameter is usually defined by knowledge from the literature or observations. The prior probability of a parameter is often assumed to follow a uniform distribution between a minimum and maximum bound, but can also have another distribution (e.g., Gaussian) within the range of the parameter. In addition to the prior range and distribution, the choice of cost function and the method to generate new sets of parameters at each step of simulation can affect the posterior distributions of model parameters, and predictions based on these.

To do this exercise, select unit 8 and Exercise 1 in the main window of CarboTrain. Select the output directory, and click the Run Exercise button. Model outputs will then appear in the `DA_with_no_obs.zip` folder in the output directory you set.

Before looking at the model outputs, let's get familiar with the model files and workflow of the data assimilation system. You can explore the data assimilation system by looking into the `Source_code/TECO_2.3` folder. For this exercise, you use the name-list file `workshop_nml/teco_workshop_da_no_obs.txt`. Settings in this file are passed to the script of `TECO v2.3` (open the file `Source_code/TECO_2.3/TECO2.3.f90` with a text editor such as Notepad++ to view the Fortran source code of `TECO v2.3`). Parameter default values and ranges may be found in the `Source_code/TECO_2.3/input/` folder. The `TECO` model

will use parameter values in `SPRUCE_pars.txt` as the initial parameter values. Minimum and maximum values of parameters to build the prior parameter uncertainties are sourced from `SPRUCE_da_pars.txt`. The latter file also specifies whether a parameter is fixed or not. If the value under a parameter is 0, it means the parameter is fixed and will not change during the model run. If the value under a parameter is 1, it means the parameter value will be sampled from its prior distribution during the model run. Since no measurements are selected for the current exercise, all prior parameter values created at the parameter generation step (Lines 1065–1082 in `TECO2.3.f90`) will be accepted as the posterior parameter values (Lines 8083–8086).

Change the directory back to `DA_with_no_obs.zip` (may need to be unzipped before opening) to examine the model files for this exercise. You can find model input files for this exercise in `DA_with_no_obs/input`, model output files in `DA_with_no_obs/output`, an R script used for visualizing model outputs in `DA_with_no_obs/R` code for `DA Unit 8.R`, and plots showing model outputs in `DA_with_no_obs/plot/DAUnit8`.

Subfolders under `DA_with_no_obs/output` receive model outputs from forward runs (subfolder `SPRUCE`), model outputs without data assimilation (subfolder `DA_nomeasure`), model outputs from assimilation of carbon pool measurements (subfolder `DA_cpool`), model outputs from assimilation of carbon flux measurements (subfolder `DA_cflux`), and model outputs from assimilation of both carbon pool and carbon flux measurements (subfolder `DA_cpool_cflux`).

In this exercise we wish to examine outputs from a model run with default parameter values and ranges. If you open the folder `DA_nomeasure`, you will find the `Paraest.txt` file, which contains the posterior parameter values. The folder also contains files such as `Simu_dailyflux001` and `Simu_dailyflux50`, which contain forward

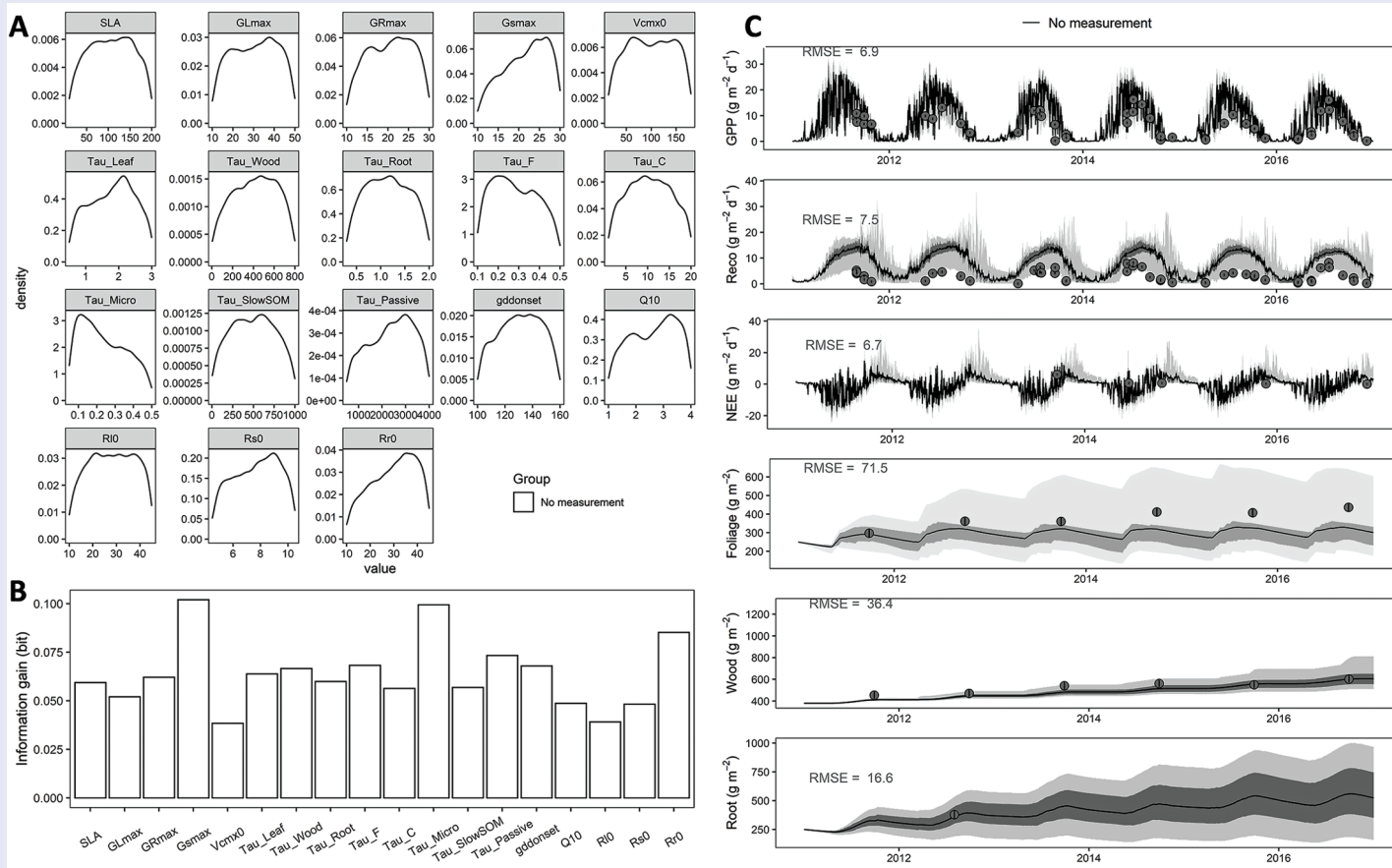


Figure 32.1. Results after model run without data assimilation. (A) Posterior distributions of parameter values. (B) Information gain. (C) Model simulations. Full name and units of parameters can be seen in “Source_code/TECO_2.3/annotations in TECO model/parameter_file_annotations.jpg”. In (C), points and error bars (not clearly visible as range is smaller than the point size) indicate the mean and estimated standard deviation of measurements, respectively. Black line, deep gray band, and light gray band indicate the mean, first to third quantiles, and range of 50 simulations, respectively. RMSE indicates root mean square error.

simulations of land carbon dynamics using randomly selected sets of posterior parameter values.

Posterior distributions of model parameters, and predictions based on these, are available in the folder `DA_with_no_obs/plot/DAUnit8`. The file `"1_Density_no_measurement.png"` corresponds to Figure 32.1A and B, below, while `"1_Simulation_no_measurement.png"` corresponds to Figure 32.1C. For each parameter, posterior values tend to cluster near the mean value rather than distribute uniformly within the parameter range (Figure 32.1A). The parameter values cluster together because of the location optimization by random walk in the data assimilation system (Line 8228–8243 in `TECO2.3.F90`). The location optimization technique generates new parameter values near the parameter values of the previous simulation by narrowing parameter spaces with a distance criterion. The purpose is to accelerate the progress of optimization. Similarly, model predictions of land carbon dynamics tend to cluster near the means rather than distribute uniformly (Figure 32.1C).

As discussed in Chapter 29, the relative information content of a model on predictions and parameters can be quantified and compared before and after data assimilation using the Shannon information index. In brief, the entropy of null knowledge on parameters (or predictions), H_0 , can be calculated as:

$$H_0 = \log_2 n \quad (32.1)$$

where n is the number of bins of posterior parameter values. A bin is a subgroup of values within the posterior parameter range. Here we set n to 200, meaning we divide the parameter range into 200 equally spaced intervals between the minimum and maximum bound of the posterior parameter distribution. We

could choose another arbitrary value for n . Since the base is set to be 2, the unit of H would be bit. The entropy of the probability distribution function of parameters (or predictions) generated by a model alone (or a model with assimilation of data), $H(X_m)$, is calculated as:

$$H(X_m) = \sum_{i=1}^n p(x_{m,i}) \log_2 p(x_{m,i}) \quad (32.2)$$

where X_m is a parameter (or a carbon pool or flux) simulated by the model alone, $x_{m,i}$ is the mean value of X_m in a bin. Relative information of the model (I_m) is calculated as:

$$I_m = H_0 - H(X_m) \quad (32.3)$$

Similarly, the entropy of the PDF of a parameter (or a carbon pool or flux) simulated by a model following assimilation of datasets, $H(X_{md})$, is calculated as:

$$H(X_{md}) = \sum_{i=1}^n p(x_{md,i}) \log_2 p(x_{md,i}) \quad (32.4)$$

where X_{md} is a parameter (or a carbon pool or flux) simulated by the model after assimilation of the datasets, $x_{md,i}$ is the mean value of X_{md} in a bin. Relative information of the dataset(s), I_d , is calculated as:

$$I_d = H_0 - H(X_{md}) \quad (32.5)$$

Model biases in simulating C pools and fluxes can be quantified using the Root Mean Square Error (RMSE), which is calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\text{Prediction}_i - \text{Measurement}_i)^2}{N}} \quad (32.6)$$

where N is the number of measurements, $Prediction_i$ means the i th predicted value of a C pool or flux, $Measurement_i$ means the i th measured value of a C pool or flux. Model information quantified using the above method is shown in Figure 32.1B.

QUESTIONS:

1. Will relative information change if you change the logarithm base and/or the number of bins?
2. Which indexes are used to indicate model bias and uncertainty in simulations?

EXERCISE 2: Model run with assimilation of carbon pool measurements

For this exercise, you select unit 8 and Exercise 2 in CarboTrain. Select the output directory and click Run Exercise. The data assimilation system will read carbon pool measurements in the file `input/SPRUCE_cpool.txt` and compare the measurements against the simulated carbon pools (Lines 7554–7597 in `TECO_2.3.f90`), in order to decide whether a set of parameter values should be accepted as posterior parameter values or not using the cost function. You can find model outputs in `DA_with_cpools.zip`.

You can find three plots in the folder `DA_with_cpools/plot/DAUnit8`. These show comparisons between the results of the model when run without data assimilation (as in Exercise 1) and with assimilation of carbon pool measurements. The figures compare posterior parameter values (“2_Density_C_pools.png”, Figure 32.2A and B) and simulated carbon dynamics (mean values in “2_Simulation_C_pools.png” and “2_1_Simulation_C_pools_uncertainty.png”, Figure 32.2C). You can see that the posterior distribution of specific leaf area is bell-shaped (Figure 32.2A). Information on specific leaf area is essentially higher following assimilation of carbon pool measurements (Figure 32.2B). Both results suggest that specific leaf area is well informed by the carbon pool measurements. The turnover times of leaf and root are also informed by the carbon pool

measurements (Figure 32.2A and B). Model parameters such as basal respiration rate of leaf, wood, and root are not strongly informed by the carbon pool measurements (Figure 32.2A and B). These results indicate that carbon pool measurements have much more information on specific leaf area and the turnover time of carbon pools compared to other parameters of the model.

After assimilating the carbon pool measurements, both model bias and uncertainty in simulating leaf carbon pool are markedly reduced (Figure 32.2C), suggesting that simulated leaf carbon pool is well informed by carbon pool measurements. Model biases and uncertainties in simulating GPP, Reco, and NEE also tend to decrease (Figure 32.2C), probably because of constrained photosynthesis rate and autotrophic respiration which are closely related to leaf carbon pool. Surprisingly, model biases and uncertainties in simulating wood and root carbon pools are not reduced after assimilating the carbon pool measurements (Figure 32.2C). This is due to the relatively short simulations and data time series, relative to the timescale of vegetation structural development; carbon pool measurements can potentially constrain model predictions of carbon pools but need to be high frequency and sufficiently long term. Since carbon fluxes are proportional to carbon pool sizes in the `TECO v2.3` model, model simulations of carbon fluxes are also informed by carbon pool measurements.

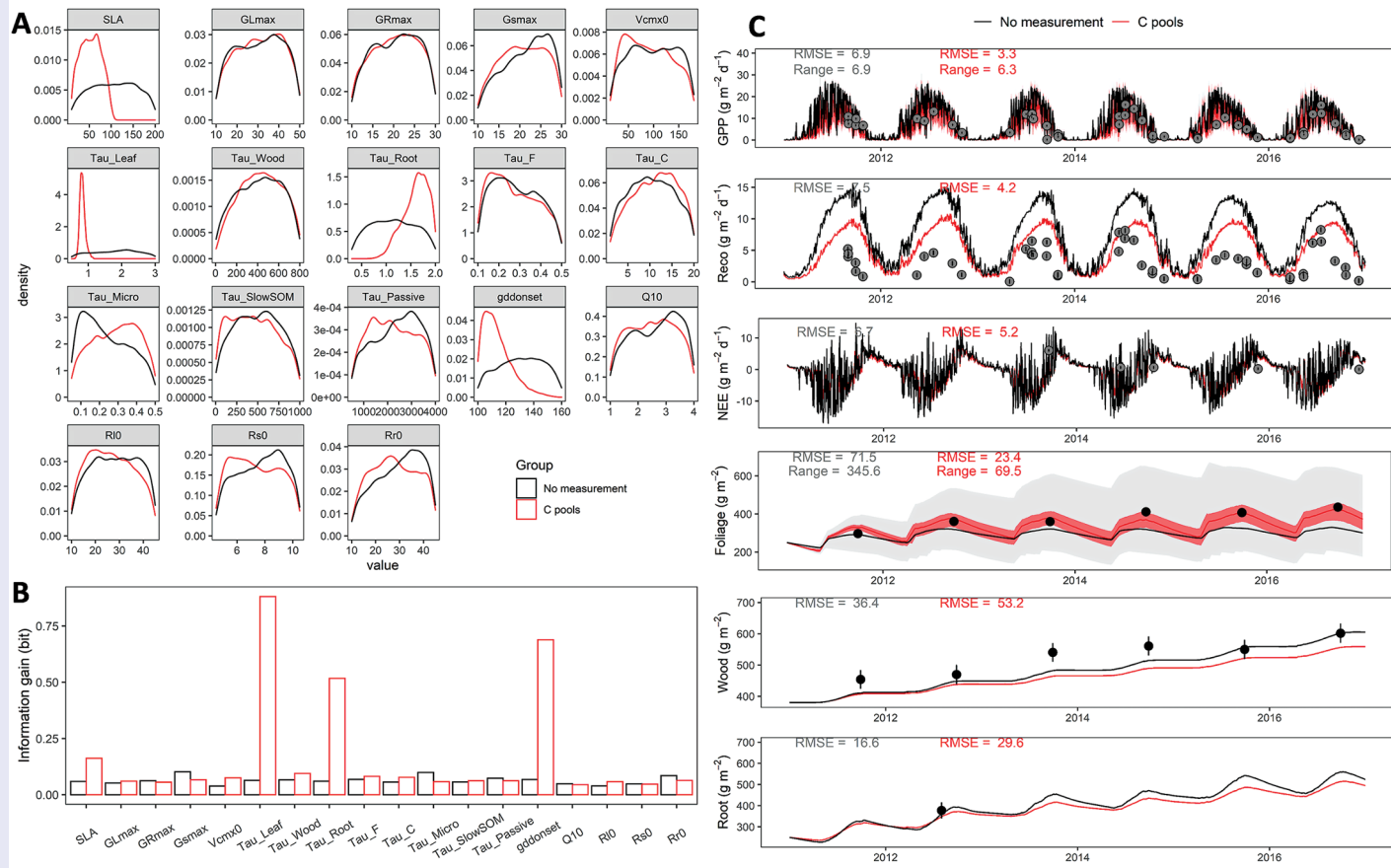


Figure 32.2. Comparison of parameters and simulations before and after assimilation of carbon pool measurements. (A) Posterior distribution of parameter values. (B) Information gain. (C) Simulations after assimilating carbon pool measurements. Full name and units of parameters can be seen in “Source_code/TECO_2.3/annotations in TECO model/parameter_file_annotations.jpg”. In (C), point and error bar (smaller than the point size) indicate the mean and the estimated standard deviation of measurements, respectively. The point and error bar are shown in gray if they were not assimilated into the model and are shown in black if they were assimilated into the model. Solid line, deep gray band, and light gray band indicate the mean, first to third quantiles, and range of 50 simulations, respectively. Simulation uncertainty is shown only for GPP and foliage biomass C. RMSE indicates root mean square error.

QUESTIONS:

1. Which measurement has more information to constrain parameters related to leaf carbon pool?

2. Why are parameters related to wood carbon pool and root carbon pool not informed by the carbon pool measurements used in this study?

EXERCISE 3: Model run with assimilation of carbon flux measurements

For Exercise 3, you select unit 8 and Exercise 3 in CarboTrain. Select the output directory and click Run Exercise. The data assimilation system will read carbon pool measurements in the file `input/SPRUCE_cflux.txt` and compare the measurements against the simulated carbon fluxes (Lines 7526-7549 in `TECO_2.3.f90`), in order to decide whether a set of parameter values should be accepted as posterior parameter values or not using the cost function. You can find model outputs in the folder `DA_with_cfluxes.zip`. There are three plots in the folder `DA_with_cfluxes/plot/DAUnit8`. These show comparisons between the results of the model when run without data assimilation (as in Exercise 1), with assimilation of carbon pool measurements (Exercise 2) and assimilation of carbon flux measurements. The figures compare posterior parameter values ("`3 Density_C_pools.png`", Figure 32.3A and B) and simulated carbon dynamics ("`3 Simulation_C_pools.png`" and "`3_2 Simulation_C_pools_uncertainty.png`", Figure 32.3C).

After assimilating the carbon flux measurements, several model parameters are well constrained (Figure 32.3A). The informed parameters are mainly related to photosynthesis, plant respiration, and autotrophic respiration, suggesting that carbon flux measurements are best at constraining model predictions of

carbon fluxes. Moreover, parameters related to fast-cycling carbon pools (e.g., leaf carbon pool) are more informed by the carbon flux measurements than parameters related to slow-cycling carbon pools (e.g., stem carbon pool and soil passive carbon pool) (Figure 32.3A and B). This is expected because of the short period (six years) of the carbon flux measurements. With a longer period of measurements, and a longer simulation, model parameters related to slow-cycling carbon pools may be informed.

After assimilating the carbon flux measurements into the TECO model, model biases and uncertainties in simulating NEE, GPP, and Reco are largely reduced (Figure 32.3C), suggesting that the carbon flux measurements can well inform model simulations of carbon fluxes. However, model biases and uncertainties in simulating carbon pools didn't decrease after assimilating the carbon flux measurements (Figure 32.3C). Overall, we see that a combination of parameter values adjusted according to carbon flux measurements can result in good simulations of carbon fluxes but have little information on simulating carbon pools.

QUESTIONS:

1. Which model parameters are well informed by carbon flux measurements?
2. Does yearly sum of carbon fluxes inform the model differently from daily sum of carbon fluxes?

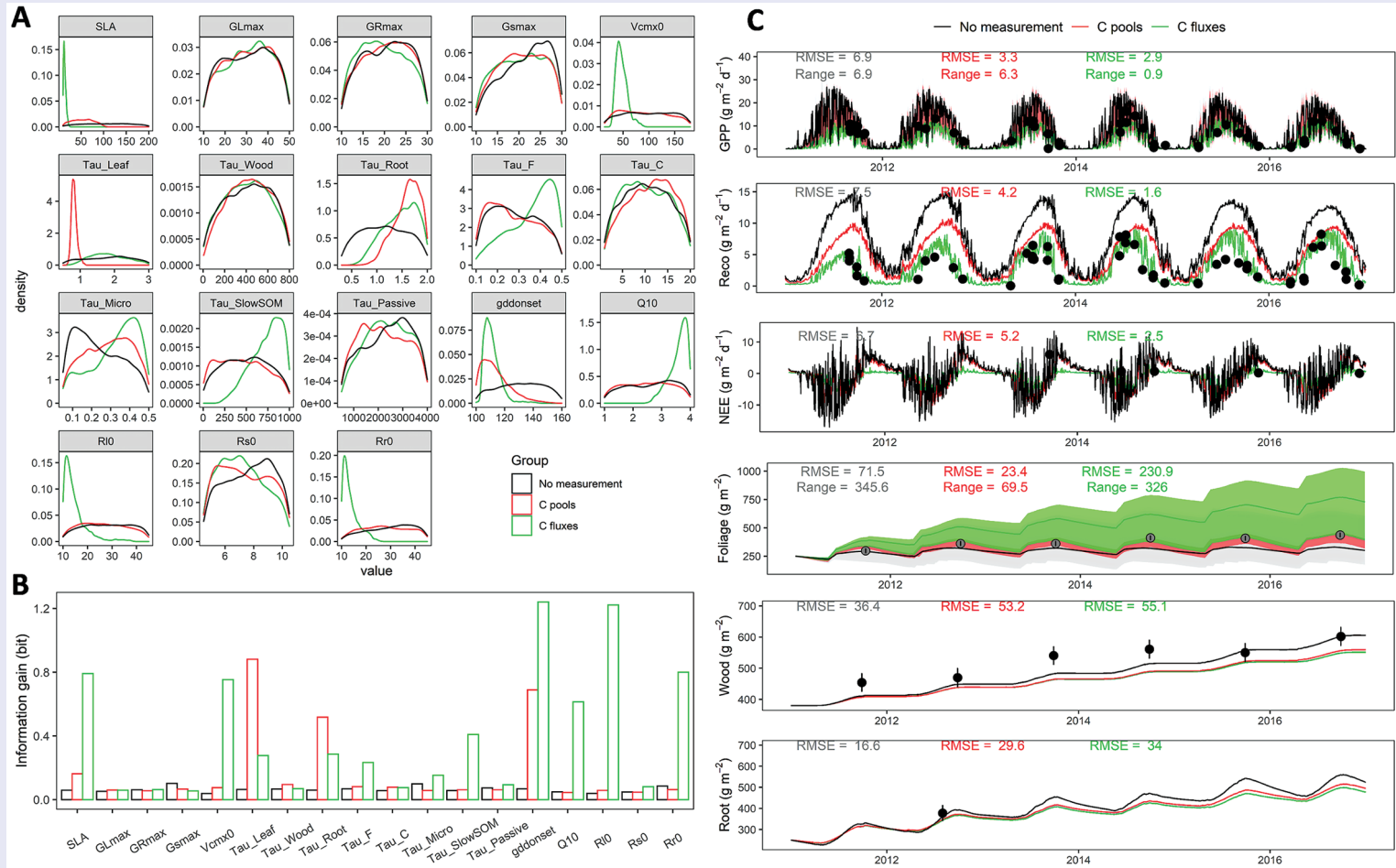


Figure 32.3. Comparison among model runs without data assimilation, and with separate assimilation of pool and flux measurements. (A) Posterior distribution of parameter values. (B) Information gain. (C) Simulations after assimilating carbon pool measurements. Full name and units of parameters can be seen in “Source_code/TECO_2.3/annotations in TECO_model/parameter_file_annotations.jpg”. In (C), black point and error bar (smaller than the point size) indicate the mean and the estimated standard deviation of measurements, respectively. Solid line, deep gray band, and light gray band indicate the mean, first to third quantiles, and range of 50 simulations, respectively. Simulation uncertainty is shown only for GPP and foliage biomass C. RMSE indicates root mean square error.

EXERCISE 4: Model run with assimilation of carbon pool and flux measurements

Based on the results of Exercises 2 and 3, you might expect the best constrained model parameters and predictions of carbon dynamics to be achieved by assimilating both pool and flux measurements into the model. We will explore whether that is the case in this exercise.

Select unit 8 and Exercise 4 in CarboTrain. Select the output directory and click *Run Exercise*. The data assimilation system will read both carbon pool and carbon flux measurements in the folder *input* and compare the measurements against the simulated carbon fluxes. You can find model outputs in the folder *DA_with_cpools_and_cfluxes.zip*. There are three plots in the folder *DA_with_cpools_and_cfluxes/plot/DAUnit8*. These show comparisons between the results of the model when run without data assimilation (as in Exercise 1), with assimilation of carbon pool measurements (Exercise 2), assimilation of carbon flux measurements (Exercise 3) and assimilation of both pool and flux measurements. The figures are a comparison of posterior parameter values (“4_Density_all.png”, Figure 32.4A and B) and simulated carbon dynamics (“4_Simulation_all.png” and “4_3_Simulation_C_pools_and_fluxes_uncertainty.png”, Figure 32.4C).

As expected, assimilating both pool and flux measurements constrains model parameters and predictions more than assimilating no measurements or pools or fluxes separately (Figure 32.4A and B). Optimized parameter values can differ between model runs when different datasets are assimilated (Figure 32.4A). For example, optimized values of specific leaf area are different between data assimilation with carbon pool measurements and data assimilation with carbon flux measurements. Similarly, optimized values of leaf turnover time are different depending which of these datasets is assimilated into the model. These results suggest that we should be cautious when interpreting optimized parameter values obtained by data assimilation.

Several model parameters, for example, maximum growth rates of leaf, stem, and root, and turnover time of soil passive C pool, are

not constrained by any available measurements (Figure 32.4A and B). This result suggests that additional measurements are needed to constrain the model. For example, measurements of stem growth in the field may be collected to constrain the maximum growth rate of stem. Soil carbon pool size and carbon age indicated by ^{14}C signature may be needed to constrain the turnover time of soil passive carbon pool, which is typically hundreds to thousands of years. Besides collecting more measurements, model simulations can be improved by simplifying the representation of model processes that cannot be constrained by available measurements. For example, litter carbon pool is separated into fine litter carbon pool and coarse litter carbon pool in the TECO model. However, we do not have any litter measurements to constrain the turnover times of the two litter pools in this case. We might consider combining the fine and coarse litter pool to simplify the TECO model.

After assimilating carbon pool and flux measurements into the TECO model, model simulations of carbon dynamics are generally less biased than assimilating only carbon pool or flux measurements (Figure 32.4C). However, reductions of model biases by carbon pool measurements and carbon flux measurements are not additive. For example, the reduction of model bias in simulating GPP with both carbon flux and carbon pool measurements is comparable to the reduction of model bias in simulating GPP with carbon flux measurements alone, although the carbon pool measurements also have some information on the simulation of GPP (Figure 32.4C). A combination of carbon pool and flux measurements helps reduce model bias in simulating NEE more than carbon pool measurements or carbon flux measurements alone, but the reductions are again not additive (Figure 32.4C). Finally, as expected, assimilating carbon pool and flux measurements into the model reduces model uncertainties in simulating both carbon pools and carbon fluxes (Figure 32.4C).

QUESTION:

1. Why are the information contents of pool measurements and flux measurements not additive?

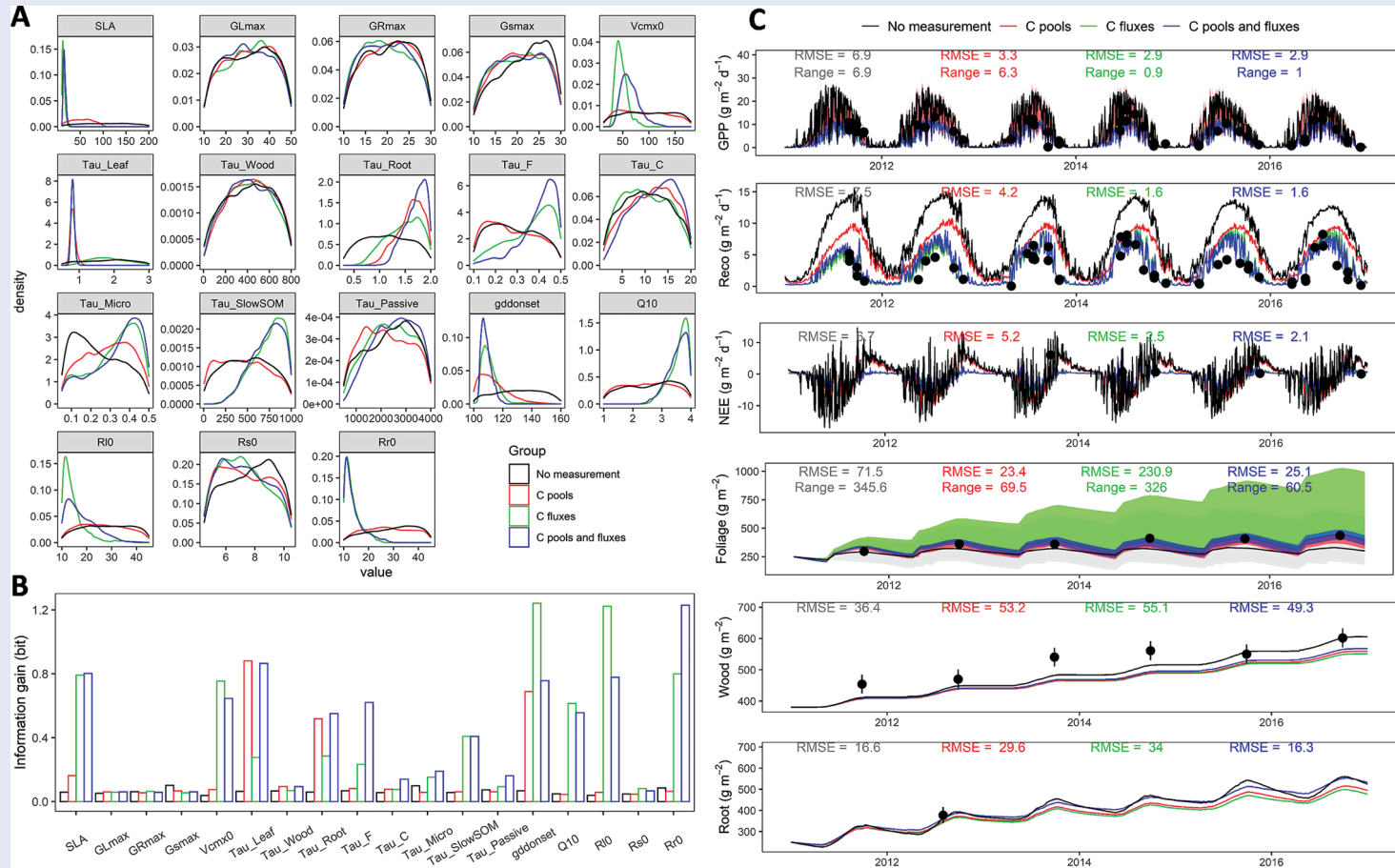


Figure 32.4. Comparison among model runs without data assimilation and with assimilation of pool and/or flux measurements. (A) Posterior distribution of parameter values. (B) Information gain. (C) Simulations after assimilating carbon pool measurements. Full name and units of parameters can be seen in “source_code/TECO_2.3/annotations in TECO model/parameter_file_annotations.jpg”. In (C), black point and error bar (smaller than the point size) indicate the mean and the estimated standard deviation of measurements, respectively. Solid line, deep gray band, and light gray band indicate the mean, first to third quantiles, and range of 50 simulations, respectively. Simulation uncertainty is shown only for GPP and foliage biomass carbon. RMSE indicates root mean square error.

EXERCISE 5: Real run of data assimilation

Advanced readers may run data assimilation on their own computer and obtain the results shown in Figures 32.1–32.4. To obtain outputs corresponding to Exercise 1, you select unit 8 and Exercise 5, set the output directory, and then you click *Set Namelist* and edit the namelist file by adjusting settings as follows:

```
do_co2_da = True
```

```
use_cflux_ob = 0
```

```
use_cpool_ob = 0
```

Click *Run Exercise* to run the model. Be aware that, depending on the specifications of your computer, it usually takes 8 hours or more to finish the model run.

To obtain model output corresponding to Exercises 2, 3, and 4, edit the namelist file as follows:

Exercise 2	Exercise 3	Exercise 4
do_co2_da = True	do_co2_da = True	do_co2_da = True
use_cflux_ob = 0	use_cflux_ob = 1	use_cflux_ob = 1
use_cpool_ob = 1	use_cpool_ob = 0	use_cpool_ob = 1

When all model outputs are available, you can visualize the results using the R script “*Source_code/TECO_2.3/R code for DA Unit 8.R*” (modifications of work directory and file names may be needed).

Beside the above exercises, you can also run data assimilation by changing initial parameter values (click *Set Initial parameters* in CarboTrain), selecting parameters for data assimilation and their ranges (click *select DA pars*). You can also assimilate methane, water,

and temperature measurements into the model (click *Set Namelist*). Finally, if you want to assimilate measurements at your own site into TECO v2.3, you will need to prepare climate forcing and your measurements in the format of files in *Training_Course/Source_code/TECO_2.3/input*. If you are familiar with Fortran, you may also modify the model source code in *TECO_2.3.f90* which can be found in the folder *Training_Course/Source_code/TECO_2.3*.

SUMMARY

We have seen that different types of measurements can constrain different model parameters and model simulations of different carbon processes. The model run using default parameter values and ranges gives information about the model. Assimilating carbon pool measurements can potentially constrain carbon turnover times and model predictions of carbon pool sizes. Assimilating carbon pool measurements may also constrain model predictions of carbon fluxes if carbon fluxes are modeled to be proportional to carbon pool sizes. Assimilating carbon flux measurements can constrain model parameters related to carbon fluxes and model predictions of carbon fluxes but may

not be able to constrain model predictions of carbon pools. Assimilating both carbon pool and flux measurements can generally constrain model parameters and predictions more than assimilating either group of measurements alone, but the information contents of carbon pool and flux measurements may overlap with each other. Assimilating carbon measurements into a model can reduce not only model uncertainties but also model biases in simulating carbon dynamics during forward runs. Overall, data assimilation technique is a tool to extract information from measurements. It can guide the collection of measurements by experimentalists and be used by modelers to constrain model predictions of land carbon dynamics.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

UNIT NINE

Ecological Forecasting with EcoPAD



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER THIRTY-THREE

Introduction to Ecological Forecasting

Yiqi Luo

Cornell University, Ithaca, USA

CONTENTS

Introduction / 287
Weather Forecasting / 287
Models and Predictability of the Terrestrial Carbon Cycle / 288
Data Availability to Constrain Forecast Via Data Assimilation / 290
Workflow System to Facilitate Ecological Forecasting / 290
Suggested Reading / 292
Quizzes / 292

In this rapidly changing world, improving the capacity to forecast future dynamics of ecological systems and their services is essential for better stewardship of the earth system. This chapter introduces ecological forecasting, the next frontier of research in ecology. Using weather forecasting as an analog, this chapter discusses four elements for ecological forecasting. The four elements are: predictability of the land carbon cycle; observations to constrain forecasting; data assimilation to integrate data with models; and a workflow system to automate ecological forecasting. This chapter also describes applications of an ecological forecasting system to a warming and CO₂ experiment in northern Minnesota and a precipitation mean and variance experiment in New Mexico.

INTRODUCTION

As you have seen in Figure 21.1, to realistically forecast ecosystem responses to environmental change, we need three elements: (1) model structure to represent the real-world processes that control system functions; (2) parameterization to reflect system properties; and (3) external forcing variables that an ecosystem experiences. We have studied the matrix approach to process-based modeling for the carbon

cycle in units 1–5. This chapter will examine the predictability of the land carbon cycle, according to the matrix equation, to understand the expectation of how well carbon forecasting can be achieved. We have also studied data assimilation for parameter estimation in units 6–8. This chapter will explore the availability of observations to achieve accuracy in ecological forecasting with different levels of model complexity. Training in this unit (i.e., this chapter, Chapter 34, and practice 9 in Chapter 35) is focused on a workflow system, Ecological Platform of Assimilating Data (EcoPAD) into model, to link real-time forcing and automate ecological forecasting.

WEATHER FORECASTING

Before we get into ecological forecasting, let us learn something from weather forecasting. Probably everyone is very familiar with weather forecasting. First, please take a moment to answer a few multiple-choice questions. How frequently do you look at weather forecasting? A. never; B. once every a few days; C. once a day; and D. a few times a day. You may make your own choice. Why do you look at weather forecasting? A. deciding what clothes to wear; B. deciding what kinds of outdoor activities to do; C. deciding whether you will do some field

research; or D. doing something else. What do you think the benefits are weather forecasting brings to society? A. saves lives; B. supports emergent management; C. mitigates the impact of and prevents economic losses from high-impact weather; D. facilitates financial revenue in energy, agriculture, transport, and recreational sectors; or E. all the above. These questions show that weather forecasting has become part of our lives, influences our daily activities, and is relevant to many aspects of our society.

According to a review paper by Peter Bauer et al. (2015) published in *Nature*, weather forecasting skills have been steadily improving. The skill reaches 98% by 2014 for a three-day weather forecast and about 60% for a seven-day weather forecast. I have personal experience on the accuracy of the weather forecast. Many of you may also notice how accurate the weather forecast has become.

Numeric weather prediction as a scientific discipline has been developing for more than one hundred years. The major milestones of weather prediction include knowing the laws of physics to make weather forecasting possible in 1901, developing and using super-computing in the 1970s, using satellite and other observations in the 1980s and using data assimilation in the 1990s. For example, it is relatively well known that the physical processes that determine weather dynamics include energy and water fluxes, momentum dynamics, and land surface conditions, among others.

Numeric weather prediction uses extensive observations from radar and other observations in data assimilation to generate weather patterns. Observations are used to constrain initial values every few hours. The data assimilation methods include 3D-var, 4D-var, and nowadays ensemble Kalman Filter. Data assimilation with complex weather models is computationally expensive. Accuracy and resolution of numerical weather prediction models increase over time as computational power exponentially increases. In short, success in weather forecasting depends on understanding of physical laws to develop models, collecting satellite and other data, using data assimilation to constrain initial values every a few hours, and relying on supercomputing to carry out calculation of the numeric models.

Similar to weather forecasting, ecological forecasting also needs process-based models, observations, data assimilation, and supercomputing. The process-based models offer model structure whereas observations are assimilated into

the process-based models for parameterization through data assimilation via supercomputing.

MODELS AND PREDICTABILITY OF THE TERRESTRIAL CARBON CYCLE

Process-based models of the terrestrial carbon cycle have different levels of complexity but are examined in units 1–5 for their general properties through the matrix approach. One of the key properties of terrestrial carbon dynamics is the convergence toward equilibrium over time, even if external forcing and disturbances often push the carbon cycle to be in disequilibrium. Using this intrinsic property, Luo et al. (2015) examined the predictability of the terrestrial carbon cycle. While the rate of approach to equilibrium, and equilibrium itself, is relatively predictable given knowledge about carbon input rates, loss rates, the initial conditions, and governing environmental constraints, there are three levels of predictability: high, medium, and low, plus two cases: less known and unknown about the predictability for individual processes (see Table 33.1).

For example, some external variables exhibit cyclic changes, typically causing the carbon flux rates, such as photosynthesis and respiration, to vary with the same period as the forcing (Table 33.1). The responses of terrestrial carbon to daily and seasonal cyclic forcing should be highly predictable. However, interannual variability in the terrestrial carbon cycle, as reflected in eddy-flux measurement and variations in the growth rate of atmospheric CO_2 , is less known for its underpinning mechanisms, making it difficult at present to evaluate its predictability.

Disturbance events, such as wildfire and climate extremes themselves, however, have an inherent random component (e.g., chances of a hurricane) making the predictability of individual events relatively low. Likewise, the severity of disturbance impacts on the carbon cycle is not very predictable, either. The recovery dynamics following a disturbance, however, appear to be highly predictable given adequate knowledge of the carbon influx rates, the residence times, and the pool sizes following disturbance (Table 33.1). Moreover, there is evidence that some ecosystems may recover to an alternative steady state following disturbance. Our lack of understanding of why this occurs limits our assessment of its consequences for carbon cycle predictability.

Most of the direct effects of climate changes on the terrestrial carbon cycle can be predicted via relatively simple response functions in ESMs.

TABLE 33.1

Intrinsic predictability of response patterns of the terrestrial carbon cycle to five classes of external forcing. The predictability of the carbon cycle measures a degree to which the response pattern is predictable given one class of external forcing. The predictability is usually judged by the sensitivity (e.g., diverging vs. converging) of systems behavior in response to various classes of perturbation and external forcing. In general, carbon cycle responses per se are more predictable than external forcing, which causes much high uncertainty in predicting carbon cycle responses to climate change

External forcing		Response of the terrestrial carbon cycle		
Class	Example	General pattern	Component	Intrinsic predictability
Cyclic environment	Diurnal, seasonal, and interannual	Cyclic	Diurnal and seasonal	High
			Interannual	Less known
Disturbance event	Fire, land use, insect outbreak, and storms etc.	Pulse-recovery	Time of events happening	Low
			Immediate impacts of disturbance events on carbon cycle	Medium
			Recovery	High
			Recovery to original or new equilibrium	Less known
Climate change	Rising [CO ₂] _a , climate warming, altered precipitation	Gradual	Direct impacts	High
			Indirect impacts via induced changes in disturbance regimes and ecosystem states	Less known
Shifts in Disturbance regimes	Regional, long-term patterns of fire, land use, insect outbreak, and storm etc.	Disequilibrium	Joint probability to describe disturbance regimes and their shifts	Unknown
			Impacts of shifted disturbance regimes on mean carbon storage	High
Ecosystem state change	Forest to cropland, grassland to cropland, reforestation, etc.	Abrupt changes	When and where ecosystem states change	Less known
			Carbon cycle change with ecosystem states	High

From Luo et al. (2015)

However, climate change also causes indirect effects on the terrestrial carbon cycle, such as changes in plant species composition, microbial priming, and respiratory acclimation. The indirect effects are much less well understood, making it currently unclear just how predictable they are (Table 33.1). Moreover, climate change may also induce shifts in disturbance regimes and changes in ecosystem states, which is less predictable as discussed below.

Disturbance regimes can be quantified by joint probabilistic distributions of disturbance frequency and severity, which, in turn, can be used to generate a probability distribution of ecosystem carbon storage. The mean of the probability distribution determines the realizable carbon storage capacity under a given regime, reflecting the mean carbon storage capacity over a sufficiently long time period or over a sufficiently large area. This means carbon storage capacity could thus be predictable. However, we do not have enough knowledge to predict when the disturbance regime changes by direct or indirect anthropogenic forcing.

When ecosystem states change, rates of carbon cycling among the plant, litter, and soil carbon pools also change. Given the change in vegetation structures and corresponding parameters, a consequent change in the carbon cycle is quantifiable. However, while vegetation state changes have been studied, their relationships with those carbon cycle parameters remain poorly understood.

Overall, many processes of the terrestrial carbon cycle are intrinsically predictable. For these processes, forecasting is expected to be highly achievable. However, the indirect effects of climate change on terrestrial carbon cycling become less predictable, especially those which, via changes in species composition and disturbance regimes, lead to ecosystem state changes. In addition, individual disturbance events usually occur stochastically even within a stationary disturbance regime. For such processes, forecasting is expected to be more uncertain.

DATA AVAILABILITY TO CONSTRAIN FORECAST VIA DATA ASSIMILATION

There are plenty of data available to support forecasting the carbon cycle. For example, eddy-flux networks provide half-hourly data streams over hundreds of sites to support near-term forecasting of carbon cycle dynamics over daily, seasonal, and interannual time scales.

Data is also available on long-term processes, such as disturbance events and subsequent recovery. Forecasting disturbance events themselves is not easy at this stage. However, data is available to test forecasting of recovery processes. As disturbance regimes and their impacts on carbon cycle take place over quite long time scales, it is clear how real or near-term forecasting would be useful to research or management. But it is feasible to use available data to test the capability of forecasting. Plenty of data are also available on ecosystem state changes for studying ecological forecasting.

There are many global change experiments ongoing right now. They offer great opportunities to test forecasting capability at some of the experimental sites. We have been forecasting responses of the carbon cycle to experimental treatments at five levels of temperature and two levels of atmosphere CO₂ concentration at the SPRUCE site since 2016 (See Chapter 25). We are also setting up the forecasting system for a drought experiment at Sevilleta Long Term Ecological Research (LTER) site in New Mexico.

Data from observational networks and experiments are integrated into models via data assimilation before ecological forecasting is made. Units 6–8 describe basic concepts, procedure, and application cases of data assimilation.

WORKFLOW SYSTEM TO FACILITATE ECOLOGICAL FORECASTING

Ecological forecasting is usually carried out in an automatic fashion. We have developed a workflow system for near-time forecasting. The system is Ecological Platform for Assimilating Data (EcoPAD) into models (Huang et al. 2019) (also see Chapter 34). EcoPAD is a software system that links sensor networks to ecological forecasting. It integrates eco-informatics, web-technology, ecological models, data assimilation techniques, and visualization. EcoPAD was designed to promote interactions among modeling, experimentation and observations to gain the best science.

Data and models are integrated through a data assimilation system before the trained models are used for forecasting, optimization of measurement plans, and uncertainty analysis. The data to be integrated can come from the real-time sensor networks or from spreadsheets with records of hand measurement (Figure 33.1). EcoPAD offers three

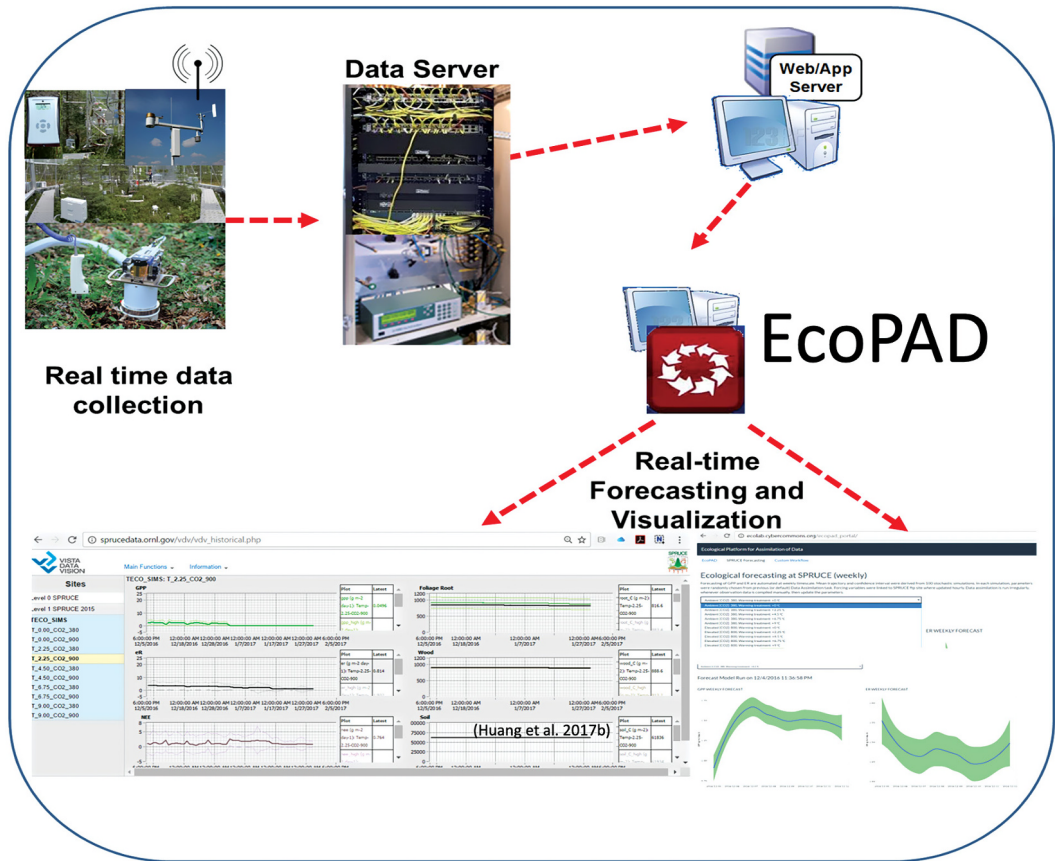


Figure 33.1. EcoPAD to streamline data ingestion from sensors and servers, model simulation, data assimilation, forecasting, and visualization. By timely updating of the parameter values, EcoPAD enables close interactions between experimenters and the modelers.

options, which are simulation, data assimilation and forecast, in response to a request from users. Once a user makes a request, EcoPAD can execute the task automatically to generate results. The generated results can be visualized in real-time as well. Thus, EcoPAD is an interactive software system for researchers to automatically execute model simulation, data assimilation, and ecological forecasting in real- or near-time.

We have applied EcoPAD to the SPRUCE experimental site located in Northern Minnesota (see Chapter 25). SPRUCE is a whole ecosystem warming and CO₂ enrichment project. It has five levels of temperature treatments and two levels of CO₂ concentrations. The experiment follows a gradient design with five chambers for five levels of temperature treatments at ambient CO₂ concentration and five chambers at elevated CO₂ concentration. The project is very well equipped

with lots of real-time sensors and involves more than 100 scientists who perform many kinds of measurements. The real-time sensors send data to data servers. Data servers also store the data from hand measurements. EcoPAD automatically ingests data from data servers through a web app server for data assimilation and forecasting. The forecast results are automatically sent to two sites, one at the SPRUCE site and one at our lab website, for visualization. EcoPAD has done ecological forecasting automatically at midnight on Saturday every week since June 2016.

The forecasting variables include snow cover, soil thermal dynamics, and frozen depth (Huang et al. 2017); many carbon cycle variables, such as gross primary production, net primary production, net ecosystem production, and ecosystem respiration (Jiang et al. 2018); and methane flux and pathways (Ma et al. 2017).

We are applying EcoPAD for data assimilation and ecologist forecasting at Sevilleta Long-Term Ecological Research (LTER) site in New Mexico. Sevilleta LTER site currently has a few experiments going on. These experiments are mainly related to precipitation and nitrogen fertilization. In addition, there are several long-term eddy-flux towers, measuring carbon, water, and energy fluxes for years. We are developing the capability to do real-time or near-time data assimilation and ecological forecasting at those experimental and eddy-flux sites.

In fact, EcoPAD can be used as a smart experiment-modeling system. First, the system can predict what ecosystems may respond to treatments once you have selected a site and decided your experimental plan. When we were writing a proposal to continue the LTER study at Sevilleta, New Mexico, we used the TECO model to do pre-experiment analysis on possible ecosystem responses to increasing variability in precipitation. The modeling results were included in the proposal. Once you get funding to do the experiment, you can use EcoPAD to assimilate the data you are collecting to constrain model forecasts. The model forecasts what ecosystem responses may likely be for the remaining period of your experiment. The forecast ecosystem responses can be used as references for you to design your measurement plan. At the SPRUCE project, Shuang Ma's forecast results stimulated discussion about how much methane may be released through bubbling. Discussion on this issue lasted for a few weeks. The extensive discussion led to improvement of Shuang's methane model, new ideas to design additional measurements of methane concentrations along the soil profiles, and more collaborations between experimental and modeling teams. Moreover, the uncertainty analysis with EcoPAD can tell us what those important datasets are and which need more measurement in order to understand the system dynamics. During the course of our study, we can use EcoPAD to periodically update the forecast by repeating steps from data assimilation to forecasting to improvement of measurement and model. We are using EcoPAD to do weekly forecasts at the SPRUCE project. During this process, we improve the models, the experiments, and the data assimilation system.

In summary, carbon cycle forecasting is a new frontier of research in ecology. As many processes are highly predictable, forecasting carbon cycle dynamics is expected to be highly achievable. Plenty of data are available to constrain forecasting with data assimilation. Workflow systems to automate data-model integration and ecological forecasting are becoming available. The challenge remains to identify societally relevant issues that carbon cycle forecasting can address.

SUGGESTED READING

Jiang Jiang, Yuanyuan Huang, Shuang Ma, Mark Stacy, Zheng Shi, Daniel M. Ricciuto, Paul J. Hanson, Yiqi Luo. 2018. Forecasting responses of a northern peatland carbon cycle to elevated CO₂ and a gradient of experimental warming. *Journal of Geophysical Research: Biogeosciences*, **123**: 1057–1071.

QUIZZES

1. What are the elements leading to the success in weather forecasting?
2. We need forcing variables to be consistent between models and field sites for realistic forecast because
 - a. the model needs forcing variables to drive simulations
 - b. it is easy to get forcing variables from field sites
 - c. environment variables that drive the model prediction have to represent what ecosystems experience
 - d. we can measure environmental variables at field sites
3. What is a workflow system?
4. Pre-experiment modeling analysis is useful because
 - a. it gives some general ideas on what ecosystem responses may look like
 - b. it will give us the precise prediction of ecosystem responses
 - c. we do not have to carry out the experiment anymore
 - d. it helps us adjust the measurement plan

Ecological Platform for Assimilating Data (EcoPAD) for Ecological Forecasting

Yuanyuan Huang

Climate Science Centre, CSIRO, Canberra, Australia

CONTENTS

Why Do We Need EcoPAD? / 293
General Structure of EcoPAD / 294
Applications of EcoPAD: The Example of SPRUCE / 298
Suggested Reading / 300
Quizzes / 300

Tremendous scientific endeavors in ecology have been driven by the goal of forecasting future ecological dynamics. This chapter introduces the web-based Ecological Platform for Assimilating Data into model (EcoPAD) to facilitate ecological forecasting. The objectives of this lecture are to understand why we need a platform like EcoPAD, the structure of the platform, and how to use EcoPAD to facilitate ecological forecasting.

WHY DO WE NEED ECOPAD?

Ecological research is driven by the quest of understanding the dynamics of biota and their interactions with the environment. To understand ecological patterns, processes and functions, we conduct field or manipulative experiments. From these experiments, we combine a wide range of approaches to obtain relevant observational data (e.g., through real-time sensor, laboratory measurements, remote-sensing and video-based observations). For example, the National Ecological Observatory Network (NEON) of the United States monitors ecosystems across the United States by collecting hundreds of data products including organismal counts and measurements, water and soil quality, energy fluxes, and remotely sensed vegetation indices. Many similar observational

networks now provide us with a wide range of ecological relevant datasets for different regions. To name a few, FLUXNET tracks the exchanges of carbon dioxide, water vapor, and energy between the biosphere and atmosphere for a network of flux tower sites around the world, while DroughtNet focuses on the responses of terrestrial ecosystems to drought. Knowledge obtained from data and experiments enables us to make inferences about ecological dynamics under novel situations. The inference could be based on complex mathematical models built upon data as well as simple relationships derived from data. We call inference under novel conditions prediction. Forecasting is a type of prediction in which we make predictions about the future.

Ecological forecasting is not only valuable for contributing to scientific advances but is also practically valuable in guiding resource management and decision-making towards a sustainable future. The practical need for ecological forecasting is particularly urgent for our current rapidly changing world, which is experiencing unprecedented food insecurity, natural resource depletion, biodiversity loss, climate changes, and pollution of air, waters, and soils. This practical need has brought a growing number of forecasting-oriented studies, for example, on fisheries, crop yield, species

dynamics, algal blooms, phenology, pollinator performance and biodiversity. Ecological forecasting is valuable in almost every subdiscipline of ecology. Recent progress especially in available data, improved process understanding, data assimilation techniques, and advanced cyber-infrastructure, is converging to transform ecological research into advanced quantitative forecasting. In practice, however, ecological forecasting remains largely aspirational, with the number of forecasting studies lagging behind demand. One bottleneck is the lack of infrastructure to enable timely integration of data into models. EcoPAD is designed as a solution to widen this bottleneck. It provides a fully interactive infrastructure to facilitate ecological forecasting, especially near-time ecological forecasting based on iterative data–model integration.

EcoPAD (https://ecolab.nau.edu/ecopad_portal/, last access: November 2020) serves to link ecological experiments and data with models and provides easily accessible and reproducible data-model integration with interactive web-based simulation, data assimilation, and forecasting. The system is designed to streamline web request-response, data management, modeling, prediction, and visualization to boost the overall throughput of observational data, promote data-model integration, facilitate communication between modelers and experimenters, inform ecological forecasting, and improve scientific understanding of ecological processes. EcoPAD facilitates estimation of model parameter values, evaluation of model structure, assessment of information content of datasets, and understanding of uncertainties revealed by model-data fusion exercises. Additionally, EcoPAD automates data management, model simulation, data assimilation, ecological forecasting, and result visualization. It provides an open, convenient, transparent, flexible, scalable, traceable, and readily portable platform to systematically conduct data-model integration towards better ecological forecasting. The automated near-time ecological forecasting through EcoPAD updates periodically in a manner similar to weather forecasting. This design of EcoPAD enables it to function as a smart interactive model-experiment (ModEx) system (Figure 34.1). ModEx forms a feedback loop in which field experiments guide modeling and modeling influences experimental focus. Information is constantly fed back between modelers and experimentalists, and simultaneous efforts from both parties advance and shape understanding towards

better forecasts. ModEx can: (1) predict what an ecosystem's response might be to treatments once the experimenter has selected a site and decided the experimental plan; (2) assimilate the data collected during the experiment to constrain model predictions; (3) project the expected ecosystem responses in the rest of the experiment; (4) tell experimenters which priority datasets to collect in order to better understand the system; (5) periodically update the projections; and (6) improve the models, the data assimilation system, and field experiments during the process.

In addition to forecasting and facilitating interaction between modeling and experimental communities, EcoPAD is desirable because of the potential service it can bring to society. Forecasting with carefully quantified uncertainty is helpful in providing support for natural resource managers and policy makers. It is always difficult to bring complex mathematical ecosystem models to end users who do not have training in modeling, which creates a gap between current scientific advances and public awareness. The web-based interface of EcoPAD makes modeling as easy as possible without losing the connection to the mathematics, knowledge and data behind the models. In this way, infrastructure like EcoPAD has the potential to transform environmental education and encourage citizen science in ecology and climate change.

GENERAL STRUCTURE OF ECOPAD

The essential components brought together by EcoPAD include experiments and data, ecological models, data assimilation techniques, and the scientific workflow (Figure 34.1).

Data are the foundation of ecological modeling and forecasting. We have entered the “big data” age, characterized by the ready availability of different, often extensive datasets across various temporal-spatial scales. These datasets might have high temporal resolution, such as time series from real-time ecological sensors, or extensive spatial coverage from remote sensing sources and data stored in geographic information systems. Data may contain information related to environmental forcing (e.g., precipitation, temperature, incoming radiation), site characteristics (e.g., soil texture and species composition), or biogeochemistry of soils and waters. EcoPAD offers systematic data management to digest diverse data streams. Datasets in EcoPAD are derived from research

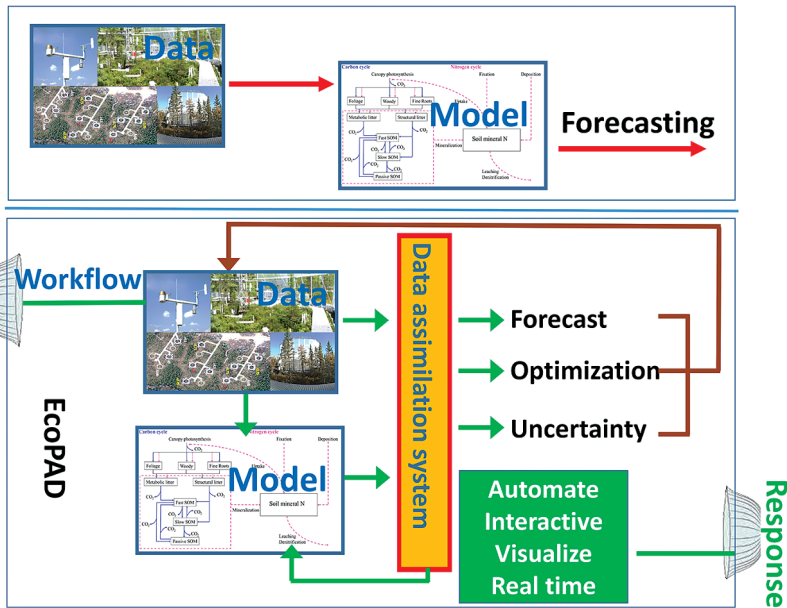


Figure 34.1. Schema of approaches to forecast future ecological responses under (a) current practice; and (b) the Ecological Platform for Assimilation of Data (EcoPAD). Current practice makes use of observations to develop and/or calibrate models to make predictions. EcoPAD goes further by linking models to data through a formalized, iterative cycle using a fully interactive platform. EcoPAD consists of four major components: experiment and data, model, data assimilation, and the scientific workflow (green arrows or lines). Data and model are iteratively integrated through its data assimilation systems to improve forecasting. Its near-real time forecasting results are shared among research groups through a web interface to guide new data collections. The scientific workflow enables web-based data transfer from sensors, model simulation, data assimilation, forecasting, result analysis, visualization, and reporting, encouraging user-model interactions, especially for experimentalists and end users having a limited background in modeling. Adapted from Huang et al. (2019).

projects in comma-separated value (csv) files or other loosely structured data formats. These datasets are first described and stored with appropriate metadata via either manual operation or scheduled automation from sensors. Data are generally separated into two groups. One comprises forcing variables to drive modeling, the other, observations used for data assimilation. Scheduled sensor data are appended to existing data files with prescribed frequency. Attention is given to how the particular dataset varies over space and time. When the spatio-temporal variability is understood, it is then placed in metadata records that allow for query through EcoPAD's scientific workflow.

The workflow and data assimilation system of EcoPAD are relatively independent of any specific ecological model. To illustrate the integration of models with EcoPAD, we take the Terrestrial ECOSystem (TECO) model as a general example. Linkages among the workflow, data assimilation system and ecological model are based on messaging. For example, the data assimilation system

generates parameters that are passed to ecological models. The state variables simulated from ecological models are passed back to the data assimilation system. Models may have different formulations. As long as these models take in the same parameters and simulate the same state variables, they are functionally identical from the point of view of the data assimilation system. TECO simulates ecosystem carbon, nitrogen, water, and energy dynamics. The original TECO model has four major submodules (canopy, soil water, vegetation dynamics, and soil carbon and nitrogen) (Weng and Luo, 2008) and is further extended to incorporate methane biogeochemistry and snow dynamics (Huang et al., 2017; Ma et al., 2017).

Data assimilation (Chapter 21) provides a framework to combine models with data to estimate model parameters, test alternative ecological hypotheses through different model structures, assess the information content of datasets, quantify uncertainties, derive emergent ecological relationships, identify model errors, and improve ecological

predictions. Under the Bayesian paradigm, data assimilation techniques treat the model structure, the initial and parameter values as priors that represent our current understanding of the system (Chapter 22). As new information from observations or data becomes available, model parameters and state variables can be updated accordingly. The posterior distributions of estimated parameters or state variables are imprinted with information from the model, observations (or data) as the chosen parameters are constrained to reduce mismatches between observations and model simulations. Future predictions benefit from such constrained posterior distributions through forward modeling. As a result, the probability density functions of predicted future states following data assimilation normally have narrower spreads than those without data assimilation. EcoPAD can accommodate different data assimilation techniques since the scientific workflow of EcoPAD is independent of the specific data assimilation algorithm. One example of a data assimilation method is the Markov chain Monte Carlo (MCMC) introduced in Chapter 22.

The scientific workflow of EcoPAD wraps around one or more user-specified ecological

models and data assimilation algorithms and acts to move datasets in and out of structured and cataloged data collections (metadata catalog), while leaving the logic of the ecological models and data assimilation algorithms untouched (Figure 34.2). When a user makes a request through the web browser or command line utilities, the scientific workflow takes charge of triggering and executing corresponding tasks, pulling data from a remote server, running a particular ecological model, automating forecasting, or making the result easily accessible and understandable to users through web based graphic displays (Figure 34.2). The workflow system is portable across operation system and programming language and is built to be scalable to meet the demands of the model and the end-user community. The essential components of the scientific workflow of EcoPAD include the metadata catalog, web application-programming interface (API), the asynchronous task or job queue (Celery), and the container-based virtualization platform (docker) (Figure 34.2). The workflow system of EcoPAD also provides structured result access and visualization. Scientific workflow is a relatively new concept in the ecology literature

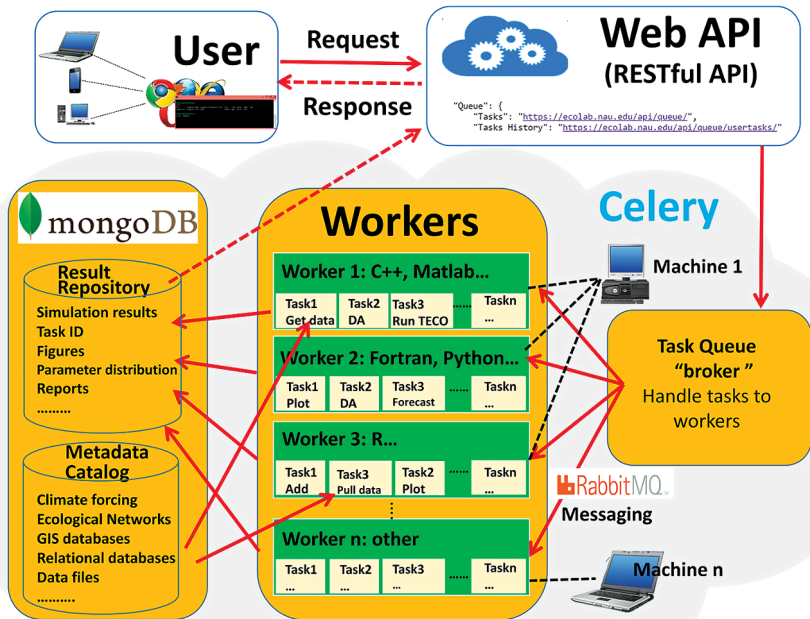


Figure 34.2. The scientific workflow of EcoPAD. The workflow wraps ecological models and data assimilation algorithms with the docker containerization platform. Users trigger different tasks through the representational state transfer (RESTful) application-programming interface (API). Tasks are managed through the asynchronous task queue, Celery. Tasks can be executed concurrently on a single or more worker servers across different scalable IT infrastructures. MongoDB is a database software that takes charge of data management in EcoPAD, and RabbitMQ is a message broker. Adapted from Huang et al. (2019).

but is essential to realize real or near-real time forecasting. Thus, we describe it in detail below. Readers who are not interested in technical details may skip the following paragraphs and jump to the section: “Applications of EcoPAD: The Example of SPRUCE”.

Datasets can be placed and queried in EcoPAD via a common metadata catalog, which allows for effective management of diverse data streams. The EcoPAD metadata scheme includes a description of the data product, access pattern, and time stamp of last metadata update. MongoDB (<https://www.mongodb.com/>, last access: November 2020), a NoSQL database technology, is employed to manage heterogeneous datasets to make documentation, query and storage fast and convenient. Through MongoDB, measured datasets can be easily fed into ecological models for various purposes such as to initialize the model, calibrate model parameters, evaluate model structure, and drive model forecasts. For datasets from real-time ecological sensors that are constantly updating, EcoPAD can be set to automatically fetch new data streams with adjustable frequency according to research needs.

The “gateway” of EcoPAD is the Representational State Transfer (RESTful) application programming interface (API). It can deliver data to a wide variety of applications and enables a wide array of user interfaces and data dissemination activities. Once a user makes a request, such as through clicking on relevant buttons from a web browser, the request is passed through the RESTful API to trigger specific tasks. Thus, the API bridges communication between the client (e.g., a web browser or command line terminal) and the server (Figure 34.2). The API exploits the HyperText Transfer Protocol (HTTP) such that data can be retrieved and ingested from EcoPAD through the use of simple HTTP headers and verbs (e.g., GET, PUT, POST, etc.). Since HTTP is also understood by web servers and clients, a user can incorporate summary data from EcoPAD into a website with a single line of html code. Users are able to access data directly through programming environments like R, Python, and MATLAB. Simplicity, ease of use, and interoperability are among the main advantages of the API, which enables web-based modeling.

The task queue is a mechanism used to distribute work across work units such as threads or machines. EcoPAD uses Celery (<https://github.com/celery/celery>, last access: November 2020) as an asynchronous task or job queue that runs in

the background (Figure 34.2). Celery communicates through messages, and EcoPAD takes advantage of RabbitMQ (<https://www.rabbitmq.com/>, last access: November 2020) to manage messaging. After the user submits a command, the request or message is passed to Celery via the RESTful API. These messages may trigger different tasks, which include but are not limited to pulling data from a remote server where original measurements are located, accessing data through a metadata catalog, running model simulations with user specified parameters, conducting data assimilation that recursively updates model parameters, forecasting future ecosystem status, and postprocessing model results for visualization. The broker inside Celery receives task messages and hands out tasks to available Celery “workers” that perform the actual tasks (Figure 34.2). Celery workers are in charge of receiving messages from the broker, executing tasks, and returning task results. The worker can be a local or remote computation resource (e.g., the cloud) that has connectivity to the metadata catalog. Workers can be distributed into different information technology infrastructures, which makes the EcoPAD workflow expandable in accommodating more computational resources. Each worker can perform different tasks depending on the tools installed in each worker. One task can also be distributed to different workers. In such a way, the EcoPAD workflow enables the parallelization and distributed computation of actual modeling tasks across various IT infrastructures and is flexible in implementing additional computational resources by connecting additional workers.

Another key feature that makes EcoPAD easily portable and scalable among different operation systems is the utilization of a container-based virtualization platform, the docker (<https://www.docker.com/>, last access: January 2019). The docker can run many applications that rely on different libraries and environments on a single kernel with its lightweight containerization. Tasks that execute TECO in different ways are wrapped inside different docker containers that can “talk” with each other. Each docker container embeds the ecosystem model into a complete file system that contains everything needed to run an ecosystem model: the source code, model input, run time, system tools, and libraries. Docker containers are both hardware-independent and platform-independent, and they are not confined to a particular language, framework, or packaging

system. Docker containers can be run from a laptop, workstation, virtual machine, or any cloud compute instance. This is done to support the widely varied number of ecological models running in various languages (e.g., MATLAB, Python, Fortran, C, and CCC) and environments. In addition to wrapping the ecosystem model into a docker container, software applied in the workflow, such as Celery, RabbitMQ, and MongoDB, are all lightweight and portable encapsulations through docker containers. Therefore, EcoPAD is readily portable to different environments.

EcoPAD enables structured result storage, access, and visualization to track and analyze data-model fusion practice. Upon the completion of the model task, the model wrapper code calls a postprocessing callback function. This callback function allows model-specific data requirements to be added to the model result repository. Each task is associated with a unique task ID and model results are stored within the local repository that can be queried by the unique task ID. The storage and query of model results are realized via the MongoDB and RESTful API (Figure 34.2). Researchers are able to review and download model results and parameters submitted for each model run through a web-accessible URL (link). The EcoPAD web page also displays a list of historical tasks (with URL) performed by each user. All current and historical model inputs and outputs are available to download, including the aggregated results produced for graphical web applications. In addition, EcoPAD also provides a task report that contains an all-inclusive recap of submitted parameters, task status, and model outputs with links to all data and graphical results for each task. Such structured result storage and access make sharing, tracking, and referring to modeling studies instantaneous and clear.

APPLICATIONS OF ECOPAD: THE EXAMPLE OF SPRUCE

The SPRUCE experiments and datasets were introduced in Chapter 25. Here, we demonstrate the use of the EcoPAD infrastructure as a way to assimilate multiple streams of data from the SPRUCE experiment to the TECO model using the MCMC algorithm and forecast ecosystem dynamics in both near time and for the next ten years. A similar example was presented in Chapter 26. The forecasting system for SPRUCE is available at: https://ecolab.nau.edu/ecopad_portal/

(last access: November 2020). From the web portal, users can check our current near- and long-term forecasting results, conduct model simulation, data assimilation, and forecasting runs, and analyze and visualize model results. We set up the system to automatically pull new data streams every Sunday from the SPRUCE FTP site that holds observational data and updates the forecasting results based on new data streams. Updated forecasting results for the following week are customized for the different manipulative treatments of the SPRUCE experiments and displayed in the EcoPAD-SPRUCE portal. At the same time, these results are sent back to SPRUCE communities and displayed together with near term observations for experimentalists to study.

In the SPRUCE project, we take advantage of this platform to stimulate interactive communication between modelers and experimentalists, study the acclimation of ecosystem carbon cycling to experimental manipulations, partition uncertainty sources in forecasting, improve the biophysical estimation for better forecasts, and explore how the updated model and data contribute to reliable forecasting. Our case studies confirm that realistic model structure, correct parameterization, and accurate external environmental conditions are critical for forecasting carbon dynamics. The following section describes how the updated model and data contribute to reliable forecasting in the SPRUCE experiment. For other applications, the interested reader is referred to Huang et al. (2019).

In the SPRUCE project, the system automatically conducts data assimilation with the new observational data streams from each week, successively improving the model parameterization. With constantly adjusted model and external forcing and weekly archived model parameters, model structure, external forcing, and forecasting results, the contribution of model and data updates to forecasting accuracy can be tracked by comparing the previous week's forecasted simulations to the current one informed by data from that week. Figure 34.3 illustrates how updated external forcing (compared to stochastically generated forcing) and shifts in ecosystem state variables shape the predictions. "Updated" means the real meteorological forcing monitored from the site's weather station. In the absence of observations, stochastically generated forcing is used as a proxy for future meteorological conditions. Future precipitation and air temperature are generated

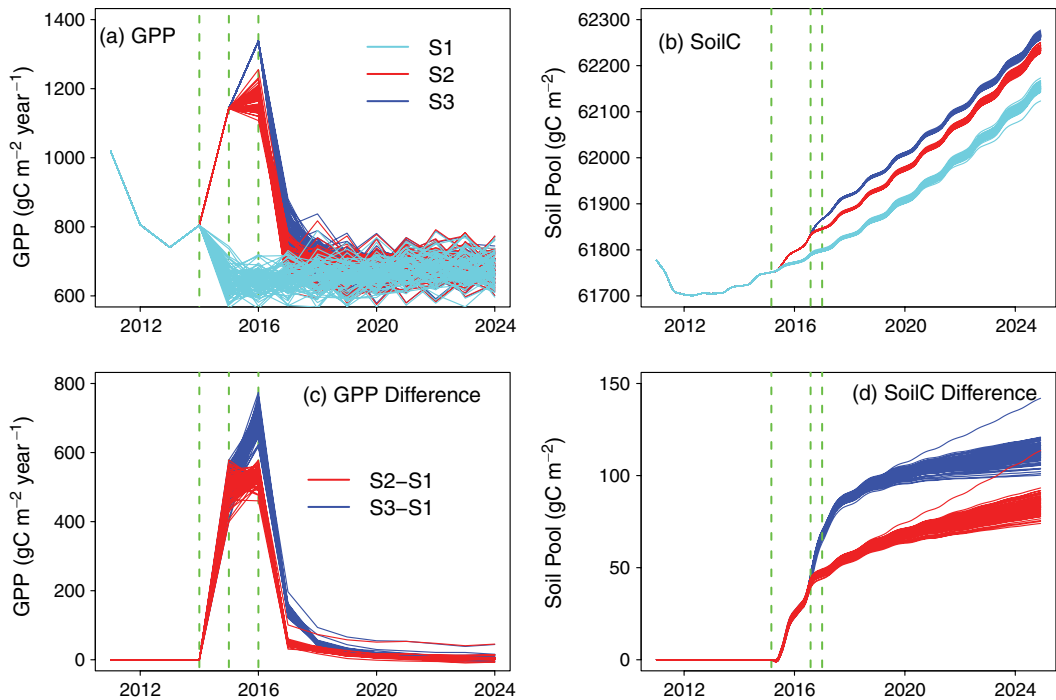


Figure 34.3. Updated v. original forecasting of gross primary production (GPP; panels a, c) and soil organic C content (SoilC; panels b, d). The upper panels show three series of forecasts with updated v. stochastically generated weather forcing. “Updated” = forced by actual meteorology from field weather stations. Cyan lines indicate forecasting with 100 stochastically generated weather forcing timeseries from January 2015 to December 2024 (S1); red lines correspond to forecasting updating with measured weather forcing from January 2015 to July 2016, followed by forecasting with 100 stochastically generated weather forcing timeseries from August 2016 to December 2024 (S2); blue lines show updated forecasting with measured weather forcing from January 2015 to December 2016, followed by forecasting with 100 stochastically generated weather forcing timeseries from January 2017 to December 2024 (S3). Panels (c) and (d) display mismatches between updated forecasting (S2, 3) and the original forecasting (S1). Red displays the difference between S2 and S1 (S2–S1), and blue shows the discrepancy between S3 and S1 (S3–S1). Dashed green lines indicate the start of forecasting with stochastically generated weather forcing. Note that panels (a) and (c) are plotted on a yearly timescale and panels (b) and (d) show results on a monthly timescale. Adapted from Huang et al. (2019).

by vector autoregression using a historical dataset (1961–2014) monitored by the weather station. Photosynthetically active radiation (PAR), relative humidity, and wind speed are randomly sampled from the joint frequency distribution at a given hour each month. Detailed information on weather forcing is available in Jiang et al. (2018). TECO is trained through data assimilation with observations from 2011–2014 and used to forecast GPP and total soil organic carbon content at the beginning of 2015.

For demonstration purposes, Figure 34.3 shows three series of forecasting results instead of updates from every week. Series 1 (S1) records forecasted gross primary production (GPP) and soil carbon with stochastically generated weather forcing from January 2015–December 2024 (Figure 34.3a, b,

cyan). Series 2 (S2) records simulated GPP and soil carbon with observed (updated) climate forcing from January 2015–July 2016 and forecasted GPP and soil carbon with stochastically generated forcing from August 2016–December 2024 (Figure 34.3a, b, red). Similarly, the stochastically generated forcing in Series 3 (S3) starts from January 2017 (Figure 34.3a, b, blue). For each series, predictions were conducted with randomly sampled parameters from the posterior distributions and stochastically generated forcing. 100 mean values are displayed (across an ensemble of forecasts with different parameters) corresponding to 100 forecasts with stochastically generated forcing.

GPP is highly sensitive to climate forcing. The differences between the updated (S2, 3) and initial forecasts (S1) reach almost $800 \text{ gC m}^{-2} \text{ yr}^{-1}$

(Figure 34.3c). The discrepancy is strongly dampened in the following 1–2 years. The impact of updated forecasts is close to zero after approximately five years. However, the soil carbon pool shows a different pattern. The soil carbon pool is increased by less than 150 gC m^{-2} , which is relatively small compared to the carbon pool size of ca. $62,000 \text{ gC m}^{-2}$. For soil, the impact of updated forecasts grows with time and is highest at the end of the simulation year 2024. GPP is sensitive to the immediate change in climate forcing, while the updated ecosystem state (or initial value) has a minimum impact on the long-term forecast of GPP. The impact of updated climate forcing is relatively small for soil carbon forecasts during our study period. Soil carbon is less sensitive to the immediate change in climate compared to GPP. However, the alteration of system status affects the soil carbon forecast, especially on a longer time-scale. Since we are archiving updated forecasts every week, we can track the relative contribution of ecosystem status, forcing uncertainty, and parameter distributions to the overall forecasting patterns of different ecological variables and how these patterns evolve in time. In addition, as more observations of ecological variables (e.g., carbon fluxes and pool sizes) become available, it is feasible to diagnose key factors that promote robust forecasts by comparing the archived forecasts to observations and these to the model parameters, initial values, and climate forcing used.

In addition to scientific capability, the EcoPAD system brings new opportunities to broaden user–model interactions and facilitate forecasting practice. The high complexity and long learning curve of ecological models and data–model fusion techniques frequently discourage researchers and impede progress in forecasting practice. EcoPAD is designed to reduce these hurdles. It can be accessed from a web browser and does not require any coding by the user, which opens the door for non-modelers to work with models. The online

storage of results lowers the risk of data loss. The results of each model run can be easily tracked and shared with a unique ID and web address. In addition, the web-based workflow saves time for experts through automated model running, data assimilation, forecasting, structured result access, and instantaneous graphic outputs, allowing the researcher to focus on a thorough exploration of results. At the same time, advanced users have the flexibility to scrutinize or modify model code, embed a different ecological model, change the data assimilation algorithm or add new forecasting properties.

In summary, EcoPAD provides an effective infrastructure with its interactive platform that rigorously integrates merits from models, observations, statistical advance, information technology and human resources from experimentalists and modelers to practitioners and the general public. This facilitates progress in ecological forecasting and analysis. That being said, ecological forecasting and the EcoPAD platform are both at their early development stage. We need more creative ideas and community efforts to realize the potential of this promising field.

SUGGESTED READING

Huang, Y., Stacy, M., Jiang, J., Sundi, N., Ma, S. et al. 2019. Realized ecological forecast through an interactive Ecological Platform for Assimilating Data (EcoPAD, v1.0) into models. *Geoscientific Model Development* 12: 1119–1137.

QUIZZES

1. What is ecological forecasting?
2. What key challenges and barriers does EcoPAD help overcome?
3. What are the four major components of EcoPAD?
4. List four potential applications of EcoPAD?

CHAPTER THIRTY-FIVE

Practice 9

ECOLOGICAL FORECASTING AT THE SPRUCE SITE

Jiang Jiang

Nanjing Forestry University, Nanjing, China

CONTENTS

Introduction / 301

Dataset Preparation for EcoPAD / 302

Accessing and Working with EcoPAD-SPRUCE / 302

This practice aims to help readers gain familiarity with ecological forecasting by using EcoPAD to perform ecological forecasting at the SPRUCE experiment. The web portal of EcoPAD-SPRUCE provides automated ecological forecasting at a weekly time scale. From the web portal, users can check current near- and long-term forecasting results, conduct model simulations, data assimilation, and forecasting runs, and analyze and visualize model results. There are three exercises in this practice, using either CarboTrain or EcoPAD. We delve into how constrained posterior parameters influence forecast uncertainty; how different forcing influences forecast uncertainty; and how ecosystem forecast responds to warming and elevated CO₂ with fully specified uncertainties.

INTRODUCTION

To generate a realistic projection of terrestrial carbon dynamics, we need to have three elements perfectly aligned (Chapter 33). First, we need a good model structure, which represents underlying ecosystem processes. Then, we need a good parameterization method, such as data assimilation discussed in units 6–8, to estimate parameter values. Finally, we also need a workflow system to link real-time forcings to the model (Chapter 34). In this practice,

we will focus on the third element, using EcoPAD to link forcing data to carbon cycle models that usually need climatic forcing to drive the projections.

The Ecological Platform for Assimilating Data into model (EcoPAD) provides this functionality. EcoPAD is described in detail in Chapter 34. Traditionally, modelers usually tune a model, validate the model with data, and then generate model output as prediction. This is called forward modeling. In addition to forward modeling, EcoPAD can perform data assimilation and ecological forecasting. EcoPAD links data and model via data assimilation to optimize parameter estimation. The system can update parameter estimations when new data becomes available to generate real-time forecasting. From an ensemble of parameters, EcoPAD can also produce an ensemble of forecasts, instead of just one projection. This method allows us to quantify the uncertainty of forecasting. The system can provide feedback and link the experimental and modeling communities, informing experimentalists on which data sets are needed to further improve model predictions; and modelers on which parts of a model are responsible for inaccurate or poorly constrained predictions, and may need to be improved.

EcoPAD requires two categories of datasets, observation data and forcing data, to be able to

generate forecasts. Observation data is used to optimize the model parameters through data assimilation. Observations might include vegetation or soil inventory data, laboratory measurements, or high temporal resolution sensor data such as ecosystem flux measurements from the FLUXNET database. Evaluation of different datasets for their effectiveness in constraining model parameters is described in Chapter 29. The other category of data sets is used as forcing to drive a model. The temporal resolution of forcing data should be same as the time step of the model used for projections, and the data should cover the same time period as the historical and future part of the simulations.

In this practice, EcoPAD is applied to the SPRUCE Experimental Forest in Northern Minnesota, USA. The SPRUCE experiment is described in detail in Chapter 25. SPRUCE is an ongoing project that focuses on long-term responses of northern peatland to climate warming and increased atmospheric CO₂ concentration. The project generates a large variety of observational datasets that reflect ecosystem dynamics on different scales. These datasets are available from the project web page and file transfer protocol (FTP) site. EcoPAD accesses this FTP site directly and automatically to download the data.

DATASET PREPARATION FOR EcoPAD

In EcoPAD-SPRUCE, observational datasets are sourced from SPRUCE archives and stored in the EcoPAD metadata catalog for running the TECO model and conducting data assimilation. The forcing datasets to drive TECO in EcoPAD are hourly climate data, which can be separated into two parts, namely past climate and projection of future climate. Climate data are automatically downloaded from the SPRUCE FTP site every week. The future climate timeseries are generated as an ensemble of future climate using vector auto-regressive modeling. Observational data can be updated any time when available. In this training version, however, we use pretreatment datasets from 2011 to 2014 to investigate how constrained posterior parameters influence forecasting uncertainty, and partition the uncertainty sources.

Pretreatment observations include three data sets of community-scale flux measurements (gross primary production, GPP; net ecosystem exchange, NEE; and ecosystem respiration, Reco) in 1.2 m internal diameter chambers, six data sets of plant

biomass growth and carbon content (foliage, wood and root), one data set of carbon in peat soil, and leaf phenological data. During 2011–2014, CO₂ flux observations were collected monthly during the growing season at ambient plots of the experimental site. A total of 30 data measurements were collected in August, September, and October 2011; May through November 2012; July, September, and October 2013; and June and July 2014. Three annual data points from 2012 to 2014 for plant foliage, woody biomass and aboveground net primary production, NPP, were estimated from inventory data. Biomass data were compiled by combining allometric data for shrubs, all ground layer species, and trees. Only one data point each was collected for fine-root and peat soil C. We collected leaf-out dates, which are calculated as growing degree-days above a threshold in the TECO model. The standard deviations reported in these data sets were also compiled to estimate uncertainties for each data stream.

The compiled observational data file `SPRUCE_obs.txt` can be found in any EcoPAD simulation results under the `input` folder. It is also available in GitHub repository (https://github.com/ou-ecolab/teco_spruce/tree/master/input). The first column of the data file is “days”, which records the day number when data were collected, using January 1st, 2011 as the first day. The subsequent columns contain data values for each variable in the dataset. If the data are not available or missing, the value is shown as -9999.

External forcing of the TECO model includes hourly climate data of photosynthetically active radiation (PAR), air temperature, soil temperature, precipitation, relative humidity, vapor pressure deficit, and wind speed. We generated an ensemble of 300 trajectories of ten year forcing variables from 2015 to 2024 as inputs for the forecasting period. The generated climate data are archived in the EcoPAD metadata catalog. For each operational forecasting, the system updates climate data streams from the SPRUCE FTP site to replace stochastically generated forcing.

ACCESSING AND WORKING WITH EcoPAD-SPRUCE

EcoPAD-SPRUCE web portal, the forecasting system for SPRUCE, is available at https://ecolab.nau.edu/ecopad_portal/. From the web portal, users can check current near- and long-term forecasting results, conduct model simulations, data

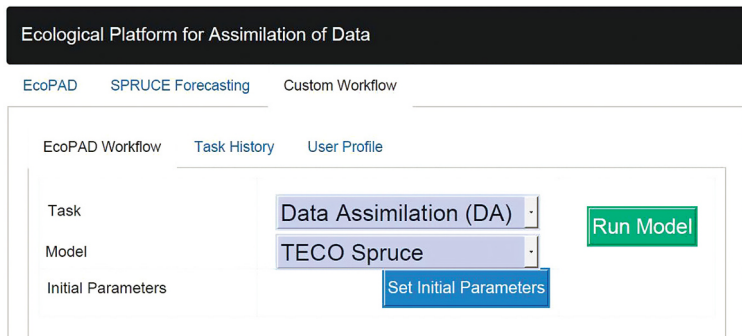


Figure 35.1. The Custom Workflow web portal of EcoPAD applied for the SPRUCE project. Users can select among simulation, data assimilation and forecasting modes from the task drop-down box to run ecological models in the background.

assimilation, and forecasting runs, and analyze and visualize model results.

The front page of the EcoPAD-SPRUCE portal includes animation demos and a brief description of the system. The animation demos display the dynamic change of GPP, Reco, foliage carbon (foliage C), wood carbon (wood C), root carbon (root C) and soil carbon (soil C) under 10 manipulative warming and elevated atmospheric CO₂ treatments. Each animation shows observations in the data assimilation period during which parameters are constrained (2011–2014) as well as model results with uncertainty from data assimilation and ten years forecasting from an ensemble of model runs. Warming generally increases GPP, Reco and carbon pools. Users can also get a sense of how uncertainties in forcing variables, such as light, temperature, and precipitation, that drive carbon fluxes in terrestrial ecosystems, affect uncertainty of predictions.

Under the *SPRUCE Forecasting* menu, forecast-ing GPP and Reco is automated at weekly tim-escape. Mean trajectory and confidence interval

are derived from 100 stochastic simulations. In each simulation, parameters are randomly chosen from previous (or default) data assimilation tasks. Data assimilation is run irregularly, when-ever observation data is compiled manually, to constrain and update the parameters. Users can choose forecasting results from different warm-ing and elevated atmospheric CO₂ treatments in a drop-down box.

Under the *Custom Workflow* menu, users can choose between three different modes to run the TECO model: *Simulation*, *Data Assimilation (DA)* and *Forecasting* (Figure 35.1). Each task requested by a user is assigned a unique task ID. Users can check information such as task ID, timestamp, param-eters, result status, and result URL from a web-enabled report once the task is submitted under the *Task History* tab. If the task status shows “SUCCESS”, users can check datasets relevant to task outputs from the result URL. The URL directs users to the location (result repository) where information related to model runs is stored.

EXERCISE 1: How do constrained posterior parameters influence forecasting uncertainty?

There are two options to do this exercise: using either CarboTrain or EcoPAD. If you choose to use CarboTrain, please open the software, and select unit 9 and Exercise 1. An output folder should also be specified. Options allow you to set parameter initial values and ranges for data assimilation. Click *Run Exercise* to perform data assimilation to constrain parameters, followed by forecasting. Depending on the power of your

computer, it may take a long time to run data assimilation. You can run the data assimilation procedure for about ten minutes and then press ‘Ctrl+C’ on your keyboard to produce a set of parameters that may not be well constrained. The system will continue to run the forecasting code and plot results. You can find all the results in the user-defined output folder.

Alternatively, you may perform the task in the EcoPAD web version. Go to the webpage at https://ecolab.nau.edu/ecopad_portal_workshop_v2/, and follow Ex 1.1, Ex 1.2 and Ex 1.3, below, to complete all the steps.

Ex 1.1 Running the task of forecasting

- Click the custom workflow, and choose *Forecasting* under the task drop-down box.
- Specify which parameters to use from previous DA. The system has preset default DA parameters.
- Run the exercise, and wait around ten minutes to get a “SUCCESS” run.
- Open the results link, and check the results in the Output Folder.

QUESTIONS:

- Do the trajectories of forecasting results match the real data?
- What’s the difference between forecasting results among different variables, such as GPP, Reco, Foliage, Wood, and Soil?

Ex 1.2 Running the task of data assimilation

- Select *Data Assimilation (DA)* in the task drop-down box.
- Click *Set DA initial Parameters*. There are 47 parameters. Up to 18 may be chosen to participate in data assimilation.
- Click *Run model*, and wait around ten minutes to get a “SUCCESS” run.
- Open the results link, and check the results in the Output Folder.

- Repeat the above steps for pool-based versus flux-based parameters and review the updated posterior distribution of parameters.
- Repeat Ex 1.1 using updated parameters to show forecasting results.

QUESTION:

What key differences are seen in the forecasting results following assimilation of pool-based versus flux-based parameters?

Ex 1.3 Effects of different posterior distribution of parameter values

- Repeat Ex 1.2 to perform data assimilation on any of the 18 parameters.
- Repeat Ex 1.1, but choose other parameters by clicking the button *Set Data Assimilation* instead of the default set. If data assimilation has not yet been conducted (Ex 1.2), the *Data Assimilation Parameter Estimate Selection* dialog will be empty.
- Check the results in the Output Folder, and compare the results with outputs from Ex 1.1.

Figure 35.2 lists some of the previous data assimilations. The metadata tags show what

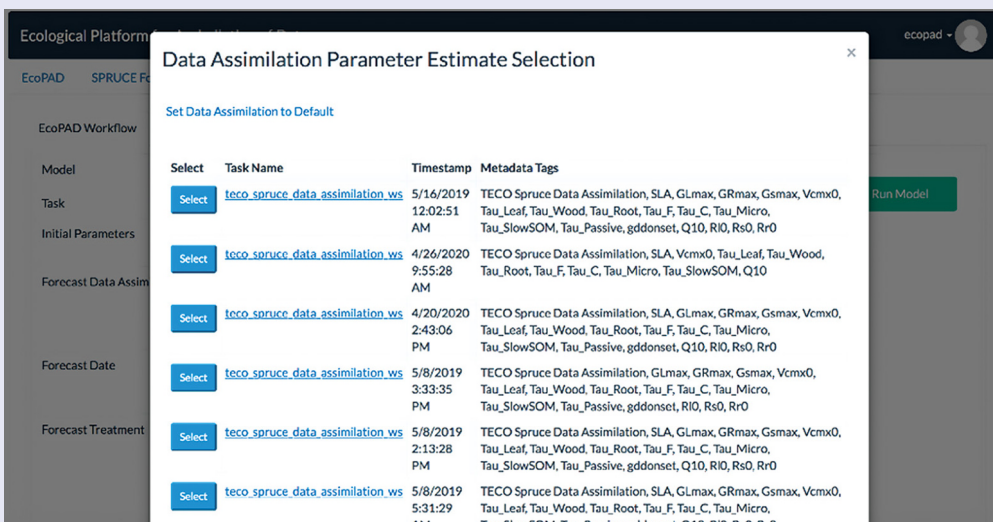


Figure 35.2. Example of outputs from previous data assimilations.

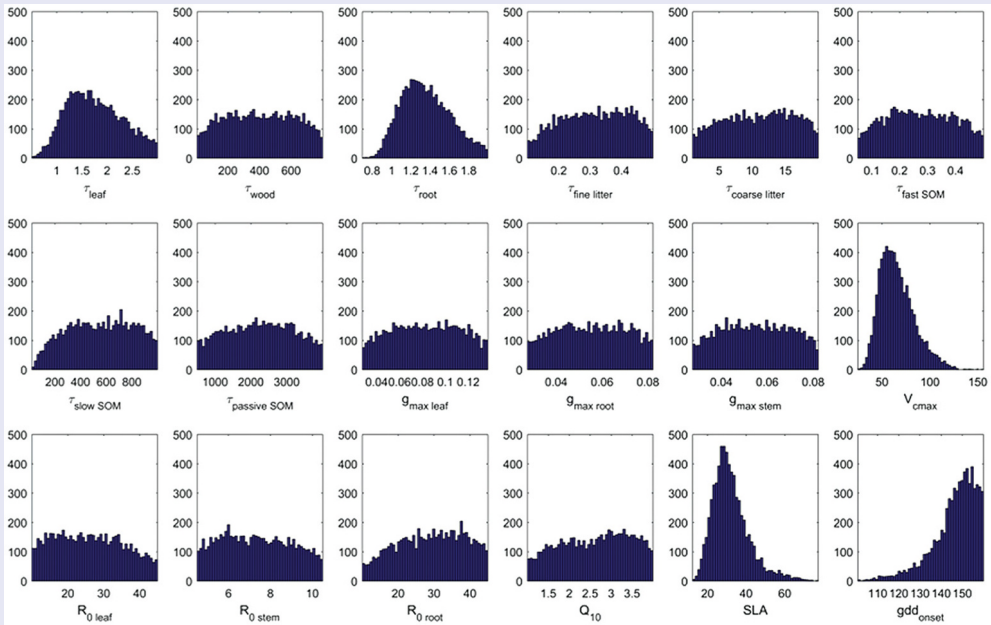


Figure 35.3. Histogram of the posterior distribution of each parameter from a data assimilation run.

parameters were chosen to do data assimilation. Users can select the constrained parameters for forecasting. To check detailed results of previous data assimilation, users can find the datasets under the *Task History* tab according to the Timestamp.

A unique feature of the data assimilation portal is that users can pick whatever parameters are to be constrained among the pool of 18 parameters. Users can change the initial value and range of any parameter to be used for data assimilation.

Figure 35.3 shows an example of forecasting using posterior parameters when all the 18 parameters were chosen for data assimilation.

QUESTIONS:

1. Do the forecasting results depend on parameter posterior distributions?
2. How do the forecasting results change depending on which parameters were constrained by data assimilation?

EXERCISE 2: How does different forcing influence forecasting uncertainty?

External forcing variables such as temperature, precipitation and radiation regulate various carbon cycle processes (e.g., plant photosynthesis, water use, and soil organic matter decomposition), and therefore influence carbon stocks in different compartments of an ecosystem. A challenge for precisely predicting the future state of an ecosystem is the low predictability of future trajectories of forcing variables. This

exercise addresses the implications of uncertainty in forcing for carbon cycle predictions.

There are two options to do this exercise: using either CarboTrain or EcoPAD. To make the task simple, CarboTrain provides two sets of forcing variables, one is fixed forcing; the other is random forcing chosen from a forcing pool.

- a. Select unit 9 and Exercise 2 in the main window of CarboTrain.
- b. Choose *fixed forcing* under the *Forcing Variables* tab, click on *Run Exercise*.

- c. Repeat the above steps but choose *random forcing* under the *Forcing Variables* tab.

If you prefer to do the task in EcoPAD web version, go to the webpage at http://ecolab.nau.edu/ecopad_portal_workshop_v2.

- a. Choose *Forecasting with Different Forcing files* under the *Task* drop-down box.

- b. Choose *fixed forcing* under the *Number of Forcing Sets* tab, and run the exercise.
- c. Repeat the above steps but choose *random forcing* under *Number of Forcing Sets* tab.

QUESTION:

How do the forecasting results differ when using different forcing files, comparing fixed forcing and random forcing?

EXERCISE 3: Uncertainty in forecasting ecosystem responses to warming and elevated CO₂

Forecasting is not very informative without fully specified uncertainties. In the SPRUCE experiment, forecasting is especially useful if it can help us to anticipate how long from the commencement of the experimental treatments it may take before carbon pool changes may be expected to be significantly different between temperature or CO₂ treatments. It takes time for carbon pool sizes to adjust in response to different treatments. In general, if the magnitude of uncertainty in forecast of one pool is small, the statistical power to detect the treatment effects on the pool is large. And the time points to observe statistical difference among treatments is short, and vice versa.

Again, there are two options to do this exercise: using either CarboTrain or EcoPAD. In CarboTrain, the steps are:

- a. Select unit 9 and Exercise 3 in the main window of CarboTrain.
- b. Click *Set DA folder* to specify which parameters to use from previous data assimilation outputs.
- c. Enter the date for *Forecast date end*. The current forecasting allows any date up to 2024/12/31.

- d. Enter the warming treatment within the range 0.0°C–9.0°C, and CO₂ adjustment within the range 380–900 ppm.
- e. Click *Run exercise* and compare forecasting simulation results among different treatments.

If you prefer to do the task in EcoPAD web version, go to the webpage at http://ecolab.nau.edu/ecopad_portal_workshop_v2.

- a. Click the custom workflow, and choose *Forecasting* under the task drop-down box.
- b. Click *Set Data Assimilation* to specify which parameters to use from previous data assimilation outputs.
- c. Under the *Forecast Treatment* tab, enter the warming treatment within the range 0.0°C–9.0°C, and CO₂ adjustment within the range 380–900 ppm.
- d. Run the exercise and compare forecasting simulation results among different treatments.

QUESTIONS:

1. How do the ecological carbon dynamics respond to warming and elevated CO₂?
2. How do the uncertainties of the forecast vary with treatments?

UNITTEN

Process-based Machine Learning and Data- driven Modeling (PRODA)



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Introduction to Machine Learning and Neural Networks

Toby Dylan Hocking

Northern Arizona University, Flagstaff, USA

CONTENTS

Introduction and Applications of Machine Learning / 309
 K-fold Cross-Validation for Evaluating Prediction/Test Accuracy / 310
 Other Applications 311
 Avoiding Under/Overfitting in a Neural Network for Regression / 312
 Comparing Neural Networks for Image Classification / 314
 Cross-Validation for Evaluating Predictions of Earth System Model Parameters / 315
 Suggested Reading / 317
 Quizzes / 317

In this chapter we introduce basic concepts and algorithms from machine learning. We explain how neural networks can be used for regression and classification problems, and how cross-validation can be used for training and testing machine learning algorithms.

INTRODUCTION AND APPLICATIONS OF MACHINE LEARNING

Machine learning is the domain of computer science which is concerned with efficient algorithms for making predictions in all kinds of big data sets. A defining characteristic of supervised machine learning algorithms is that they require a data set for training. The machine learning algorithm then memorizes the patterns present in those training data, with the goal of accurately predicting similar patterns in new test data. Many machine learning algorithms are domain-agnostic, which means they have been shown to provide highly accurate predictions in a wide variety of application domains (computer vision, speech recognition, automatic translation, biology, medicine, climate science, chemistry, geology, etc.).

For example, consider the problem of image classification from the application domain of

computer vision. In this problem, we would like a function that can input an image, and output an integer which indicates class membership. More precisely, let us consider the MNIST and Fashion-MNIST data sets (Figure 36.1), in which each input is a grayscale image with height and width of 28 pixels, represented as a matrix of real numbers $\mathbf{x} \in \mathbb{R}^{28 \times 28}$ (LeCun et al., 1998, Xiao et al., 2017). In both the MNIST and Fashion-MNIST data sets each image has a corresponding label which is an integer $y \in \{0, 1, \dots, 9\}$. In the MNIST data set each image/label represents a digit, whereas in Fashion-MNIST each image/label represents

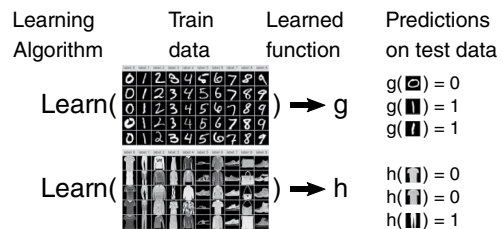


Figure 36.1. A learning algorithm inputs a train data set, and outputs a prediction function, g or h. Both g and h input a grayscale image and output a class (integer from 0 to 9), but g is for digits and h is for fashion.

a category of clothing (0 for T-shirt/top, 1 for Trouser, 2 for Pullover, etc). In both data sets the goal is to learn a function $f: \mathbb{R}^{28 \times 28} \rightarrow \{0, 1, \dots, 9\}$ which inputs an image \mathbf{x} and outputs a predicted class $f(\mathbf{x})$ which should ideally be the same as the corresponding label y .

As mentioned above, a big advantage of supervised learning algorithms is that they are typically domain-agnostic, meaning that they can learn accurate prediction functions f using data sets with different kinds of patterns. That means we can use a single learning algorithm LEARN on either the MNIST or Fashion-MNIST data sets (Figure 36.1, left). For the MNIST data set the learning algorithm will output a function for predicting the class of digit images, and for Fashion-MNIST the learning algorithm will output a function for predicting the class of a clothing image (Figure 36.1, right). The advantage of this supervised machine learning approach to image classification is that the programmer does not need any domain-specific knowledge about the expected pattern (e.g, shape of each digit, appearance of each clothing type). Instead, we assume there is a data set with enough labels for the learning algorithm to accurately infer the domain-specific pattern and prediction function. This means that the machine learning approach is only appropriate when it is possible/inexpensive to create a large, labeled data set that accurately represents the pattern/function to be learned.

How do we know if the learning algorithm is working properly? The goal of supervised learning is **generalization**, which means the learned prediction function f should accurately predict $f(\mathbf{x}) = y$ for any inputs/outputs (\mathbf{x}, y) that will be

seen in a desired application (including new data that were not seen during learning). To formalize this idea, and to compute quantitative evaluation metrics (accuracy/error rates), we need a test data set, as explained in the next section.

K-fold Cross-Validation For Evaluating Prediction/ Test Accuracy

Each input \mathbf{x} in a data set is typically represented as one of N rows in a “design matrix” with D columns (one for each dimension or feature). Each output y is represented as an element of a label vector of size N , which can be visualized as another column alongside the design matrix (Figure 36.2, left). For example, in the image data sets discussed above we have $N = 60,000$ labeled images/rows, each with $D = 784$ dimensions/features (one for each of the 28×28 pixels in the image).

The goal of supervised learning is to find a prediction function f such that $f(\mathbf{x}) = y$ for all inputs/outputs (\mathbf{x}, y) in a test data set (which is not available for learning f). So how do we learn f for accurate prediction on a test data set, if that test set is not available? We must assume that we have access to a train data set with the same statistical distribution as the test data. The train data set is used to learn f , and the test data can only be used for evaluating the prediction accuracy/error of f .

Some benchmark data sets which are used for machine learning research, like MNIST and Fashion-MNIST, have designated train/test sets. However, in most applications of machine learning to real data sets, train/test sets must be created. One approach is to create a single train/test split by randomly assigning a set to each of the

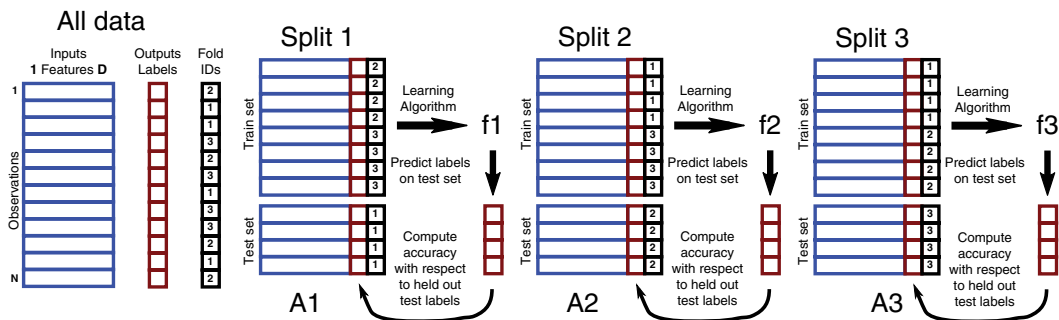


Figure 36.2. $K = 3$ fold cross-validation. Left: the first step is to randomly assign a fold ID from 1 to K to each of the observations/rows. Right: in each of the $k \in \{1, \dots, K\}$ splits, the observations with fold ID k are set aside as a test set, and the other observations are used as a train set to learn a prediction function (f_1 – f_3), which is used to predict for the test set, and to compute accuracy metrics (A_1 – A_3).

N rows/observations, say 50% train rows and 50% test rows. The advantage of that approach is simplicity, but the drawback is that we can only report accuracy/error metrics with respect to one test set (e.g., the algorithm learned a function which accurately predicted 91.3% of observations/labels in the test set, meaning 8.7% error rate).

In addition to estimating the accuracy/error rate, it is important to have some estimate of variance in order to make statements about whether the prediction accuracy/error of the learned function f is significantly larger/smaller than other prediction functions. The other functions to compare against may be from other supervised learning algorithms, or some other method that does not use machine learning (e.g., a domain-specific physical/mechanistic model). A common baseline is the constant function $f(\mathbf{x}) = y_0$ where y_0 is the average or most frequent label in the train data. This baseline ignores all of the inputs/features \mathbf{x} , and can be used to show that the algorithm is learning some non-trivial predictive relationship between inputs and outputs (for an example see Figure 36.4).

The K -fold cross-validation procedure generates K splits, and can therefore be used to estimate both mean and variance of prediction accuracy/error. The number of folds/splits K is a user-defined integer parameter which must be at least 2, and at most N . Typical choices range from $K = 3$ to 10, and usually the value of K does not have a large effect on the final estimated mean/variance of prediction accuracy/error. The algorithm begins by randomly assigning a fold ID number (integer from 1 to K) to each observation (Figure 36.2, left). Then for each unique fold value from 1 to K , we hold out the corresponding observations/rows as a test set, and use data from all other folds as a train set (Figure 36.2, right). Each train set is used to learn a corresponding prediction function, which is then used to predict on the held-out test data. Finally, accuracy/error metrics are computed in order to quantify how well the predictions fit the labels for the test data. Overall, for each data set and learning algorithm the K -fold cross-validation procedure results in K splits, K learned functions, and K test accuracy/error metrics, which are typically combined by taking the mean and standard deviation (or median and quartiles). Other algorithms may be used with the same fold assignments, in order to compare algorithms in terms of accuracy/error rates in particular data sets.

For example, Figure 36.4 uses $K = 4$ -fold cross-validation to compare four learned functions on an image classification problem. The accuracy rates of the “dense” and “linear” functions, $97.4 \pm 1.6\%$ and $96.3 \pm 1.9\%$ (mean \pm standard deviation) are not significantly different. Both rates are significantly larger than the accuracy of the “baseline” constant function, $16.4 \pm 1.4\%$, and smaller than the accuracy of the “conv” function, $99.3 \pm 1.1\%$. We can therefore conclude that the most accurate learning algorithm for this problem, among these four candidates, is the “conv” method (which uses a convolutional neural network, explained later). It is important to note that statements about which algorithm is most accurate can only be made for a particular data set, after having performed K -fold cross-validation to estimate prediction accuracy/error rates.

OTHER APPLICATIONS

So far we have only discussed machine learning algorithms in the context of a single prediction problem, image classification. In this section we briefly discuss other applications of machine learning. In each application the set of possible inputs \mathbf{x} and outputs y are different, but machine learning algorithms can always be used to learn a prediction function $f(\mathbf{x}) \approx y$. Jones et al. (2009) proposed to use interactive machine learning for cell image classification in the CellProfiler Analyst system. This application is similar to the previously discussed digit/fashion classification problem, but with only two classes (binary classification). In this context the input is a multi-color image of cell $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$ where h, w are the height and width of the image in pixels, and $c = 3$ is the number of channels used to represent a color image (red, green, blue). The output $y \in \{0, 1\}$ is a binary label which indicates whether or not the image contains the cell phenotype of interest.

Some email programs use machine learning for spam filtering, which is another example of a binary classification problem. When you click the “spam” button in the email program you are labeling that email as spam ($y = 1$), and when you respond to an email you are labeling that email as not spam ($y = 0$). The input \mathbf{x} is an email message, which can be represented using a “bag-of-words” vector (each element is the number of times a specific word occurs in that email message).

Russell et al. (2008) proposed the LabelMe tool for creating data sets for image segmentation, which is more complex than the previously discussed image classification problems. In this context the input $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$ is typically a multi-color image, and the output $\mathbf{y} \in \{0, 1\}^{h \times w}$ is a binary mask (one element for every pixel in the image) indicating whether or not that pixel contains an object of interest.

Machine learning can be used for automatic translation between languages. In this context the input is a text in one language (e.g., French) and the output is the text translated to another language (e.g., English). The desired prediction function f inputs a French text and outputs the English translation.

Machine learning can be used for medical diagnosis. For example, Poplin et al. (2017) showed that retinal photographs can be used to predict blood pressure or risk of heart attack. Since the output y is a real number (e.g., blood pressure of 120 mm mercury), we refer to this as a regression problem.

AVOIDING UNDER/OVERFITTING IN A NEURAL NETWORK FOR REGRESSION

In this section we begin by explaining the prediction function and learning algorithm for a simple neural network. We then demonstrate how the number of iterations of the learning algorithm can be selected using a validation set, in order to avoid underfitting and overfitting.

We consider a simple regression problem for which the input $x \in \mathbb{R}$ is a single real number ($D = 1$ feature/column in the design matrix), and the output $y \in \mathbb{R}$ is as well. Using a neural network with a single hidden layer of U units, two unknown **parameter** vectors are apparent which need to be learned using the training data, $\mathbf{w} \in \mathbb{R}^U$ and $\mathbf{v} \in \mathbb{R}^U$. The prediction function f is then defined as:

$$f(x) = \mathbf{w}^T \sigma(x\mathbf{v}) = \mathbf{w}'z, \quad (36.1)$$

where $\sigma : \mathbb{R}^U \rightarrow \mathbb{R}^U$ is a non-linear activation function, and $z \in \mathbb{R}^U$ is the vector of hidden units. Typical activation functions include the logistic sigmoid $\sigma(t) = 1/(1 + \exp(-t))$ and the rectifier (or rectified linear units, ReLU) $\sigma(t) = \max(0, t)$. The prediction function is learned using gradient descent, which is an algorithm

that attempts to find parameters \mathbf{w}, \mathbf{v} which minimize the mean squared error between the predictions and the corresponding labels in the N train data:

$$L(\mathbf{w}, \mathbf{v}) = \frac{1}{N} \sum_{i=1}^N [W^T \sigma(x_i \mathbf{v}) - y_i] \quad (36.2)$$

Gradient descent begins using uninformative parameters $\mathbf{w}_0, \mathbf{v}_0$ (typically random numbers close to zero), then at each iteration $t \in \{1, \dots, T\}$ the parameters are improved by taking a step of size $\alpha > 0$ in the negative gradient direction,

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \alpha \nabla_{\mathbf{w}} L(\mathbf{w}_{t-1}, \mathbf{v}_{t-1}) \quad (36.3)$$

$$\mathbf{v}_t = \mathbf{v}_{t-1} - \alpha \nabla_{\mathbf{v}} L(\mathbf{w}_{t-1}, \mathbf{v}_{t-1}) \quad (36.4)$$

The algorithm described above is referred to as “full gradient” because the gradient descent direction is defined using the full set of N samples in the train set. Other common variants include “stochastic gradient” (gradient uses one sample) and “minibatch” (gradient uses several samples). When doing gradient descent on a neural network model, one “epoch” includes computing gradients once for each sample (e.g., 1 epoch = 1 iteration of full gradient, 1 epoch = N iterations of stochastic gradient).

In the algorithm above, the number of hidden units U , the number of iterations T , and the step size α must be fixed before running the learning algorithm. These **hyper-parameters** affect the learning capacity of the neural network. An important consideration when using any machine learning algorithm is that you most likely need to tune the hyper-parameters of the algorithm in order to avoid underfitting and overfitting. **Underfitting** occurs when the learned function f neither provides accurate predictions for the train data, nor the test data. **Overfitting** occurs when the learned function f only provides accurate predictions for the train data (and not for the test data). Both underfitting and overfitting are bad, and need to be avoided, because the goal of any learning algorithm is to find a prediction function f which provides accurate predictions in test data.

How can we select hyper-parameters which avoid overfitting? Note that the choice of

hyper-parameters such as number of hidden units U and iterations T affect the learned function f , so we cannot use the test data to learn these hyper-parameters (by assumption that the test data are not available at train time). Then, how do we know which hyper-parameters will result in learned functions which best generalize to new data?

A general method which can be used with any learning algorithm is splitting the train set into subtrain and validation sets, then using grid search over hyper-parameter values. The subtrain set is used for parameter learning, and the validation set is used for hyper-parameter selection. In detail, we first fix a set of hyper-parameters, say $U = 50$ hidden units and $T = 100$ iterations. Then the subtrain set is used with these hyper-parameters as input to the learning algorithm, which outputs the learned parameter vectors \mathbf{w} , \mathbf{v} . Finally, the learned parameters are used to compute predictions $f(x)$ for all inputs x in the validation set, and the corresponding labels y are used to evaluate the accuracy/error of those predictions. The procedure is then repeated for another hyper-parameter set, say $U = 10$ hidden units with $T = 500$ iterations. In the end we select the hyper-parameter set with minimal validation error, and then retrain using the learning algorithm on the full train set with those hyper-parameters. A variant of this method is to use K -fold cross-validation to generate K subtrain/validation splits, then compute mean

validation error over the K splits, which typically yields hyper-parameters that result in more accurate/generalizable predictions (when compared to hyper-parameters selected using a single subtrain/validation split). Note that this K -fold cross-validation for hyper-parameter learning is essentially the same procedure as shown in Figure 36.2, but we split the train set into subtrain/validation sets (instead of splitting all data into train/test sets as shown in the figure).

For example, we simulated some data with a sine wave pattern (Figure 36.3), and used the R package `nnet` to fit a neural network with one hidden layer of $U = 50$ units (Venables and Ripley, 2013). We demonstrate the effects of under/overfitting by varying the number of iterations/epochs from $T = 1$ to 1000. In this example $K = 4$ -fold cross-validation was used, so each data point was randomly assigned a fold ID integer from 1 to 4. The result for only the first split is shown, so observations assigned fold ID=1 are considered the validation set, and other observations (folds 2–4) are considered the subtrain set (which is used at input to the `nnet` R function which implements the gradient descent learning algorithm). We then used the `predict` function in R to compute predictions for subtrain and validation data, and analyzed how the prediction error changes as a function of the number of iterations/epochs T of gradient descent. The data exhibit a

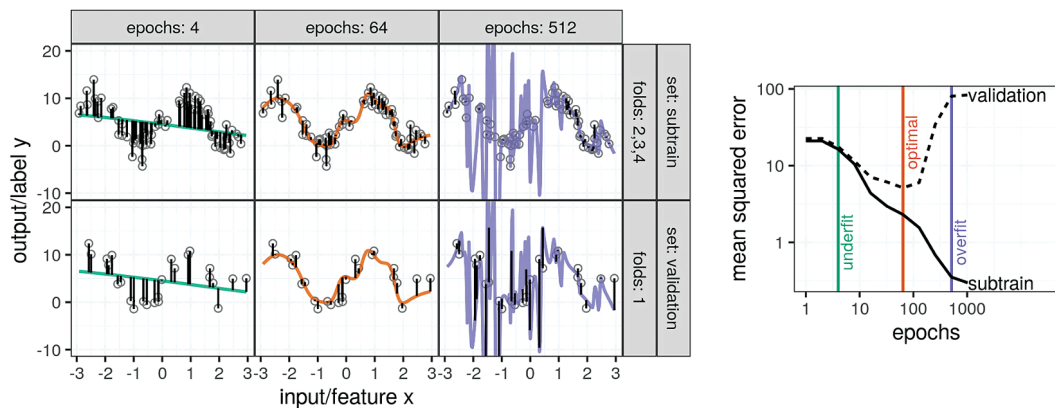


Figure 36.3. Illustration of underfitting and overfitting in a neural network regression model (single hidden layer, 50 hidden units). Left: noisy data with a nonlinear sine wave pattern (grey circles), learned functions (colored curves), and residuals/errors (black line segments) are shown for three values of epochs (panels from left to right) and two data subsets (panels from top to bottom). Right: in each epoch the model parameters are updated using gradient descent with respect to the subtrain loss, which decreases with more epochs. The optimal/minimum loss with respect to the validation set occurs at 64 epochs, indicating underfitting for smaller epochs (green function, too regular/linear for both subtrain/validation sets) and overfitting for larger epochs (purple function, very irregular/nonlinear so good fit for subtrain but not validation set).

nonlinear sine wave pattern, but the learned function for $T = 4$ iterations/epochs is mostly linear (underfitting, large error on both subtrain/validation sets). For $T = 512$ iterations/epochs the learned function is highly non-linear (overfitting, small error for the subtrain set but large error for the validation set). When the error rates are plotted as a function of a model complexity hyper-parameter such as T (Figure 36.3, right), we see the characteristic U shape for the validation error, and the monotonic decreasing train error. The hyper-parameter with minimal validation error is $T = 64$ iterations/epochs; smaller T values underfit or are overly regularized, and larger T values overfit or are under-regularized.

Overall, in this section we have seen how a neural network for regression can be trained using gradient descent (for learning parameter vectors, given fixed hyper-parameters) and sub-train/validation splits (for learning hyper-parameter values to avoid under/overfitting).

COMPARING NEURAL NETWORKS FOR IMAGE CLASSIFICATION

In this section we provide a comparison of several other neural networks for image classification. In general, in a neural network with $L-1$ hidden layers we can represent the prediction function as the composition of L intermediate f_l functions, for all layers $l \in \{1, \dots, L\}$:

$$f(\mathbf{x}) = f_L \left[\dots f_1 \left[\mathbf{x} \right] \right] \quad (36.5)$$

Each of the intermediate functions has the same form:

$$f_l(t) = A_l \left(\mathbf{W}_l^t \right), \quad (36.6)$$

where A_l is an activation function and $\mathbf{W}_l \in \mathbb{R}^{u_l \times u_{l-1}}$ is a weight matrix with elements that must be learned based on the data. This model includes several hyper-parameters which must be fixed prior to learning the neural network weights:

- The number of layers L .
- The activation functions A_l .
- The number of units per layer u_l .
- The sparsity pattern in the weight matrices \mathbf{W}_l .

The number of units in the input layer is fixed, $u_0 = D$, based on the dimension of the inputs $\mathbf{x} \in \mathbb{R}^D$. The number of units in the output layer u_L is also fixed based on the outputs/labels y . The numbers of units in the hidden layers (u_1, \dots, u_{L-1}) are hyper-parameters which control under/overfitting. Increasing the numbers of hidden units u_l results in larger weight matrices \mathbf{W}_l , which in general means more parameters to learn, and larger capacity for fitting complex patterns in the data. The sparsity pattern of \mathbf{W}_l means which entries are forced to be zero; this technique is used in “convolutional” neural networks for avoiding overfitting and reducing training/prediction time. When the matrix is not sparse (all entries non-zero), we refer to the layer as dense or fully connected.

For example, in the previous section we used a neural network for regression with one hidden layer, which in this more general notation means using $L = 2$ intermediate functions; the input dimension is $u_0 = D = 1$, the number of hidden units is $u_1 = U = 50$, and there is a single output $u_2 = 1$ to predict. The weight matrices are dense/fully connected (no convolution/sparsity), of dimension $\mathbf{W}_1 \in \mathbb{R}^{50 \times 1}$, $\mathbf{W}_2 \in \mathbb{R}^{1 \times 50}$. The hidden layer activation function A_1 used by the `R nnet` package is the logistic sigmoid, $\sigma(t) = 1/(1 + \exp(-t))$, and the output activation for regression (real-valued outputs) is the identity, $A_2(t) = t$.

In this section we implement three other neural networks for image classification. Using the “zip.train” data set of $N = 7291$ handwritten digits (Hastie and Tibshirani, 2009), each input is a grayscale image of 16×16 pixels which means that number of input units is $u_0 = 256$. As in Figure 36.1 (top) there are ten output classes, one for each digit. For the activation function A_L in the output layer we use the “softmax” function which results in a score/probability for each of the ten possible output classes, so the number of output units is $u_L = 10$.

The three neural networks that we consider are:

linear $L = 1$ intermediate function with 2,570 parameters to learn (linear model, inputs fully connected to outputs, no hidden units/layers).

dense $L = 9$ intermediate functions with 97,410 parameters to learn (nonlinear model, each hidden layer dense/fully connected with 100 units).

sparse $L = 3$ intermediate functions with 99,310 parameters to learn (nonlinear model, one convolutional/sparse layer followed by two dense/fully connected layers).

We defined and trained each neural network using the `keras` R package (Allaire and Chollet, 2020). We used the `fit` function with argument `validation_split=0.2`, which creates a single split (80% subtrain, 20% validation). We selected the number of epochs hyper-parameter by minimizing the validation loss, and we used the selected number of epochs to re-train the neural network on the entire train set (no subtrain/validation split).

We did this entire procedure $K = 4$ times, once for each fold/split in K -fold cross-validation. Note that even though these data have a pre-defined split into “zip.train” and “zip.test” files, we used K -fold cross-validation on the “zip.train” file, yielding K train/test splits that we used to estimate mean and variance of prediction accuracy for these models (the “zip.test” file was ignored). In each split we used the test set to quantify the prediction accuracy of the learned models. It is clear that the test accuracy of all three neural networks is significantly larger than the baseline model which always predicts the most frequent class in the train set (Figure 36.4, left); they are clearly learning some non-trivial predictive relationship between inputs and outputs. Furthermore, it is clear from Figure 36.4 (right) that the dense neural network is slightly more accurate than the linear model ($p = 0.032$ in paired one-sided t_3 -test), and the sparse/convolutional neural network is significantly more accurate than the dense model ($p = 0.009$).

In summary, from this comparison it is clear that among these three neural networks, the sparse model should be preferred for most accurate predictions in this particular “zip” data set. However,

we must be careful not to generalize these conclusions to other data sets — even for some other image classification data sets such as MNIST (Figure 36.1), the most accurate algorithm may be different. For very difficult data sets, it may even be the case that these three neural networks are no more accurate than the baseline model which always predicts the most frequent class in the train set. In general, we always need to use computational cross-validation experiments to determine which machine learning algorithm is most accurate in any given data set. To learn a predictive model with maximum prediction accuracy, machine learning algorithms other than neural networks should be additionally considered (e.g., regularized linear models, decision trees, random forests, boosting, support vector machines).

CROSS-VALIDATION FOR EVALUATING PREDICTIONS OF EARTH SYSTEM MODEL PARAMETERS

As a final example application, we consider using cross-validation to evaluate a neural network that predicts carbon cycle model parameters (Tao et al., 2020). In this context there is a data set with $N = 26,158$ observations, each one a soil sample with $D = 60$ input features. There are 25 real-valued output variables to predict; each is the value of an earth system model parameter at the location of the soil sample. We want a neural network that will be able to predict the values of these earth system parameters at new locations. Tao et al. (2020) proposed using a neural network with $L = 4$ fully connected layers and dropout regularization for this task (see paper for details). In this section the “multi-task” model uses the same number of layers/units as described in that paper; the term multi-task means that the neural network outputs

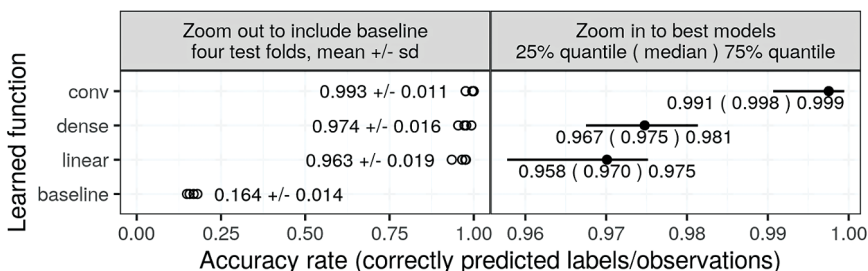


Figure 36.4. Prediction accuracy of functions learned for image classification of handwritten digits. The baseline function always predicts the most frequent class in the train set; other three learned functions are neural networks with different numbers of hidden layers (linear=0, conv=2, dense=8).

a prediction for all 25 outputs/tasks. For comparison, we additionally consider “single-task” models with the same number of hidden layers/units, but only one output unit. We expect the multi-task model to sometimes be more accurate, because of the expected correlation between outputs (earth system model parameters). To see whether or not these neural networks learn any nontrivial predictive relationship between inputs and output, we consider a baseline model which always predicts the mean of the train set label/output values (and does not use the inputs at all).

Here we show how $K = 5$ fold cross-validation can be used to evaluate how well these neural networks predict each of the outputs at new locations. We first assign a fold ID from 1 to 5 to each observation/row, either systematically using the longitude coordinate, or randomly (Figure 36.5, top). We can define a cross-validation procedure using both sets of fold IDs, in order to answer the question, “is it more difficult to predict at new longitudes, or new random locations?” We expect that predicting at new longitudes should be more difficult, because that involves more extrapolation (predicting outside the range of observed data values). In detail, for each fold ID from 1 to 5, we define the test set as the data points which have been assigned that fold ID using both methods

(longitude and random). For these data with $N = 26,158$ observations total, each fold has approximately 5000 observations, so each resulting test set has approximately 1000 observations. As described in the last section on image classification, we used the R `keras` package to compute the neural network parameters and predictions (using a maximum of 100 epochs, and a single 80% subtrain 20% validation split to choose the optimal number of epochs for re-training on the entire train set). For each fold/model/output we computed mean squared error with respect to the test set, and we plot these values for four of the 25 outputs (Figure 36.5, bottom). It is clear that some outputs are more difficult to predict than others; for `cryo` and `maxpsi` outputs the neural networks show little or no improvement over baselines, whereas for `tau4s3` and `fs2s3` outputs we observed substantial improvements over baselines. As expected, there is a difference in test error between fold assignment methods (random has lower error rates than Lon for several outputs), indicating that it is indeed easier to predict at new random locations, and harder to predict at new longitudes. Finally, the multi-task models are slightly more accurate than the single-task models, indicating that the neural network is learning to exploit the correlations between outputs. Overall

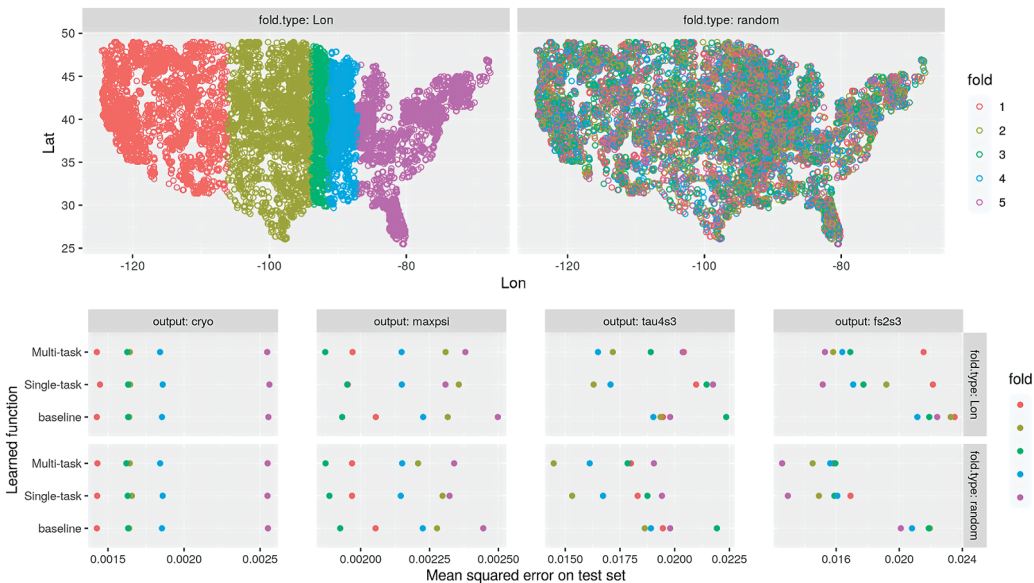


Figure 36.5. Cross-validation for estimating error rates of machine learning algorithms that predict earth system model parameters. **Top:** fold IDs were assigned to each observation using longitude (left) or randomly (right). **Bottom:** prediction error for four of the 25 outputs. Please see (Tao et al., 2020) for meanings of abbreviations (`cryo`, `maxpsi`, `tau4s3`, `fs2s3`).

this comparison has shown how cross-validation can be used to quantitatively evaluate and compare machine learning algorithms for predicting earth system model parameters.

In comparison to the neural network practice in unit 10, the main difference is that here we discussed how held-out test sets can be used to estimate prediction accuracy/error rates of learning algorithms. Chapter 38 discusses how a validation set can be used to avoid overfitting, as we have done in this chapter as well. We have additionally discussed how $K = 5$ fold cross-validation can be used to generate several train/test splits, which can be used to estimate prediction error rates for each fold/data/algorithm combination (e.g., Figure 36.5, bottom). This technique is useful since it allows us to see which algorithms are significantly more/less accurate than others on given data sets.

SUGGESTED READING

Machine learning is a large field of research with many algorithms, and there are several useful textbooks that provide overviews from various perspectives:

Reproducibility statement. Code for figures in this chapter can be freely downloaded from <https://github.com/tdhock/2020-yiqi-summer-school>

- C. M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York.
- I. J. Goodfellow, Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press, Cambridge, MA, USA.
- T. Hastie, R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, Springer Science+Business Media, New York NY, second edition.
- K. P. Murphy. 2013. *Machine Learning: A Probabilistic Perspective*. 2013. MIT Press, Cambridge, MA.

L. Wasserman. 2010. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York.

QUIZZES

1. When using a design matrix to represent machine learning inputs, what does each row and column represent? What other data/options does a supervised learning algorithm such as gradient descent need as input, and what does it yield as output?
2. When splitting data into train/test sets, what is the purpose of each set? When splitting a train set into subtrain/validation sets, what is the purpose of each set? What is the advantage of using K -fold cross-validation, relative to a single split?
3. In order to determine if any non-trivial predictive relationship between inputs and output has been learned, a comparison with a baseline that ignores the inputs must be used. How do you compute the baseline predictions, for regression and classification problems?
4. How can you tell if machine learning model predictions are underfitting or overfitting?
5. When using the `nnet` function in R to learn a neural network with a single hidden layer, do large or small values of the number of iterations hyper-parameter result in overfitting? Why?
6. When using the `nnet` function in R learn a neural network with a single hidden layer, and you do not yet know how many iterations to use, what data set should you use as input to `nnet`? How should you learn the number of iterations to avoid underfitting and overfitting? After having computed the number of iterations to use, what data set should you then use as input to `nnet` to learn your final model? Hint: possible choices for set to use are all, train, test, subtrain, validation.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER THIRTY-SEVEN

PROcess-Guided Deep Learning and DATA-Driven Modelling (PRODA)

Feng Tao

Tsinghua University, Beijing, China

Yiqi Luo

Cornell University, Ithaca, USA

CONTENTS

The Need for Optimizing Parameterization of Earth System Models /	319
The Workflow of PRODA /	320
Model Representation of SOC Content Across Observation Sites /	323
Spatial Distribution of SOC Across the Conterminous U.S. /	323
Vertical Distribution of SOC Across the Conterminous U.S. /	325
Toward More Realistic Representations of SOC Distribution /	327
Suggested Reading /	328
Quizzes /	328

This chapter describes a PROcess-guided deep learning and DATA-driven modeling (PRODA) approach to optimize parameterization of Earth system models (ESMs) using spatio-temporal datasets. PRODA involves both data assimilation to estimate parameter values and deep learning to predict spatial and temporal distributions of parameter values so as to optimize ESM prediction. An application to the Community Land Model version 5 (CLM5) using soil organic carbon (SOC) distributions in the conterminous United States illustrates the potential and utility of the PRODA approach.

THE NEED FOR OPTIMIZING PARAMETERIZATION OF EARTH SYSTEM MODELS

Earth system models (ESMs) are used to simulate historical and potential future states of climate and ecosystems. However, simulations often deviate

substantially from observations. For example, soil carbon dynamics simulated by ESMs vary widely among models and often fit poorly with observations. Modeled global soil carbon storage differs by up to six-fold among 11 models of the Coupled Model Intercomparison Project Phase 5 (CMIP5) ensemble (Todd-Brown et al. 2013). None of the models reproduces the spatial distribution of SOC stocks presented in the Harmonized World Soil Database (HWSD) (Luo et al. 2015).

Uncertainty in simulating SOC dynamics with ESMs could stem from poor parameterization, incorrect model structure, or biased external forcing (Luo and Schuur 2020, chapter 33). While model structure represents ecological processes (e.g., decomposition of soil organic matter), parameters in ESMs characterize properties of the processes, such as baseline decomposition rate at reference temperature and moisture content, or sensitivity to these drivers. The choice of parameter values can strongly influence model projections

of SOC dynamics. Parameter values in the current generation of ESMs, however, are mostly determined on an *ad hoc* basis. They may be derived from the results of field experiments, other models, or informed from scientific or grey literature (Luo et al. 2001), but rarely take into account the range of possible values encompassed by such sources.

Data assimilation techniques to estimate parameter values from observations were discussed and illustrated in earlier chapters (units 6, 7, 8). Parameter values constrained by data assimilation can improve SOC simulation in ESMs compared to the default parameter values. For instance, the global representation of SOC distribution in the Community Land Model version 3.5 (CLM3.5) was improved from explaining 27 to 41% of variation in the HWSD database by constraining model parameters with a Bayesian Markov Chain Monte Carlo (MCMC) data assimilation method (Hararuk et al. 2014). The large unexplained variation in observed SOC with ESMs is partly due to a textbook concept that parameter values of a simulation model must be constant in contrast to variables that can vary over the time course of simulation (Forrester 1961). In reality, ecosystem properties, which parameters characterize in models, constantly evolve via acclimation and adaptation. In addition, a model, no matter how complex it is, can never represent all the processes of a system at resolved scales (Luo and Schuur 2020). Interactions of processes at unresolved scales with those at resolved scales should be reflected in model parameters. Therefore, Luo and Schuur (2020) argue that parameter values in ESMs may have to vary over space and time (i.e., heterogeneous parameter values) to represent changing properties of evolving ecosystems and unresolved processes.

The advent of big ecological data provides a golden opportunity to reconcile model representations with observations and quantify the spatial and temporal features of key parameters in soil carbon cycle simulation. Meanwhile, new techniques such as deep learning have been proposed to improve performance of ESMs (Reichstein et al. 2019). By constructing computational models with multiple processing layers and allowing the models to learn representations of data from multiple levels of abstraction (LeCun et al. 2015), deep learning techniques have promising applications in Earth system science, such as pattern classification, anomaly detection, regression, and space- or time-dependent state prediction (Reichstein et al. 2019). Exploration is warranted on how to properly employ deep

learning techniques in reducing uncertainties of simulated carbon dynamics in ESMs.

Here, we propose the PROcess-guided deep learning and DATA-driven modelling (PRODA) approach to estimate spatially and temporally heterogeneous parameter values for ESMs from extensive spatio-temporal datasets ('big data') at regional or global scales. The PRODA approach estimates parameter values at individual sites via data assimilation and builds a deep learning model to upscale the site-level estimates of parameters to predict spatially heterogeneous parameters at regional and global scales so that modeled and observed SOC are maximally matched.

In this chapter, we introduce the PRODA approach by using an extensive dataset of vertical soil profiles across the conterminous United States to optimize SOC representation by CLM5. We discuss the PRODA-optimized model performance in representing SOC stock and its vertical and spatial distributions, and compare it with results of the default model simulation and after the data assimilation optimization. In particular, we highlight that the PRODA approach helps the process model to achieve the most precise SOC distribution ever represented in ESMs. An accurate SOC representation in ESMs is critical to fully understand soil carbon feedbacks to future climate change.

THE WORKFLOW OF PRODA

Three fundamental components together formulate the PRODA approach (Figure 37.1a), namely the process-based model, the site-level data assimilation, and the deep learning model. Process-based models with their predefined structure and default parameter values simulate SOC distributions using meteorological forcing data. Data assimilation is used to estimate parameter values of a process-based model with soil carbon data at sites where the observations were made. The deep learning model is used to predict optimized site-level parameter values with their associated environmental variables. Eventually, the process-based model will apply the optimized parameter values upscaled by the deep learning model to simulate SOC distributions at regional or global scales.

Process-based model: We use the matrix representation of the Community Land Model version 5 (CLM5) to facilitate data assimilation and model simulation in the PRODA approach (Figure 37.1b). CLM5 is the latest version of CLM

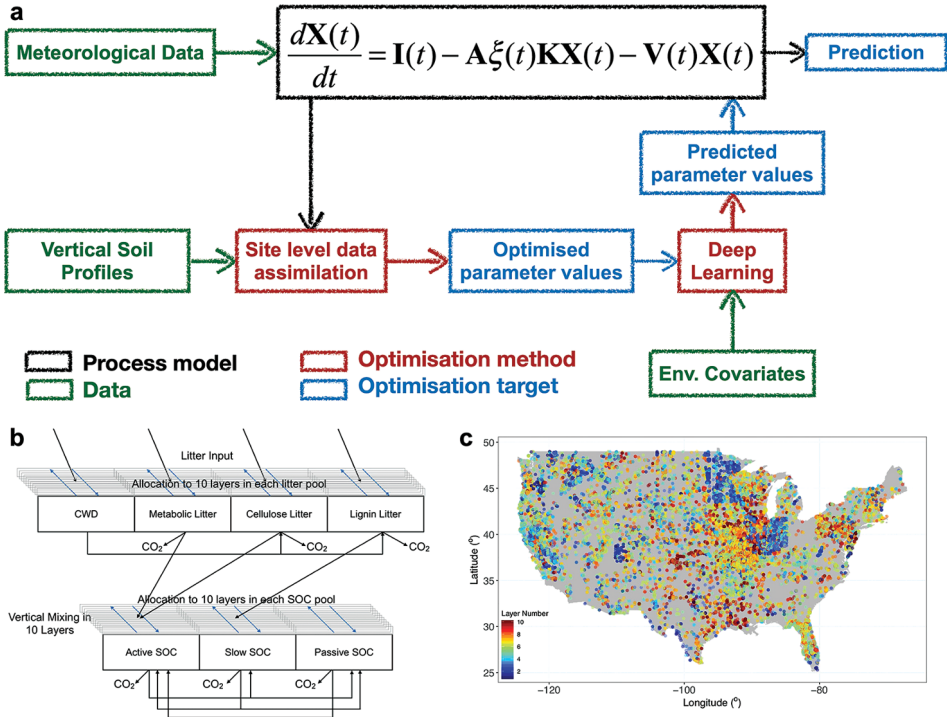


Figure 37.1. Workflow of the PRODA approach. (a) PRODA optimally matches CLM5 as the process-based model (b) with vertical SOC profiles on the conterminous United States (c). We first assimilate data at each site into CLM5 to estimate its parameters through the Markov Chain Monte Carlo method (MCMC). We further assemble the estimated site-level parameter values (i.e., the mean value of the posterior distribution after MCMC) as targets to be predicted by a multilayer neural network with environmental covariates in a deep learning model. The predicted parameters by the deep learning model are applied to CLM5 to optimize model representation of SOC distribution.

models (Lawrence et al. 2019). Its soil carbon module is similar to that in CLM4.5 (Koven et al. 2013), except that it has an option to change the number of soil layers from a default of 20. In this example, we use ten soil layers with a vertical transformation among carbon pools from the surface to a maximum depth of 3.8 m as in CLM4.5. The soil carbon component of CLM5 includes carbon transfer among four litter pools (coarse woody debris, metabolic litter, cellulose litter, and lignin litter) and three soil organic carbon pools (fast, slow, and passive SOC) in each of ten layers, totaling 70 pools. The thickness of soil layers increases exponentially from the surface layer (1.75 cm) to deep layers (151 cm), with a total depth of 3.8 m over the ten layers. Vertical carbon transfer between soil layers only occurs among the adjacent layers and represents both diffusive and advective carbon flux transportation caused by bioturbation and cryoturbation. The baseline advective rate of carbon flux is set to zero in CLM5 as a default, and this is assumed in our example as well.

We have discussed in units 1–5 that carbon balance equations in land carbon models can be unified to a matrix form. For CLM5, we use the matrix equation to describe carbon transfer among the 70 pools with state variables $\mathbf{X}(t)$ as:

$$\frac{d\mathbf{X}(t)}{dt} = \mathbf{B}u(t) + \mathbf{A}\xi(t)\mathbf{K}\mathbf{X}(t) - \mathbf{V}(t)\mathbf{X}(t) \quad (37.1)$$

where \mathbf{B} is a vector (70×1) of partitioning coefficients from C input to each of the pools (unitless), and $u(t)$ is C input rate ($\text{gC m}^{-3} \text{day}^{-1}$). \mathbf{A} represents the transfer coefficients among litter and soil pools (unitless), including the transfer coefficients from four litter pools to three soil carbon pools as well as the transfer coefficients of SOC among soil carbon pools in the same layer. $\xi(t)$ represents effects of environmental variables on decomposition of litter and soil (unitless). It includes scalars of temperature (ξ_T), soil water (ξ_w), oxygen (ξ_o), nitrogen (ξ_N), and depth (ξ_D). \mathbf{K} indicates

the decomposition rate of SOC in different litter and soil carbon pools (day^{-1}). $\mathbf{V}(t)$ represents SOC mixing among vertical soil layers through cryoturbation or bioturbation (day^{-1}). The t in parentheses indicates that the corresponding element is time-dependent. At a steady-state of the carbon cycle ($\frac{d\mathbf{X}(t)}{dt} = 0$), the SOC content of each carbon pool at each layer can be calculated as:

$$\mathbf{X}(t) = [\mathbf{A}\xi(t)\mathbf{K} - \mathbf{V}(t)]^{-1}(-\mathbf{Bu}(t)) \quad (37.2)$$

Soil carbon data and site-level data assimilation: We use vertical SOC profiles in the conterminous U.S. from the World Soil Information Service (WoSIS) dataset (www.isric.org) for the site-level data assimilation (Figure 37.1c). The depth of recorded SOC layers ranges from the surface to more than 3 metres. A total of 26,509 soil profiles with a total of 240,148 layers at different depths in the conterminous U.S. are available in this study.

In addition, we use the mean values of global net primary productivity (NPP) from 2000 to 2014 as carbon input (DAAC 2018). After running the CLM4.5 model to a steady-state by the pre-industrial climate forcing (version code of forcing database: I1850CRUCLM45BGC), ten-year records of soil temperature and soil water potential of the conterminous U.S. were obtained from the model outputs.

The site-level data assimilation constrains parameter values of CLM5 with one data set of a vertical SOC profile at each site with the Markov chain Monte Carlo (MCMC) method (as described in chapter 22). Three parallel chains are generated each containing a test run of 20,000 iterations and a formal run of 30,000 iterations. To effectively capture the vertical distribution pattern of soil content along the depths, we put weights to observations at different depths in calculating the discrepancy between modeled and observed SOC content (i.e., cost function). These weights decrease exponentially with the depth (i.e., $\text{weight}_i = e^{-|\text{depth}_i|}$, where i refers to the layer's soil depth in observations) except for the top layer and the bottom layer, where a weight of ten is assigned to accelerate calibrating the upper and lower bounds of the SOC distribution curve. To monitor the efficiency of the MCMC process, an acceptance rate threshold is set. For Markov chains whose acceptance rate is higher than 50% or lower than 15%,

the corresponding data assimilation results are rejected. After the MCMC process, the first half of the accepted parameter values in the formal run are discarded as burn-in. The Gelman-Rubin statistics of each parameter are then calculated for each soil profile to ensure the convergence of these three independent MCMC results. We randomly select one chain after eliminating the burn-in period to generate the posterior distributions of parameters. The mean value of the parameter's posterior distribution is calculated and chosen to serve as the training target in the deep learning model.

We evaluate the effectiveness of the site-level data assimilation by the coefficient of efficiency:

$$E = 1 - \frac{\sum (\text{obs}_i - \text{mod}_i)^2}{\sum (\text{obs}_i - \overline{\text{obs}})^2} \quad (37.3)$$

where obs_i and mod_i are the observed and modeled SOC content at i th soil layer of one soil profile; $\overline{\text{obs}}$ is the mean value of observed SOC content of the soil profile. In this study, we take profiles having negative E values as invalid and discard the results from the corresponding deep learning model. Moreover, at those sites where an observation is available at only one soil depth, we do not apply the data assimilation to the data point. After those data sets are excluded, 25,444 out of 26,905 soil profiles, or 94.6% of the entire dataset, are used in the PRODA approach.

Deep learning model: We design a deep learning model with multiple processing layers to predict optimized parameter values with environmental covariates. A total of 60 environmental variables that describe the climatic, edaphic and vegetation features at the observational sites is used. We used 80% of the total dataset to train and validate the neural network. After model training, we use the remaining 20% of the dataset to quantify the prediction accuracy of the deep learning model. The predicted parameter values are first compared with those retrieved in site-level data assimilation and then applied to the matrix CLM5 model to simulate soil organic carbon stock at each observational site. Meanwhile, we used the trained deep learning model to generate parameter maps across the United States based on gridded environmental covariates. The parameter maps are then applied to the matrix CLM5 to simulate the SOC distributions across the United States at a resolution of 0.5 degrees.

SOC distributions optimized by data assimilation: To analyse the significance of the spatially-explicit parameter estimation of PRODA compared with a traditional approach, we perform a batch data assimilation using all the observational dataset as one batch in the MCMC method to estimate parameter values of CLM5 with data assimilation. The estimated parameter values from this method are spatially homogeneous, in contrast with the site-level data assimilation, which is a middle step of the PRODA approach to estimate spatially heterogeneous parameter values. SOC distributions simulated by CLM5, trained by the batch data assimilation versus the results by PRODA, can then be compared.

The batch data assimilation runs three parallel MCMC chains, each containing 50,000 iterations as test run and 200,000 iterations as formal run. Weights at different depth in calculating the cost function and acceptance control are the same as those in the site-level data assimilation. After the MCMC method, we first discard the first half of the accepted parameter values of the formal run as burn-in. The Gelman-Rubin statistics for each parameter are then calculated to ensure the convergence of these three independent MCMC results. We randomly select one Markov chain after eliminating the burn-in period to generate the posterior distribution for each parameter. We then randomly sample parameter values from the posterior distributions 1,000 times and apply the sampled parameter values to the CLM5 matrix model. We estimate SOC content distributions at different sites by calculating the average of the results. The same sampled parameter values are further assigned in CLM5 to estimate SOC content distributions at each grid cell on the map of the conterminous US at a resolution of 0.5 degrees.

Reference SOC data products: We use two sets of SOC data, WISE30sec and SoilGrids250m (Hengl et al. 2017), as references to compare with spatial and vertical distributions of SOC obtained from our study over the United States. WISE30sec is an updated version of the dataset HWSD, generated by using traditional mapping methods at a resolution of 30×30 arc sec. SoilGrids250m is a global gridded soil information dataset generated by using machine learning techniques at 250 m resolution. We took data of SOC content over three depth intervals from these two datasets, 0–30 cm, 0–100 cm and 0–200 cm. All the original data with high resolution were resampled to a resolution of 0.5×0.5 degrees.

MODEL REPRESENTATION OF SOC CONTENT ACROSS OBSERVATION SITES

The original CLM5 model with default parameterization presents significant geographical biases on the estimation of SOC content in comparison with observations. Modeled SOC in the grid cell in which the site of observation was located is compared with observations (Figure 37.2a). SOC storage is systematically overestimated by the original model near the east and west coasts of the U.S. but underestimated in the Midwest. The consistency between observed and modeled SOC content is low, with $R^2 = 0.32$ and $RMSE = 15.9 \text{ kgC m}^{-3}$ (Figure 37.2b and Table 37.1).

The batch data assimilation method generates the distribution of SOC from continentally homogeneous posterior distributions of parameters estimated from all the observation data at once in data assimilation. With the batch data assimilation, the mismatch between observed and modeled SOC content in the CLM5 model is moderately reduced in the north and east parts of the U.S. (Figure 37.2c). However, geographical biases in model representation of SOC are not eliminated. CLM5 optimized by the batch data assimilation still underestimates SOC storage in the Intermontane Plateaus and southern Great Plains. Meanwhile, overestimation still exists in the Great Lakes areas and the Northeast. Overall, CLM5 after optimization by the batch data assimilation explains 43% variation in the observed SOC content with $RMSE = 11.4 \text{ kg C m}^{-3}$ (Figure 37.2d and Table 37.1).

Through the deep learning model, the PRODA approach predicts the optimized parameter values at each site across the conterminous U.S. by its environmental variables. PRODA-optimized CLM5 achieves a better representation of SOC distribution compared to the batch data assimilation. Little systematic geographical biases in estimating SOC storage are observed across the study domain (Figure 37.2e). The modeled and observed SOC content are highly correlated with $R^2 = 0.62$ and $RMSE = 9.0 \text{ kg C m}^{-3}$ (Figure 37.2f and Table 37.1).

SPATIAL DISTRIBUTION OF SOC ACROSS THE CONTERMINOUS U.S.

We take point observations (Figure 37.3a–c) and estimations from WISE30sec (Figure 37.3d–f) and SoilGrids250m (Figure 37.3g–i) as references

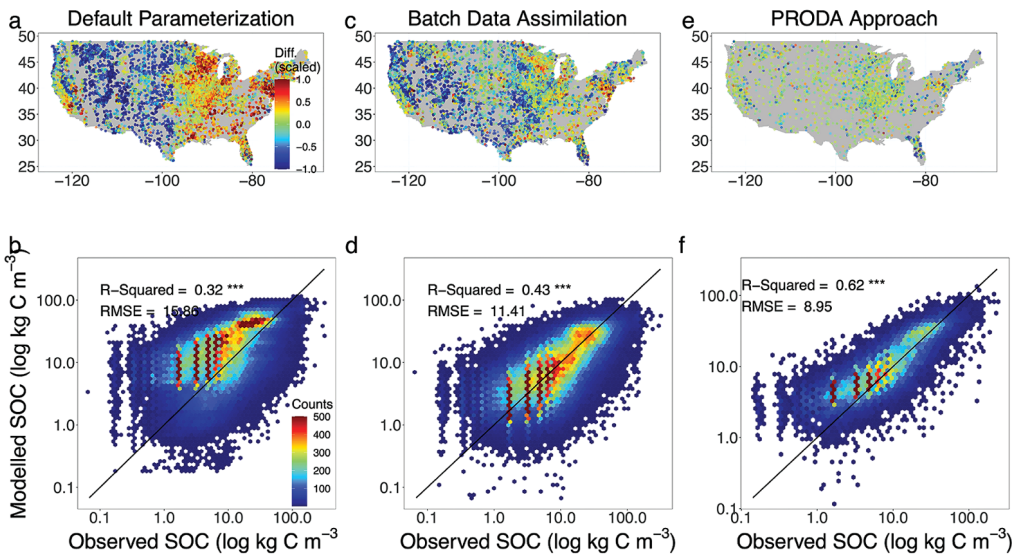


Figure 37.2. The agreement between observed and modeled SOC content with different approaches. SOC estimates modeled by CLM5 were extrapolated to the depths of observations to evaluate model performance. The upper panel indicates the deviation of the modeled SOC storage from the observation of the whole profile for each site. The lower panel shows the results of linear regression between observed and modeled vertical SOC content at different depths in different methods. In calculating the deviation of modeled SOC storage from observations, for better presentation, the positive (overestimation) and negative (underestimation) discrepancy between the observed and modeled SOC content were scaled based on the 95% quantile of the positive discrepancy and 5% quantile of the negative discrepancy, respectively. Meanwhile, only the results of the testing set were presented in PRODA approach.

TABLE 37.1
Performance of CLM5 in representing SOC distribution under different approaches

Method	Model Performance	
	R ²	RMSE (kg C/m ³)
Default CLM5	0.32	15.86
Batch Data Assimilation	0.43	11.41
PRODA Approach	0.62	8.95

Note: R² is the coefficient of determination from linear regression between the observed and modeled SOC content. RMSE is the root mean square error.

to compare the SOC estimations by CLM5 with default parameterization, optimized parameterization after the batch data assimilation, and the PRODA approach. At the continental scale, the reference data suggest large volumes of SOC in the northeast and northwest of the conterminous U.S. The magnitude of SOC content in these regions can be as high as 30 kg C m⁻² for the 0–200 cm depth interval. Meanwhile, a decreasing gradient

of SOC from the northeast to the southwest is observed. High SOC exists in areas across the Great Plains, extending from Texas to the Great Lakes.

The default CLM5 model (Figure 37.3j–l) captures the continental SOC content gradient from the northeast to the southwest but fails to reproduce sub-regional features of SOC distribution in the Great Plains. Meanwhile, SOC content in the east and northwest estimated by the original CLM5 is significantly higher than that indicated by the reference data. After optimization by the batch data assimilation, CLM5 reproduces the continental SOC gradient from the northeast to the southwest with reasonable values (Figure 37.3m–o). However, high SOC content in the Great Plains is still not well represented. The PRODA approach performs best overall, helping achieve the most realistic spatial SOC distribution (Figure 37.3p–r) in comparison with observations (Figure 37.3a–c) and data products (Figure 37.3d–i). In addition to capturing the continental SOC distribution pattern, the PRODA-optimized CLM5 presents more accurate subregional SOC distribution patterns in the Great Plains.

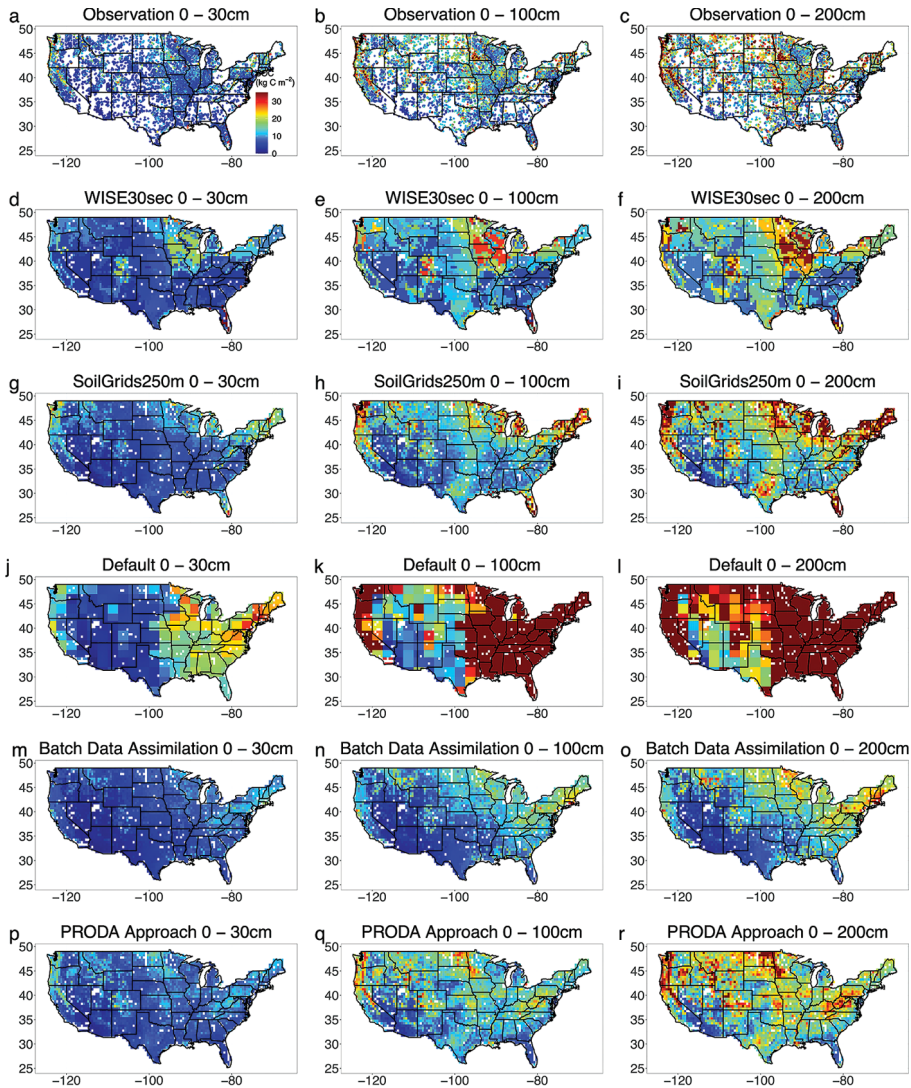


Figure 37.3. Modeled spatial SOC distributions in three depth intervals across the conterminous U.S. by different approaches and datasets.

VERTICAL DISTRIBUTION OF SOC ACROSS THE CONTERMINOUS U.S.

We take results from WISE30sec and SoilGrids250m as references in estimated SOC stocks at different depth intervals (Figure 37.4). For the first 2-meter soil, WISE30sec suggests 243 PgC and SoilGrids250m estimates 269 PgC stored as SOC. Along the soil depth, WISE30sec suggests 98 PgC at 0–30 cm depth, 81 PgC at 30–100 cm, and 64 PgC at 100–200 cm. SoilGrids250m estimates 102, 86, and 81 PgC at the same three depth intervals, respectively.

The original CLM5 model with default parameterization substantially overestimates SOC stocks

in comparison with the references at all three soil depths (Figure 37.4). Compared with the references, the overestimation becomes stronger with increasing soil depth. Both the batch data assimilation and the PRODA approach help CLM5 estimate more reasonable SOC storage compared with the original CLM5 model. We estimate 165 PgC using the batch data assimilation and 246 PgC for the first 2-meter soils using the PRODA approach.

For different vegetation types, the PRODA approach presents more accurate estimations of the vertical SOC distribution than the batch data assimilation (Figure 37.5). CLM5 underestimates the SOC content in the evergreen forest, shrubland,

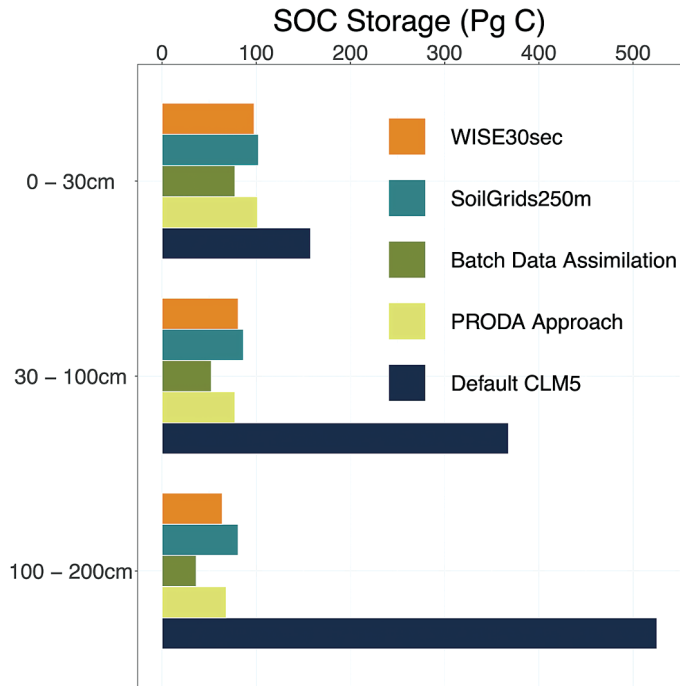


Figure 37.4. SOC storage across the conterminous U.S. at different depths estimated by different approaches and data sources.

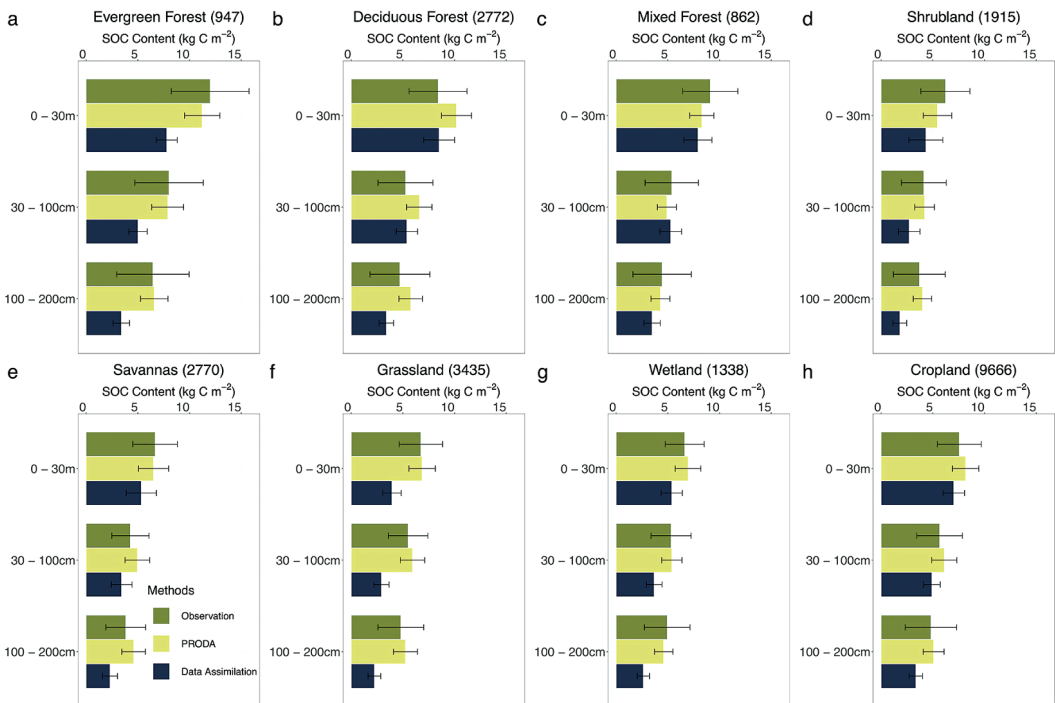


Figure 37.5. SOC storage for different vegetation types across the conterminous U.S. at different depths estimated by different approaches and data sources. The number in parentheses after vegetation type is the number of sites with that vegetation in the dataset used in this study. The error bars indicate ± 0.5 standard deviation.

savanna, grassland and wetland regions after the optimization by the batch data assimilation. The PRODA-optimized CLM5, in contrast, presents the least biased estimations in comparison with observations at all depth intervals in the aforementioned regions.

TOWARD MORE REALISTIC REPRESENTATIONS OF SOC DISTRIBUTION

This chapter has systematically explored the significance of spatially heterogeneous parameterization for the adequate prediction of SOC distribution in Earth system models, with CLM5 as a representative case. The results support the PROcess-guided deep learning and DATA-driven modelling (PRODA) as a promising approach to optimize model representation of SOC, utilising the explanatory power implicit in immense observational data. PRODA considers biogeochemical processes in the soil carbon cycle while preserving strong big data analysis ability to integrate soil data into complex models. We compared the PRODA-optimised SOC representation by CLM5 with the default model simulation and the results optimized by batch data assimilation and conclude that PRODA helped CLM5 achieve the most accurate SOC representation. Indeed, no better fit to reference data on SOC has ever been simulated by process-based models.

In the past decades, different approaches have been developed for representation of SOC distribution (Figure 37.6). Soil scientists collect soil data and develop mechanistic understanding of soil carbon cycling from field observations or experiments. The simulation modeling approach conceptualizes those mechanisms into mathematical equations and strives to simulate SOC according to process understanding. Notwithstanding the detailed description of carbon cycle processes, the models struggle to realistically simulate SOC distribution. Such unrealistic model simulations mainly arise from inadequate parameterization. Parameters that represent critical processes of the soil carbon cycle in the real world are not sufficiently constrained with widely distributed observational data. Therefore, it is difficult for process-based models to accurately represent SOC distributions. In our example, CLM5 with default parameter values substantially overestimates the total SOC storage of the conterminous U.S. and presents strong geographical biases in the representation of SOC distribution.

Batch data assimilation provides a way of incorporating observational data information into the process model to improve SOC simulation. Such data-driven optimization harmonizes site-level data information as a whole to adjust the parameter values for better representation of the SOC. We

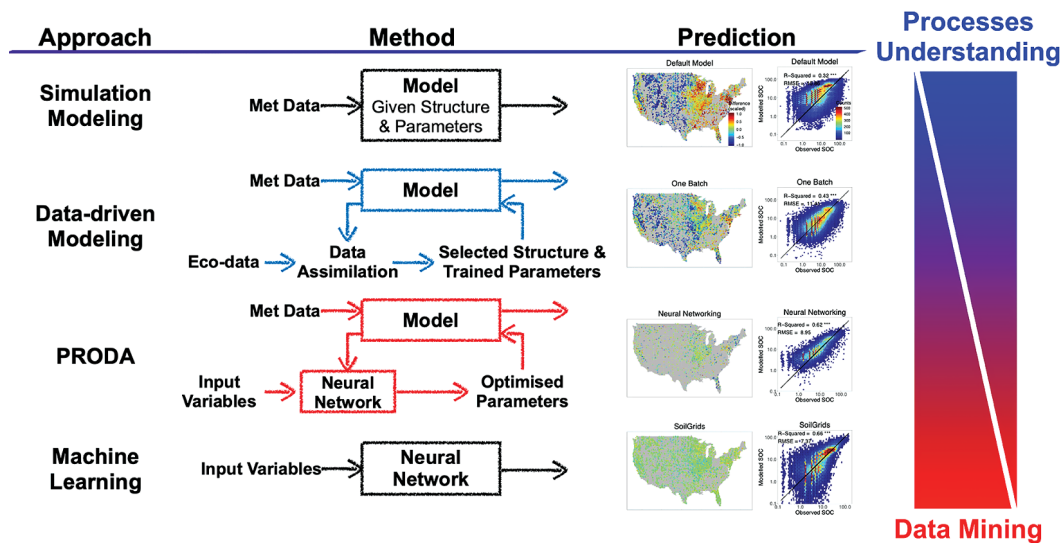


Figure 37.6. Schema of different approaches to represent SOC distributions. The PRODA approach benefits from both process understanding (as featured by simulation modelling) with the real-world information brought out by big data analysis from machine learning. The latter is primarily to obtain accurate representations of the spatial distribution of SOC and its underlying mechanisms.

have shown in the example study that the optimized CLM5 with data assimilation successfully corrects the considerable overestimation of total carbon storage across our study domain.

In terms of representing the spatial variability of SOC, however, batch data assimilation fails to capture the spatial variability of observed SOC. The spatially invariant parameter values optimized from the batch data assimilation approach are insufficient in describing the heterogeneity of SOC distribution at large scales. In our example study, geographical bias still exists after the optimization by the batch data assimilation.

The PRODA approach solves the issue of geographical bias by using a deep learning model to first fully estimate parameters at the site level using the data assimilation and then upscales the site-level estimates of parameters to the whole U.S. continent. The spatially varying parameter values retrieved from the PRODA approach contribute to a more accurate model representation of SOC across the range of ecosystem types (vegetation class, soil type, geology etc) across the continent. PRODA-optimized CLM5 simulates the most realistic SOC distribution ever simulated by process models. The high agreement between observed and modeled SOC content ($R^2 = 0.623$ across the conterminous U.S.) achieved by the PRODA approach is comparable with that for harmonization mapping in SoilGrids250m by machine learning ($R^2 = 0.635$ across the globe) (Hengl et al. 2017), and greater than the agreement between separate gridded empirical data products (Wu et al. 2019).

More importantly, the PRODA approach paves the way for more mechanistic understanding of the soil carbon cycle from big data analysis with

machine learning. Machine learning alone is good at accurately describing SOC distribution, yet previous applications used in digital soil mapping focus only on the complex statistical relationship between environmental variables and SOC. The PRODA approach not only precisely maps SOC distributions but also provides the spatial patterns of different mechanisms (as represented by different parameters) of the soil carbon cycle. In the future, disentangling how these mechanisms vary with environments and quantifying their importance to SOC storage will be essential for understanding terrestrial carbon dynamics and their feedbacks to climate change.

SUGGESTED READING

Tao, F., Z. Zhou, Y. Huang, Q. Li, X. Lu, S. Ma, X. Huang, Y. Liang, G. Hugelius, L. Jiang, R. Doughty, Z. Ren, and Y. Luo. 2020. Deep Learning Optimizes Data-Driven Representation of Soil Organic Carbon in Earth System Model Over the Conterminous United States. *Frontiers in Big Data* 3, 17.

QUIZZES

1. What is the main difference, in terms of parameterization scheme, between the batch data assimilation and the PRODA approach as described in this chapter?
2. Describe the input and output of the deep learning model in the PRODA approach?
3. What is the advantage of the PRODA approach in comparison to conventional machine learning methods in representing SOC distributions?

CHAPTER THIRTY-EIGHT

Practice 10

DEEP LEARNING TO OPTIMIZE PARAMETERIZATION OF CLM5

Feng Tao

Tsinghua University, Beijing, China

CONTENTS

- Rationale of Estimating Parameter Values by a Deep Learning Model / 329
- What Is a Neural Network? / 330
- Hyperparameters in the Neural Network / 331
- Tuning the Neural Network for Better Performance / 333
- PRODA Versus Data Assimilation Alone for Optimized SOC Distributions in CLM5 / 334

This practice offers guidance on how to use the PROcess-guided deep learning and DATA-driven modeling (PRODA) approach to integrate observations with the biogeochemical module of the Community Land Model version 5 (CLM5) to best represent regional soil organic carbon distributions. Over three exercises, we focus on how to build, train, and tune a deep learning model in the PRODA approach to predict parameters estimated from site-level data assimilation. Readers can use either the CarboTrain platform, or the original codes (via <https://www2.nau.edu/luo-lab/?workshop>) to explore different deep learning options and to understand and modify the optimization methods.

RATIONALE OF ESTIMATING PARAMETER VALUES BY A DEEP LEARNING MODEL

In Chapter 37, we discussed the performance of different approaches that assimilate observations into CLM5 to best represent soil organic carbon (SOC) stocks across the conterminous United States (Tao et al. 2020). We concluded that the PRODA approach outperforms data assimilation alone by fully interpreting the spatial heterogeneity of parameters. In the PRODA approach, parameter values are first

retrieved where the observation resides through the Markov Chain Monte Carlo (MCMC) method and then upscaled to the region by a deep learning model. Eventually, we apply the parameter values predicted by the trained deep learning model to CLM5 to simulate SOC stock and distributions (see Chapter 37 for details of the PRODA workflow).

We use various environmental variables to predict the parameter values by a neural network in the deep learning model. The rationale behind this procedure is that parameters in the process-based model can be expected to vary with environmental conditions over space and time in order to account for changing properties of evolving ecosystems and unresolved processes (Luo and Schuur 2020). The relationships between parameter values and environmental variables, however, are often not easily identified so as to be explicitly represented in model structure. Therefore, we introduce a deep learning model to explore such complex relationships. Specifically, we set the local environmental variables as the input and the estimated parameter values in the site-level data assimilation as the prediction target in a deep learning model. Then, the deep learning model is trained to best predict the parameter values based on input environmental variables (Figure 38.1).

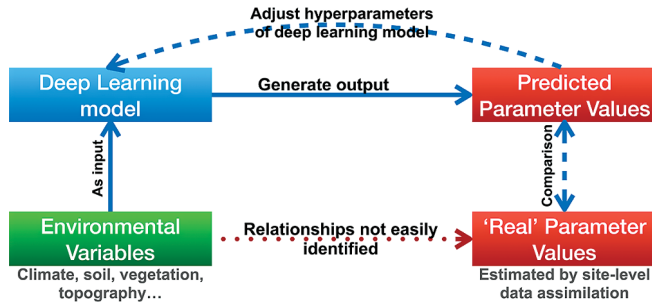


Figure 38.1. Schematic diagram of using a deep learning model to predict parameter values. The deep learning model is used to interpret relationships between environmental variables and parameter values retrieved from the site-level data assimilation.

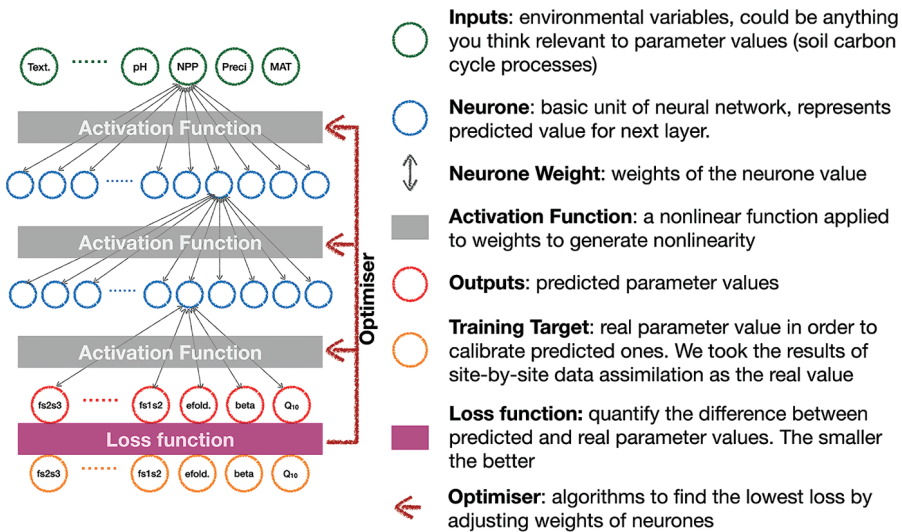


Figure 38.2. Basic elements in a typical multilayer neural network.

WHAT IS A NEURAL NETWORK?

Chapter 36 introduced the basic concepts of machine learning/deep learning. In this practice, we focus on how to build and optimize a deep learning model that is structured by a multilayer neural network. Four elements together construct the skeleton of a typical multilayer neural network, which are: **inputs**, **outputs**, **neurones**, and **neurone weights** (Figure 38.2). In this practice, the inputs are environmental variables that are relevant to the parameters in CLM5, such as climatic, edaphic, and vegetation variables. The outputs are the parameter values we predict using the trained neural network, which will be further applied in CLM5 to simulate SOC distributions. The neurones are the basic unit of a neural network. They distribute at each hidden layer of the neural network (Figure 38.2), receive information from the inputs or the previous layer,

and generate all possible predictions for the next layer or as the final outputs. Finally, the neurone weights are the weights assigned to the predictions of each neurone. We get the final neural network predictions by combining all the predictions by different neurones with their weights.

In addition to the four basic elements of a neural network, we use the **activation function** to generate the neurone weights. We generally use nonlinear activation functions to enable the neural network to explore the nonlinearities between the inputs and the final outputs.

We train the neural network to best predict our **training target** (i.e., the parameter values retrieved from the site-level data assimilation) via a certain optimization algorithm. A **loss function** is used to quantify the difference between the predicted parameter values from the “real” parameter values retrieved from the site-level data assimilation.

The lower the loss function value is, the closer the predictions are from the “real” parameter values, and the better the model performance is.

The loss function value alone, however, cannot initiate the optimization. We need an algorithm to determine how to adjust the neural network according to the results of the loss function so that the final predictions can best fit the training target. In the neural network, we have the **optimizers** to adjust the weights of neurones according to the results of the loss function. Ideally, after sufficient training, the optimizer will eventually lead the neural network to reach a point where the loss function reserves the lowest value it can pursue. We regard the neural network at this point as reaching its global optimum. Predictions by the neural network will consequently be the closest to the training targets.

HYPERPARAMETERS IN THE NEURAL NETWORK

Hyperparameters are the parameters whose values control the training process in the neural network. Hyperparameters that control the shape of a neural network include the number of hidden layers and the neuron numbers of each hidden layer. Hidden layer numbers determine the depth of a neural network. Neuron numbers of each hidden layer, at the same time, control the width of the neural network. Choices of hidden layer numbers and neuron numbers are largely empirical. You can try neural networks with different shapes and choose the one that can best predict your training target.

The epoch number determines how many times the deep learning algorithm will go through the whole dataset for training. During each epoch, the neural network can propose a set of neuron weights and adjust them by the optimizer according to the loss function results. The number of epochs ranges from hundreds to thousands in

different deep learning applications. You may try different numbers to find the best epoch number that allows the loss function value to be minimized, so that the neural network can accurately predict the training target.

In addition, the batch size defines how many training data you want to use as one batch when working through the whole training dataset in each epoch. For example, for a training set with 10,000 samples, if we set the batch size as 50, then it will take 200 iterations ($\text{iteration number} = \frac{\text{sample size}}{\text{batch size}}$) to go through the whole training set in each epoch of optimization. Possible choices of batch size vary from 1 to the size of the whole training set. Setting batch size as 16, 32, or 64 would be a plausible start.

We also need to decide several hyperparameters that control the optimization process in the neural network before initiating the training. You may need to specify the loss function, activation function, and the optimizer. We use the Keras package in Python to build and train the neural network in this practice. Keras provides multiple choices for these hyperparameters. The loss function can be the mean squared error (expressed as “mean_squared_error” in Keras) or other functions that can quantify the difference between predicted and target values. The activation function can be the Rectified Linear Unit (expressed as “ReLU” in Keras), the hyperbolic tangent function (expressed as “tanh” in Keras), the sigmoid function (expressed as “sigmoid” in Keras), other activation functions that are pre-defined by Keras, or a function that you define yourself. Keras provides several options for the optimizer. Choosing “Adam”, “Adadelta”, or “RMSprop” would generally be a good start. You can refer to the website <https://keras.io/api/> for more possibilities of the hyperparameters you can use in Keras.

EXERCISE 1

Building and training a neural network that uses environmental variables to predict parameter values in CLM5. Follow the instructions in CarboTrain:

- Select **Unit 10**
- Select **Exercise 1**
- Select **Output Folder**

d. Open Source Code

- Read section “Setting NN Structure” to get familiar with hyperparameters that we used in this neural network.
- Run Exercise**
- Check results in your Output Folder. Four figures will be generated (Figure 38.3). The figure in `loss_nn.png` describes the changes of the loss function value

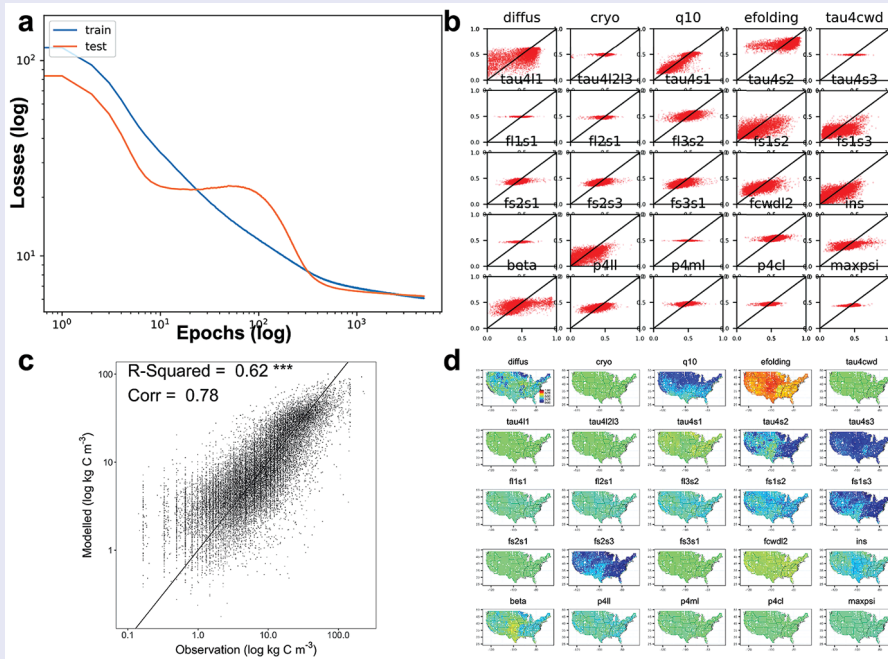


Figure 38.3. Output figures in Exercise 1.

in the training and validation set with increased epochs. `para_nn.png` describes how well the trained neural network predicts different parameters. `nn_obs_vs_mod.jpeg` indicates how well the CLM5 model can simulate SOC after applying the predicted parameter values from the trained neural network. `map_para_us.jpeg` describes the predicted parameter maps from the trained neural network.

Note: You can also choose your own computer to execute the exercise. You may download the code package via <https://www2.nau.edu/luo-lab/?workshop>, and follow the instructions:

- Read and operate the Python code via “yourpath/nau_training_proda/nn_clm_cen.py”.
- After the neural network training, operate the R code via “yourpath/nau_training_proda/NN_Indi_MAP_Project_CLM_CEN.R”.
- Operate the R code via “yourpath/nau_training_proda/nn_para_map.R”.

QUESTIONS:

- How many layers are set in the default neural network? How many neurones are distributed at each layer?
- What is the activation function used in the default neural network?
- How many epochs will the default neural network run in training? What is the batch size?
- What is the loss function used in the default neural network? Can you justify the reason why we need a loss function?
- In Figure 38.3a, which two hyperparameters in the neural network control the calculation of loss value and changes to the loss value?
- In Figure 38.3b, which parameters are predicted well by the trained neural network, and which not? Why do you think the neural network predicts some parameters well, and others not that well?

According to what you have learnt in Chapter 37, what will the results presented in Figure 38.3d be used for?

TUNING THE NEURAL NETWORK FOR BETTER PERFORMANCE

Setting up basic neural network structures and hyperparameters does not guarantee satisfactory model performance in prediction. We need to tune the neural network for better performance. We will briefly introduce some common procedures in tuning the neural network (Figure 38.4).

Suppose you find the predicted parameter values cannot fit well with the targets after training the neural network. In this case, you are recommended to first try a new model structure with more hidden layers (deeper neural network) and/or more neurons for each hidden layer (wider neural network). Expanding the depth and/or width of the neural network will generally increase its ability to interpret more complex relationships between the input environmental variables and the target parameter values. Meanwhile, you can also consider applying different activation functions or optimizers.

We may sometimes encounter the overfitting problem. Overfitting happens during training of the neural network when the loss value of the validation set stops decreasing with the loss value of the training set but begins to increase (Figure 38.5). In such a case, even though the final prediction of the training set may fit well with the target, the trained neural network cannot make precise predictions in the validation set (see a detailed discussion about the overfitting in Chapter 36). To avoid overfitting, we can adopt early stopping to stop training the neural network at reasonable epochs. After early stopping, the loss values of both the validation and training sets stay low and thus ensure the robustness of neural network predictions in both datasets.

Regularization is another option to prevent overfitting in training the neural network. Regularization applies penalties to the weights of neurones to avoid the predictions of the trained neural network relying too much on the performance of one or some small group of neurones. Regularization is a more advanced option in tuning the neural network. We may not touch on it too much in this book. If you are interested, you can refer to the website at <https://machinelearningmastery.com/how-to-reduce-overfitting-in-deep-learning-with-weight-regularization/>.

Dropout offers a further option if we do not want the performance of a small group of neurones to matter too much in the final prediction of the trained neural network. The dropout option allows us to randomly permute some certain percent of neurones in each epoch of optimization. The neural network will then be trained to not depend too much on any specific neurones in prediction, thereby improving its robustness. If you check the default setting in Exercise 1, you will find we used the dropout option in the neural network training.

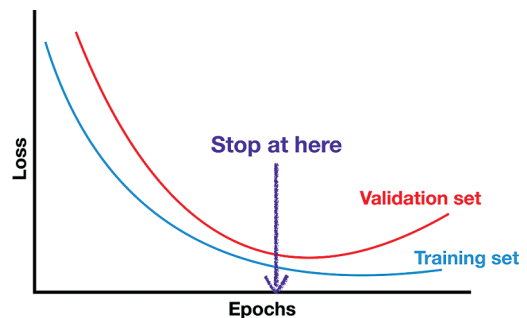


Figure 38.5. Overfitting in neural network training and early stopping option.

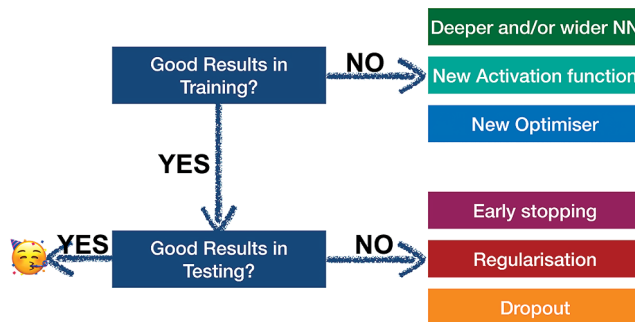


Figure 38.4. Tips on tuning the neural network for better model prediction performance. Reproduced from Lee (2016).

EXERCISE 2

Tuning the neural network used in Exercise 1. Follow the instructions in CarboTrain:

- Select **Unit 10**
- Select **Exercise 2**
- Select **Output Folder**
- Change one or more hyperparameter values. e.g., select and change the Optimizer to “adam”.
- Run Exercise**
- Check results in your Output Folder. Four figures as described in Exercise 1 will be generated.

Note: If you would like to use your own computer to do the exercise, you may follow the instructions:

- Open the python code via “yourpath/nau_training_proda/nn_clm_cen.py”.
- Change hyperparameter values at corresponding locations
- Operate “nn_clm_cen.py”
- After the neural network training, operate the R code via “yourpath/nau_training_proda/NN_Indi_MAP_Project_CLM_CEN.R”.
- Operate the R code via “yourpath/nau_training_proda/nn_para_map.R”.

QUESTIONS:

How do the output figures change compared to those generated in Exercise 1? Can you explain the reasons for these changes?

PRODA VERSUS DATA ASSIMILATION ALONE FOR OPTIMIZED SOC DISTRIBUTIONS IN CLM5

In Exercise 1 and 2, we have introduced how to build, train, and tune a neural network to best predict parameter values by environmental variables. The deep learning model is the core part of the PRODA approach. We optimize SOC representation in a process model by fully interpreting the environmental dependencies of its parameters through a deep learning model.

Data assimilation alone can also utilize multi-site observations to retrieve parameter values so that the biogeochemical model can better represent SOC distributions. Instead of optimizing parameter values at each observational site as in the PRODA approach, the retrieved parameter values with the data assimilation alone are spatially invariant. Exercise 3 will compare results of CLM5 in terms of the SOC representation when using PRODA to set parameter values, compared with data assimilation alone.

EXERCISE 3

Comparing SOC representations in CLM5 between PRODA and data assimilation alone. Follow the instructions in CarboTrain:

- Select **Unit 10**
- Select **Exercise 3**
- Select **Neural Network Task Folder**
- Select **DA alone Task Folder**
- Select **Output Folder**
- Run Exercise**
- Check results in your Output Folder. Three figures will be generated

(Figure 38.6). Figure `mod_vs_obs_nn.jpeg` shows the agreement between the PRODA-optimized and observed SOC. Figure `mod_vs_obs_ob.jpeg` describes the agreement between retrieved and observed SOC using data assimilation alone. Figure `soc_map.jpeg` shows the SOC distributions simulated by CLM5 at different depths across the conterminous United States after optimization by the PRODA approach, and following parameter estimation by data assimilation alone.

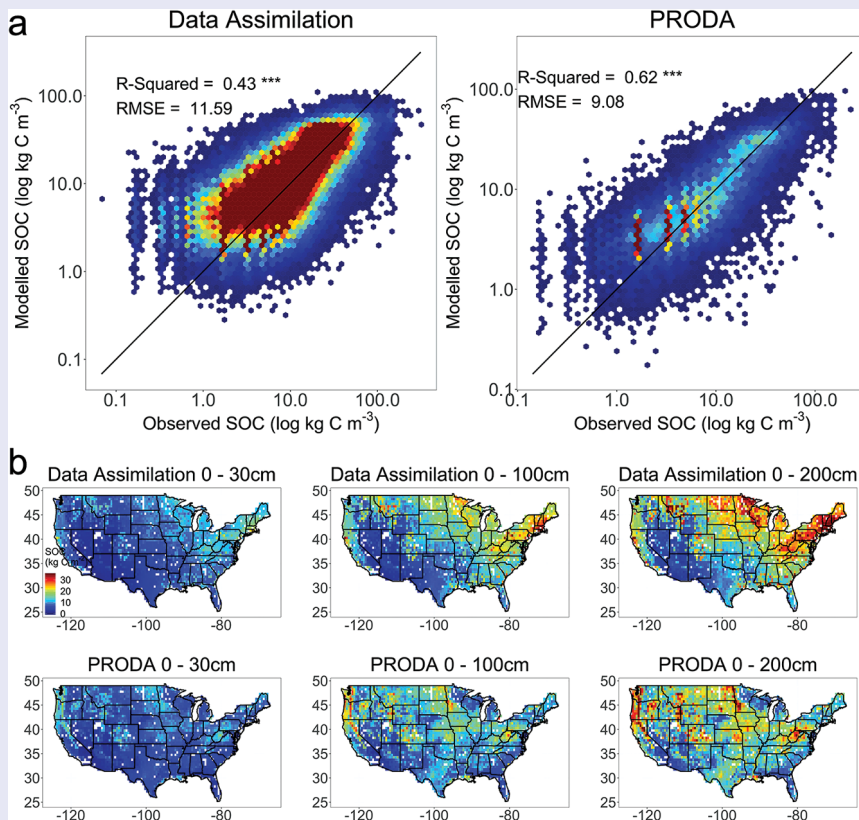


Figure 38.6. Output figures in Exercise 3. (a) Agreement between model simulations and observations; (b) SOC distributions at different depth across the conterminous United States simulated by CLM5 with parameter estimation by data assimilation alone v. PRODA.

If you would like to use your own computer to do the exercise, you may follow the instructions:

- a. Open the R code via "yourpath/nau_training_proda/NN_Indi_MAP_Project_CLM_CEN.R"
- b. Set variable "ifnn" as ifnn = 1 (PRODA results)
- c. Operate "NN_Indi_MAP_Project_CLM_CEN.R"
- d. Set variable "ifnn" as ifnn = 2 (data assimilation alone results)
- e. Operate "NN_Indi_MAP_Project_CLM_CEN.R" again
- f. Open the R code via "yourpath/nau_training_proda/Global_Projection_NN_CLM_CEN.R"
- g. Set variable "is_nn" as is_nn = 1 (PRODA results)
- h. Operate "Global_Projection_NN_CLM_CEN.R"
- i. Set variable "is_nn" as is_nn = 0 (data assimilation alone results)
- j. Operate "Global_Projection_NN_CLM_CEN.R" again
- k. Open and operate the R code via "yourpath/nau_training_proda/Different_Method_obs_vs_mod.R"
- l. Open and operate the R code via "yourpath/nau_training_proda/different_method_soil_map.R"

QUESTION:

Which approach performs better in representing SOC distribution? Why?

To summarize, this practice has explored the technical details of the PRODA approach in optimizing parameter values and retrieving SOC distributions. The deep learning model is the core part of the PRODA approach. The first two exercises illustrate the basic components in a neural network (Exercise 1) and how to tune a configured neural network for better performance (Exercise 2). In Exercise 3, we applied CLM5 to simulate the

SOC distribution across the conterminous United States after parameter estimation by data assimilation alone and further parameter optimization by the PRODA approach. We compared the agreement between simulations and observations after the two approaches. The results show the advantage of using the PRODA approach to optimize SOC distribution in biogeochemical models.

APPENDIX 1

Matrix Algebra in Land Carbon Cycle Modeling

Ye Chen

Northern Arizona University, Flagstaff, USA

The purpose of this appendix is to deliver the necessary matrix algebra foundation you will need to understand the matrix model of the land carbon cycle, introduced in Chapter 1. If you have taken the undergraduate level of matrix algebra or above, you may skip this appendix. Please note that some important topics in the following are presented in a simple way and they could be easily extended into longer sections. Professor Gilbert Strang's Linear Algebra lecture in MIT OpenCourseWare (see Recommended Reading) is a good source for extra learning material and self-study.

MOTIVATIONS

Many processes can be described using a dynamical system. In one type of dynamical system, the state \vec{x}_{n+1} of a system at time $n + 1$ can be derived from the state \vec{x}_n at time n using a transition matrix A :

$$\vec{x}_{n+1} = A\vec{x}_n$$

See the following problem for more details.

Problem 1

Consider a simplified carbon transfer model in which 3% of the carbon in the fast soil pool moves to the slow soil pool each year while 95% of the fast soil pool stays the same, and 1% of the slow soil pool moves to the fast soil pool while 97% of the slow soil pool stays the same with no other influences on the two pools. What are the pool sizes after 10 years? 50 years? 100 years?

This question will be fully answered at the end of this appendix. To see how linear

algebra can help us, we will do some basic analysis in this section.

Note that there is 2% lost in each pool when transfer occurs every year. To simplify the problem, we are not going to track what happens with that carbon once it exits the two pools.

Let $\vec{x}_n = \begin{bmatrix} f_n \\ s_n \end{bmatrix}$ be the status of the two

pools at n th year. That is, f_n is the size of the fast soil pool at n th year, and s_n is the size of the slow soil pool at n th year. Then

$$f_{n+1} = 0.95f_n + 0.01s_n$$

$$s_{n+1} = 0.03f_n + 0.97s_n$$

First of all, this linear system can be written down in the matrix form using the knowledge from sections 2 and 4 below,

$$\begin{bmatrix} f_{n+1} \\ s_{n+1} \end{bmatrix} = \begin{bmatrix} 0.95 & 0.01 \\ 0.03 & 0.97 \end{bmatrix} \begin{bmatrix} f_n \\ s_n \end{bmatrix}$$

or equivalently, $\vec{x}_{n+1} = A\vec{x}_n$, where A is the 2×2 coefficient matrix.

Given the pool size of the first year, $\vec{x}_1 = \begin{bmatrix} f_1 \\ s_1 \end{bmatrix}$, one interesting question is to find out the pool size a long time after, i.e., \vec{x}_n for a large n . In this problem, $n = 10, 50, 100$. Note that

$$\vec{x}_{n+1} = A\vec{x}_n = A(A\vec{x}_{n-1}) = A^2\vec{x}_{n-1} = \dots = A^n\vec{x}_1,$$

and finding $A^n \vec{x}_1$ for large n can be computationally intensive when the dimension of A is large. However, if we can find the eigenvalues of A as introduced in section 5, then we may be able to turn the problem of finding $A^n \vec{x}_1$ with large n into finding $\lambda^n \vec{x}_1$, where λ is an eigenvalue of A .

MATRIX OPERATIONS

Basic Operations

Definition 1

A **matrix** is a rectangular array of numbers. We say that a matrix A is an $m \times n$ matrix when it has m rows and n columns, and the **dimension** of A is $m \times n$.

For example, this is a 2×3 matrix:

$$A = \begin{bmatrix} 2 & 3 & 5 \\ 4 & 1 & -9 \end{bmatrix} \quad (\text{A1.1})$$

Definition 2

A matrix is **square** if it has the same number of rows and columns.

We can use subscript notation to refer to particular entries in a matrix: the notation A_{ij} refers to the entry of matrix A in row i and column j .

Problem 2

Let A be the matrix in equation A1.1. What is A_{13} ?

Answer: The number in row 1 and column 3 is 5.

Definition 3

The **transpose** of a matrix A , denoted A^T , is the matrix A with the rows and columns switched. That is, $(A^T)_{ij} = A_{ji}$.

For example, for the matrix in Equation A1.1

$$A^T = \begin{bmatrix} 2 & 4 \\ 3 & 1 \\ 5 & -9 \end{bmatrix}$$

We can add two matrices if they have the same dimensions. We do this by adding corresponding entries. For example,

$$\begin{bmatrix} 2 & 3 \\ 4 & -1 \end{bmatrix} + \begin{bmatrix} 5 & 0 \\ -6 & 3 \end{bmatrix} = \begin{bmatrix} 2+5 & 3+0 \\ 4-6 & -1+3 \end{bmatrix} \\ = \begin{bmatrix} 7 & 3 \\ -2 & 2 \end{bmatrix}$$

Problem 3

For two matrices A and B , is $A + B$ always the same as $B + A$?

Answer: Yes.

Matrix Multiplication

To multiply a scalar number k by a matrix, simply multiply k with every element of the matrix. For example,

$$2 \begin{bmatrix} 0 & 1 & 5 \\ 2 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 10 \\ 4 & 2 & -2 \end{bmatrix}.$$

To multiply two matrices, the number of columns of the first matrix must be the same as the number of rows of the second matrix. Let's say that we have two matrices, X , which is $m \times k$, and Y , which is $k \times n$. Then their product, denoted XY , will be an $m \times n$ matrix. Here is how to determine the elements of the matrix product XY : to get $(XY)_{ij}$ (the entry in the i th row and j th column), take the i th row of X and the j th column of Y , multiply their corresponding elements, and add the results.

Suppose we have the matrices

$$X = \begin{bmatrix} 1 & -2 & 3 \\ 0 & -1 & 6 \end{bmatrix} \quad Y = \begin{bmatrix} 0 & -2 & 1 & 0 \\ 5 & 4 & 2 & 6 \\ 1 & -3 & 0 & -1 \end{bmatrix}$$

X has dimensions 2×3 , and Y has dimensions 3×4 . Since the number of columns of X equals the number of rows of Y , we can multiply them, and the result will be a 2×4 matrix.

Now, to determine the entry in the first row and first column of the product XY , we look at the first row of X and the first column of Y , here shown highlighted in blue and red:

$$X = \begin{bmatrix} 1 & -2 & 3 \\ 0 & -1 & 6 \end{bmatrix} \quad Y = \begin{bmatrix} 0 & -2 & 1 & 0 \\ 5 & 4 & 2 & 6 \\ 1 & -3 & 0 & -1 \end{bmatrix}$$

We now take these two lists of three numbers, multiply them element-by-element, and add the results:

$$1 \cdot 0 + -2 \cdot 5 + 3 \cdot 1 = -7$$

So far, we know that the matrix product XY looks like this:

$$XY = \begin{bmatrix} -7 & ? & ? & ? \\ ? & ? & ? & ? \end{bmatrix}$$

Now let's compute $(XY)_{12}$ (highlighted in red above). Since we are trying to compute the entry of XY in the first row and second column, we take the first row of X , and the second column of Y :

$$X = \begin{bmatrix} 1 & -2 & 3 \\ 0 & -1 & 6 \end{bmatrix} \quad Y = \begin{bmatrix} 0 & -2 & 1 & 0 \\ 5 & 4 & 2 & 6 \\ 1 & -3 & 0 & -1 \end{bmatrix}$$

Multiplying them pairwise and then adding yields

$$1 \cdot -2 + -2 \cdot 4 + 3 \cdot -3 = -19$$

Now XY looks like

$$XY = \begin{bmatrix} -7 & -19 & ? & ? \\ ? & ? & ? & ? \end{bmatrix}$$

Problem 4

Finish computing the matrix product XY .

$$\text{Answer: } XY = \begin{bmatrix} -7 & -19 & -3 & -15 \\ 1 & -22 & -2 & -12 \end{bmatrix}$$

Problem 5

Use the matrices X, Y defined in the previous example, and also

$$A = \begin{bmatrix} 2 & 4 \\ -1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 3 \\ 2 & 2 \end{bmatrix}$$

Compute each of the following if it is defined, or say undefined otherwise.

1. YX 2. XA 3. AB 4. BA

Answer: 1. YX is undefined. 2. XA is undefined as the dimension of X is 2×3 , and the dimension of A is 2×3 : the inner dimension does not match. 3. $AB = \begin{bmatrix} 10 & 14 \\ 1 & -1 \end{bmatrix}$. 4.

$$BA = \begin{bmatrix} -1 & 7 \\ 2 & 10 \end{bmatrix}.$$

Problem 6

Based on the results of Problem 5, is AB always the same as BA ? Explain why.

Answer: No. What you found in Problem 5 is that matrix multiplication is not commutative. This is one way in which matrices are different from real numbers. With real numbers, you are able to switch around the order of operands being multiplied. But with matrices, you have to be very careful: you cannot do this with matrices!

It turns out, however, that matrix multiplication is associative: for any matrices X, Y , and Z , as long as they have dimensions that match up properly, it is always true that $(XY)Z = X(YZ)$. That is, when doing more than one matrix multiplication, it doesn't matter which multiplication we do first, as long as we keep them in the right order. This means that we can write things like $ABCDE$ instead of $((AB)C)(DE)$ or $A(B(C(DE)))$ or $((AB)(CD))E$ since they are all the same.

Quiz 1

- If the dimension of A is 10×20 , the dimension of B has only one column, and AB is well defined. How many columns does the matrix product AB have? (Answer: The matrix AB has one column).
- If the dimension of the matrix A is 10×25 , the dimension of the matrix C is 10×20 , and $AB = C$. What is the dimension of B ? (Answer: The dimension of B is 25×20 .)

MATRIXEQUATIONS

Identity Matrix, Inverse Matrix

Problem 7

Can you find a 2×2 matrix I which, when multiplied by any other 2×2 matrix X , yields X ? That is, $IX = XI = X$ for any 2×2 matrix X . I is called the 2×2 identity matrix (sometimes also written I_2).

$$\text{Answer: } I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Problem 8

What is the 3×3 identity matrix, I_3 ? In general, what does the $n \times n$ identity matrix I_n look like?

$$\text{Answer: } I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. I_n \text{ is an } n \times n$$

matrix with the main diagonal elements being 1, and all other elements being 0.

Problem 9

There is no such thing as a 2×3 identity matrix. Why not?

Answer: To match the dimension such that $IX = XI = X$ in the definition of identity matrix, I has to be a square matrix.

Problem 10

Multiply the following two matrices:

$$A = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix}, B = \begin{bmatrix} 5 & -3 \\ -3 & 2 \end{bmatrix}$$

What do you get? Why is this interesting?

Answer: $AB = I_2$.

Definition 4

The **inverse** of a square matrix A , written A^{-1} , is a matrix which multiplied by A results in the identity matrix: $AA^{-1} = A^{-1}A = I$.

Problem 11

As it turns out, not all square matrices have an inverse. But this should not be too

surprising. Why not? (Hint: think about the inverse of 0).

Solving Matrix Equations

Matrices which have an inverse are called **invertible** matrices, and matrices which do not have an inverse are called **singular** matrices. Why do we care whether a matrix is invertible? Well, remember what you do in algebra to solve an equation like $3x = 12$: you multiply both sides by $1/3$, the inverse of 3. In the same way, inverting matrices allows us to solve matrix equations like $AX = Y$ (where A , X , and Y are all matrices). If A is invertible, we can left multiply both sides of the equation by A^{-1} to get $X = A^{-1}Y$. Note that YA^{-1} is not the solution of X by Problem 6.

Problem 12

If A , B are invertible, solve this matrix equation for X :

$$AXB = Y$$

Answer: $X = A^{-1}YB^{-1}$

Example 1

Solve this matrix equation for X :

$$Y = BC + ADKX$$

If the matrix product ADK is invertible,

$$Y = BC + ADKX$$

$$\Rightarrow ADKX = -BC + Y$$

$$\Rightarrow X = (ADK)^{-1}(-BC + Y)$$

$$\Rightarrow X = (ADK)^{-1}(-BC) + (ADK)^{-1}Y$$

Following the steps of this example, you can solve the next problem.

Problem 13

The matrix equation below is presented in Chapter 1, Equation 1.6:

$$X'(t) = B\mu(t) + A\xi(t)KX(t)$$

Can you solve it to obtain an expression for $X(t)$ if the matrix $A\xi(t)K$ is invertible, $X'(t)$, B , A , $\xi(t)$, K and $X(t)$ are matrices, and $\mu(t)$ is a scalar?

Answer: You may have noticed this equation is similar to the matrix equation in Example 1, except the notations are different. The computation would still follow the rule of matrix operations. And to solve this equation for X , just follow the steps in Example 1. The solution is

$$X(t) = (A\xi(t)K)^{-1}(-B\mu(t)) + (A\xi(t)K)^{-1}X'(t)$$

Quiz 2

Let A , C be invertible matrices. Solve $AXC + BD = Y$ for X . (Answer: $X = A^{-1}(Y - BD)C^{-1}$.)

LINEAR SYSTEM

Definition 5

A linear equation in the variables x_1, x_2, \dots, x_n is an equation that can be written in the form:

$$a_1x_1 + a_2x_2 + \dots + a_nx_n = b$$

where a_1, a_2, \dots, a_n and b are real numbers. Thus, for example,

$$2x_1 + 3x_2 + x_3 - 1x_4 = 6$$

is a linear equation in the four variables x_1, x_2, x_3, x_4 . But the following equations are not linear:

$$2x_1 + 2x_2 + x_3x_4 = 6$$

$$2x_1^2 + 3x_2 = 6$$

Definition 6

A **system of linear equations** (or **linear system**) is a set of linear equations.

The following is a simple linear system.

Example 2

Solve the linear system:

$$(1) \quad -x_1 + \frac{1}{3}x_2 = 6$$

$$(2) \quad x_1 + x_2 = 10$$

This system can be easily solved by hand using forward and backward substitution. The solution is

$$x_1 = -2, x_2 = 12.$$

Note that with the matrix multiplication, the system is equivalent to

$$x_1 \begin{bmatrix} -1 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \end{bmatrix}$$

Therefore, the solution can be written as

$$-2 \begin{bmatrix} -1 \\ 1 \end{bmatrix} + 12 \begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \end{bmatrix}$$

Here is another approach using the matrix inverse. The linear system can be written down in the matrix form:

$$\begin{bmatrix} -1 & \frac{1}{3} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \end{bmatrix}$$

We denote this matrix form as

$$Ax = b$$

To find the solution of this linear system for x , we left multiply this equation by A^{-1} if it exists, then

$$x = A^{-1}b$$

EIGENVALUES AND EIGENVECTORS

Definition 7

Let A be a square matrix.

A vector \vec{x} is an **eigenvector** of A if $\vec{x} \neq 0$ and there is a scalar λ such that $A\vec{x} = \lambda\vec{x}$.

A scalar λ is an **eigenvalue** of A if there is a vector $\vec{x} \neq \vec{0}$ such that $A\vec{x} = \lambda\vec{x}$.

Remark 1 By this definition, if λ is an eigenvalue of A and \vec{x} is a corresponding eigenvector, then we must have the following:

$$A^n \vec{x} = A^{n-1}(A\vec{x}) = A^{n-1}(\lambda\vec{x}) = \lambda A^{n-1}\vec{x} = \lambda A^{n-2}(A\vec{x}) = \dots = \lambda^n \vec{x}$$

Next, we are going to apply Definition 7 to verify the eigenvalues and eigenvectors of a matrix.

Example 3

Confirm that both $\vec{v}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and $\vec{v}_2 = \begin{bmatrix} 1/3 \\ 1 \end{bmatrix}$ are eigenvectors for the matrix

$$A = \begin{bmatrix} 0.95 & 0.01 \\ 0.03 & 0.97 \end{bmatrix}, \text{ and find the corresponding eigenvalues.}$$

Solution: As

$$A\vec{v}_1 = \begin{bmatrix} 0.95 & 0.01 \\ 0.03 & 0.97 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.94 \\ 0.94 \end{bmatrix} = 0.94\vec{v}_1$$

by the definition of eigenvalue, \vec{v}_1 is an eigenvector for A with corresponding eigenvalue $\lambda_1 = 0.94$.

Similarly, it is easy to verify that \vec{v}_2 is an eigenvector for A with corresponding eigenvalue $\lambda_2 = 0.98$.

In this example, the eigenvectors are already given, so it is relatively easy to determine the eigenvalues. In general, to determine the eigenvalues and eigenvectors by hand could be very hard or even impossible when the dimension of A is large. Luckily, modern programming environments like MATLAB and Python provide functions for this purpose.

Example 4

Let us continue finding the solution for Problem 1. Let $\vec{x}_1 = \begin{bmatrix} 6 \\ 10 \end{bmatrix}$ and compute the pool sizes after 50 years.

Solution: By Example 2, we have

$$\begin{bmatrix} 6 \\ 10 \end{bmatrix} = -2 \begin{bmatrix} -1 \\ 1 \end{bmatrix} + 12 \begin{bmatrix} 1/3 \\ 1 \end{bmatrix}, \text{ which is}$$

equivalent to $\vec{x}_1 = -2\vec{v}_1 + 12\vec{v}_2$. So,

$$A^n \vec{x}_1 = A^n(-2\vec{v}_1 + 12\vec{v}_2) = -2A^n\vec{v}_1 + 12A^n\vec{v}_2$$

By Remark 1 and Example 3, $A^n\vec{v}_1 = \lambda_1^n\vec{v}_1$, $A^n\vec{v}_2 = \lambda_2^n\vec{v}_2$. We then have

$$A^n \vec{x}_1 = -2 \cdot 0.94^n \vec{v}_1 + 12 \cdot 0.98^n \vec{v}_2$$

By Problem 1, to determine the pool sizes after 50 years, it is equivalent to find \vec{x}_{51} , where $\vec{x}_{51} = A^{50}\vec{x}_1$. By the analysis we just did, we know that this can be computed easily:

$$\begin{aligned} \vec{x}_{51} &= A^{50}\vec{x}_1 = -2 \cdot 0.94^{50}\vec{v}_1 \\ &\quad + 12 \cdot 0.98^{50}\vec{v}_2 \approx \begin{bmatrix} 1.55 \\ 4.28 \end{bmatrix} \end{aligned}$$

Therefore, after 50 years, the size of the fast pool is 1.55, and the size of the slow pool is 4.28.

Quiz 3

Continue with Problem 1. Let $\vec{x}_1 = \begin{bmatrix} 2 \\ 6 \end{bmatrix}$.

Compute the pool sizes after 20 years. (Answer:

As $\vec{x}_1 = 6\vec{v}_2$, we then have $A^n\vec{x}_1 = 6 \cdot 0.98^n\vec{v}_2$. So,

$$\vec{x}_{21} = 6 \cdot 0.98^{20}\vec{v}_2 \approx \begin{bmatrix} 1.34 \\ 4.01 \end{bmatrix}. \text{ After 20 years, the size}$$

of the fast pool is 1.34, and the size of the slow pool is 4.01.)

SUGGESTED READING

Gilbert Strang. 18.06SC Linear Algebra. Fall 2011. Massachusetts Institute of Technology: MIT OpenCourseWare, <https://ocw.mit.edu>. License: Creative Commons BY-NC-SA. <https://ocw.mit.edu/courses/mathematics/18-06sc-linear-algebra-fall-2011/index.htm>.

APPENDIX 2

Introduction to Programming in Python

Xin Huang

Northern Arizona University, Flagstaff, USA

This appendix is intended to equip readers with little or no programming experience with basic skills to write and read programs in the Python language. Python is a powerful programming language widely used in many applications. This appendix introduces basic programming knowledge in Python including variables, operators, function, class, module, etc. All code examples in this appendix come from Python files (`test_p2.py`, `GeneralModel.py` and `model.py`), which will be used in practice chapters 8, 12, and 16. Python 3.7 is preferred for the practice chapters of this book. Readers are not expected to be an expert in programming but to acquire the basic ability to read/write Python codes for the practice chapters.

WHAT IS PYTHON AND HOW DOES IT RUN?

Python is a programming language used for general-purpose software engineering. It is one of the most popular languages among scientists, engineers, and mathematicians for the following reasons. First, Python has simple syntax similar to the English language. It is easy to read, write, learn, and maintain Python code. Second, Python can work seamlessly on different platforms such as Windows, Linux and Mac. Third, Python is an interpreted language, which means that it executes the code line-by-line. If any error occurs, it will stop further execution until programmers are able to locate the errors and fix them. Fourth, Python comes with many great standard libraries to provide users with a vast choice of functions needed for their tasks.

Similar to English, Python is a language composed of vocabulary and syntax. The vocabulary in

English comprises different words. The vocabulary in Python comprises operators, variables/operands and keywords. As in common languages, the right syntax must be applied when linking different parts of the vocabulary into meaningful statements. In English, the sentence ‘learn python I’ is not syntactically valid. In Python, the expression “hello”+9 is not syntactically valid. Similar to learning English, readers will learn the most common vocabulary and syntax of Python from this appendix.

Before reading further, please refer to Appendix 3 to install Python on your computer.

There are two ways to run a Python program. One way is to use an interactive shell window. The shell will prompt `>>>` and wait for the user to type in a line of Python code. Given the code input, the shell will execute this code, display the results and wait for the next code input. For all but the simplest operations, we normally wish to gather consecutive lines of code in a Python source file (`xxx.py`). We may execute a Python source file with a command like `python xxx.py`. All code lines in this source file will be executed one after the other until the end of the program is reached. In the practice chapters, we will use this second way.

THE FIRST PYTHON PROGRAM

A Python program is a collection of code that manipulates data. Figure A2.1 is a code segment in the `test_p2.py` program. Each code line except for the annotation lines is called an expression. These expressions will be executed by the Python interpreter. There are two ways to denote an annotation. One is a single-line annotation starting with `#` as shown in Figure A2.1. The other is multiple-line annotations using `'` or `"` symbols. The expression

if `__name__ == '__main__':` in line 6 of Figure A2.1 indicates the start of a Python program. If one line contains multiple code expressions as line 12–13, it is recommended to use semicolon (;) to separate different expressions.

One of the most common expressions is to assign a value to a variable like lines 8–26 in the sample code of Figure A2.1. Taking line 12 as an example, the variable `f31` will store the value of 0.72. This value can be retrieved by this variable name after this code line. A variable must be assigned a value before any expressions that use it. Readers will learn more about variable types in the next section.

Another common expression is to print out on the screen. Figure A2.2 is to display the last elements in the `res` variable (a two-dimensional array) on the screen. Programmers often use the print expression to check values stored in variables, which is a useful approach for debugging the Python program.

Python uses colons to indicate an indented code block, which often appears in while-loop,

for-loop, if-else condition procedure and function definitions. Readers will learn more about these procedure definitions in the following section. The code groupings in these procedures are indicated by white space or indentation (Figure A2.3). The right level of indentation is important – too much or too little space will induce an error.

VARIABLES AND OPERATORS

As we learned in the previous section, variables are one of the primary vocabulary elements in the Python language. Similar to words, variables are containers to store information such as numbers and string values. The information type decides the variable type. Generally, the variable types in Python are integers (e.g., 1, 2, 3, ..), rational numbers (e.g., 3.1415926), complex numbers (e.g., 12 + 0.2i), strings (e.g., “helloworld”), Boolean values (i.e. TRUE or FALSE) and NaN. The final variable type is special, and it only has one value that is

```

6  if __name__ == '__main__':
7
8      output_folder = sys.argv[1]
9
10     B = np.array([0.45, 0.55, 0, 0, 0, 0, 0]).reshape([7,1]) # allocation
11
12     f31 = 0.72; f41 = 0.28; f42 = 1; f53 = 0.45; f54 = 0.275; f64 = 0.275;
13     f65 = 0.296; f75 = 0.004; f56 = 0.42; f76 = 0.03; f57 = 0.45;
14
15     A = np.array([-1, 0, 0, 0, 0, 0, 0,
16                  0, -1, 0, 0, 0, 0, 0,
17                  f31, 0, -1, 0, 0, 0, 0,
18                  f41, f42, 0, -1, 0, 0, 0,
19                  0, 0, f53, f54, -1, f56, f57,
20                  0, 0, 0, f64, f65, -1, 0,
21                  0, 0, 0, 0, f75, f76, -1]).reshape([7,7]) # tranfer
22
23
24     #turnover rate per day of pools: foliage, wood, metabolic litter, structural
25     #litter, soil microbial,slow soil, passive soil
26     temp = [0.00176, 0.000100104, 0.021468, 0.000845, 0.008534, 8.976e-005, 0.0000154782]

```

Figure A2.1. Code example of defining an annotation with #. This code segment is from the `test_p2.py` program.

```

67     print(res[:,year-1]) # print result of the last year

```

Figure A2.2. Code example of print expression.

```

56     for i in range(1, 4):
57         for j in range(1, 4):
58             if (i-1)*3 + j > 7:
59                 break
60             ax = plt.subplot(3, 3, (i-1) * 3 + j)
61             ax.plot(x, res[(i-1) * 3 + j - 1,:])
62             plt.xlabel("year", fontsize = 12)
63             plt.ylabel(pool_names[(i-1) * 3 + (j-1)] + " pool (%g C m-2%)", fontsize = 12)

```

Figure A2.3. A Python program uses organized spaces to indicate code grouping. This code example includes two for-loop code blocks (red and green arrows and texts) and an if-else code block (blue arrow and text).

none. Other programming languages such as Java require you to declare a variable type before using it. Python, however, does not require variable declarations. Variables with the same name can be used in different places in a Python program to store different types of values.

Operators are another primary vocabulary element in Python. Table A2.1 shows a collection of binary operators performing calculations on two variables. They may work on numbers or Boolean values. The final two operators are special. The numerical calculation and assignment are performed at the same time. The expression `a/=b`, for example, first calculates the value of `a` divided by `b` and then assign this as a new value for the variable `a`.

Another operator collection is about control flow. Control flow is to decide which and how expressions are to be executed. We will use examples in English to introduce two common control flows in Python. The first one is conditional control flow. The example in English is: ‘If tomorrow is sunny, I will go hiking, otherwise I will stay at home’. In the conditional control flow, only one set of statements will be executed, either ‘go hiking’ or ‘stay at home’. The syntax of conditionals in Python is illustrated in Figure A2.4a. It starts with a `<BooleanExpr>` expression whose value is either `TRUE` or `FALSE`. If the value of this expression is `TRUE`, then `<ExpressionT1>`, ..., `<ExpressionTk>` will be executed; If it is `FALSE`, then another set of expressions `<ExpressionF1>`, ..., `<ExpressionFk>` will be executed. A code example of if-else control flow is shown in Figure A2.4b. If the variable type of `self.input_fluxes` is an array, the variable `self.tmp_input_fluxes` will be assigned a specific element in this array variable. Otherwise, `self.tmp_input_fluxes` saves the whole values in `self.input_fluxes`.

The second type of control flow is a loop, including a for-loop (Figure A2.5) or a while-loop (Figure A2.6). A loop repeats a set of statements over and over until a termination status is reached. An example in English is ‘repeat taking medicine until you feel better’. The syntax of for-loop control flow is shown

TABLE A2.1
A collection of binary operators in python

Operator	Description
<code>a+b</code>	sum
<code>a-b</code>	difference
<code>a*b</code>	product
<code>a/b</code>	division
<code>a//b</code>	The integral part of the quotient when <code>a</code> is divided by <code>b</code>
<code>a%b</code>	The remainder when <code>a</code> is divided by <code>b</code>
<code>a**b</code>	<code>a</code> to the power of <code>b</code>
<code>a b</code>	True when either <code>a</code> or <code>b</code> is True
<code>a&b</code>	True when both <code>a</code> and <code>b</code> is True
<code>not a</code>	True if <code>a</code> is False
<code>a==b</code>	True when <code>a</code> equals to <code>b</code>
<code>a!=b</code>	True when <code>a</code> doesn't equal to <code>b</code>
<code>a+=b</code>	<code>a=a+b</code>
<code>a/=b</code>	<code>a=a/b</code>

in Figure A2.5a. It starts from retrieving the first value from a `<listExpr>` to a variable `<var>`, then executing `<Expression1>`, ..., `<ExpressionN>`. This set of statements will be repeatedly executed when every value in `<listExpr>` is retrieved. At the end, the `<var>` will store the last element in `<listExpr>`. The code example of a for-loop (Figure A2.5b) uses a useful procedure, `range(10)`, as the `<listExpr>`. Generally, `range(n)` returns integers from 0 up to `n-1`. The for-loop iterates this numerical list from 0 to 9, assigns each element in the list to a variable `t` and prints the value of `t` to the screen.

Generally, any for-loop control flow can be rewritten into a while-loop flow. The code example in Figure A2.6b generates the same results as that in Figure A2.5b. The special characteristic of

```

(a)
1  if <BooleanExpr>:
2      <ExpressionT1>
3      ...
4      <ExpressionTk>
5  else:
6      <ExpressionF1>
7      ...
8      <ExpressionFk>

(b)
37 if type(self.input_fluxes) == np.ndarray:
38     self.tmp_input_fluxes = self.input_fluxes[idx]
39 else:
40     self.tmp_input_fluxes = self.input_fluxes

```

Figure A2.4. The statement form of conditional control flow (a), and its code example in `GeneralModel.py` (b).

<p>(a)</p> <pre> 1 for <var> in <listExpr>: 2 <Expression1> 3 ... 4 <ExpressionN> </pre>	<p>(b)</p> <pre> 1 for t in range(10): 2 print(t) </pre>
--	--

Figure A2.5. The statement form of for-loop control flow (a), and its code example (b).

<p>(a)</p> <pre> 1 while <BooleanExpr>: 2 <Expression1> 3 ... 4 <ExpressionN> </pre>	<p>(b)</p> <pre> 1 t=0 2 while t<10: 3 print(t) 4 t=t+1 </pre>
--	---

Figure A2.6. The statement form of while-loop control flow (a), and its code example (b).

a while-loop is that programmers do not need to know in advance how many times the set of expressions needs to be repeated. The syntax of a while-loop control flow is shown in Figure A2.6a. It starts by evaluating `<BooleanExpr>` whose value is either `TRUE` or `FALSE`. If the value of this expression is `TRUE`, the set of expressions will be executed and the `<BooleanExpr>` will be evaluated again. If the expression of `<BooleanExpr>` gets `FALSE`, the execution of this loop will be terminated. Generally, programmers initialize a variable before the while loop, such as variable `t` in the code example (Figure A2.6b). The value of this variable is changed each time by the set of expressions in the loop block to be repeated and the new value is tested in the `<BooleanExpr>` to decide whether the loop is to be terminated or not.

The `break` keyword offers another mechanism to exit a loop which is currently being executed. It works both with a for-loop or a while-loop. To make the program more readable, it is suggested to avoid the `break` keyword. As mentioned above, colons are used to indicate an indented code block in the control flow. Expressions after the colon have to be indented to one level (Figure A2.3). For example, in an if-else control flow, expressions after `if` and `else` keywords are at the same level of indentation and this indentation level is below the `if` and `else`.

ADVANCED VARIABLES AND OPERATORS

The variables and operators we have seen so far are known as primitive variables and operators. They are capable of coping with simple programming tasks. However, if we only use the primitive variables and operators, programs can quickly become long and messy. For longer programs, a modular design

is preferred to make code clear and readable, and easier to modify or debug. Modularity entails aggregating primitive variables and operators into more advanced variables or operators. Advanced variables are centered around the organization of data. These advanced variables include `list`, `array`, `dictionary`, and `set`. Because the practice chapters of this book mainly use the `list` type, this appendix will only introduce `list`. For other advanced variables, please refer to suggested reading materials at the end of this appendix. Advanced operators are collections of primitive operations such as basic arithmetic, conditional evaluation and recursion. This appendix will introduce three advanced operators: `function`, `class`, and `module`.

The List Variable

A list is a collection of ordered and mutable variables such as numbers, strings, or boolean values. A list is written using square brackets `[]`, with elements separated by commas. Figure A2.7 shows an example of defining a list and some common operations on the list. Each element in the list is numbered starting from 0. Therefore, the first element is at index 0, the second element is at index 1, and the final index is one less than the size (number of elements) of the list. We can retrieve the final element of a list called `varList` using the expression `varList[len(varList)-1]`. We can also select a continuous part of the list through slicing. For example, `varList[0:3]` returns the first three elements (i.e., `varList[0]`, `varList[1]`, `varList[2]`) of `varList`. Remember that the index after the colon in the slicing operator (i.e., 3) will not be included in the result. The slicing operator always returns a new list. Therefore, `varList[0]` returns a variable value while `varList[0:1]` returns a list which only includes one variable

```

1 L=[2020,"training",[3,4]] #define a list with 3 elements. The third element is another list
2 L[0]=2021 # modify the first element to 2021
3 L[0:2] # return the first and second elements, [2020,"training"]
4 L[0:1] # return a new list, [2020]
5 L[0] # return a number, 2020
6 len(L) # return the length of the list, 3
7 L.extend([1,2]) # add more elements to L, now L is [2020, "training",[3,4],1,2]
8 L.remove(2020) # remove the element 2020, now L is ["training",[3,4]]

```

Figure A2.7. Common operators of list variable type.

`varList[0]`. Some other common operations are illustrated in Figure A2.7, such as getting the length of a list, inserting and removing elements.

The Function Operator

As an advanced operator, function can be viewed as a collection of primitive variables and operators. Specifically, one function `f1` can call another function `f2`. As a result, all basic operators in `f2` are included in `f1`. One powerful characteristic of a function is that it can be used as a black box and we only care about the argument inputs and outputs returned. Another feature of function is reusability of code. For example, if we want to decide whether a numerical value is even or odd, we need to write four lines of code for each value. The number of code lines is four times the number of values (Figure A2.8a). Gathering these four lines of code in a function helps keep the code clean and efficient. In Figure A2.8b, we put all expressions used in determining if a number is even or odd into a function called `Evenodd`. We can invoke

the function by typing its name, followed by an argument in parentheses. Each time the function is called, it determines whether the argument is even or odd and prints out the result on the screen.

Figure A2.9 shows the syntax of a function definition. We write a sequence of expressions inside the function and give that function a name. The sequence of expressions can be executed at any point in the Python program by calling the function name. In the example of Figure A2.9, the function name is `get_df`. In function definition, variables in parentheses `()` are the input and become available for use by expressions inside the function body. These input variables are also called parameters or arguments. When calling a function, we need to pass a variable storing specific values to the function. In the example of Figure A2.8, the `Evenodd` function requires a numerical value. If we want a value to be returned after a function is called, we need to add a `return` expression at the end of the function definition like line 81 in Figure A2.9. Remember, the indentation in a function definition is important. In the function,

<p>(a)</p> <pre> 1 if 103%2==0: 2 print('even') 3 else: 4 print('odd') 5 if 1100%2==0: 6 print('even') 7 else: 8 print('odd') 9 if 530%2==0: 10 print('even') 11 else: 12 print('odd') 13 if 79%2==0: 14 print('even') 15 else: 16 print('odd') </pre>	<p>(b)</p> <pre> 1 def Evenodd(x): 2 if x%2==0: 3 print('even') 4 else: 5 print('odd') 6 7 Evenodd(103) 8 Evenodd(1100) 9 Evenodd(530) 10 Evenodd(79) </pre>
--	--

Figure A2.8 Code example to decide whether each of the numbers 103, 1100, 530, and 79, is even or odd, and display the answer on the screen. (a) Version using simple code operators; (b) Version using a function to keep the code clean and efficient.

```

77 def get_df(self):
78     res = self.Y.y.T
79     df = pd.DataFrame(res)
80     df.columns = self.create_headers()
81     return df

```

Annotations for Figure A2.9:

- start with `def` keyword (points to line 77)
- name of the function (points to `get_df`)
- inputs (points to `self`)
- return outputs (points to `return df`)

Figure A2.9. The syntax of function definition from `GeneralModel.py`.

```

1 def func1(a,b,c):
2     ...
3
4 def func2(a,b,c):
5     func3()
6     ...
7
8 def func3():
9     ...

```

```

1 if __name__ == '__main__':
2     a=1
3     b=1
4     c=3
5     func1(a,b,c)
6     func2(a,b,c)

```

Annotations for Figure A2.10:

- Indicate the start of the program (points to line 1 of the second code block)
- Call each function (points to lines 5, 6, and 8 of the first code block)

Figure A2.10. A workflow of function calls.

all expressions after the `def` keyword and comma have to be indented one level below.

A function can call another function. With more and more functions defined, which one is the highest-level function to call others? As we learned in the first section, a code line with the expression `if __name__ == '__main__'` indicates the start of a Python program no matter where it occurs in a source code file. The code block with the start of program execution is called the main program. In Figure A2.10, three functions (`func1`, `func2`, `func3`) are defined. The main program calls `func1` and `func2` and `func2` further calls `func3`. The variables in `func1`, `func2`, and the main program have the same names: `a`, `b`, `c`. How to distinguish them? Python uses namespace to do so. Namespace is an isolated scope where variables are valid. So `func1` and `func2` have local namespaces and the main program has a global namespace as well. The `a`, `b`, and `c` variables can have different values in different namespaces. If we want to change a variable created in the global namespace, it is required to clarify the global property of the variable by using the `global` keyword before the name of the variable.

The Class Operator

A class is a mixture of variables and functions, which are called attributes. Just like a function, once defined, can be used many times, a class may

be defined, and then multiple instances can be made. Each class instance is called a class object and maintains its own attributes, which provides a way to reuse code to keep the program clean and efficient. This programming style is called object-oriented programming.

Figure A2.11a shows the syntax of a class definition. The `BaseClassName` is an abstract class to support inheritance. A more detailed introduction to inheritance is available in the recommended reading materials. In practice, most expressions inside a class definition are function definitions. Figure A2.11b lists all functions defined in the `GeneralModel` class from `GeneralModel.py`. A special function is for instantiation and its name is `__init__` (spelt with double underscores on both sides of `init`). The instantiation function is first executed whenever a class object is created and used for the first time within a Python program. The instantiation function in the `GeneralModel` class (Figure A2.11b) requires six parameters and `self` keyword represents the object itself. Figure A2.11c shows an example to initialize a `GeneralModel` object (`mod`). Typically, the expression is the class name followed by a list in parentheses of the parameters required by the instantiation function. The syntax to access attributes of the class object (i.e., variables and functions) is `obj.name` where `obj` is the class object and `name` is the variable name or function name. Figure A2.11c shows some examples of calling functions defined in `GeneralModel`.

```

(a)
1 class ClassName(BaseClassName):
2     <Expression1>
3     ...
4     <ExpressionN>
(b)
1 class GeneralModel(Model):
2     def __init__(self, times, B, A, K, iv_list, input_fluxes, xi=1):
3         ...
4     def ode_solver(self):
5         ...
6     def get_input(self, t, y):
7         ...
8     def right_hand_equation(self, t, y):
9         ...
10    def get_x(self):
11        ...
12    def get_df(self):
13        ...
14    def write_output(self, filename):
15        ...
16    def get_diagnostic_variables(self):
17        ...
18    def sas_u_spinup(self):
19        ...
20    def create_headers(self):
21        ...
22    def get_x_df(self):
23        ...
24    def get_pool_n(self):
25        ...
26    def get_c_input(self):
27        ...
(c)
46 mod = GeneralModel(times, B, A, K, iv_list, input_fluxes)
48 res = mod.get_x()

```

Figure A2.11. (a) Syntax of class definition; (b) functions defined in `GeneralModel` class; (c) code expressions to initialize a `GeneralModel` object and to call a function attribute of the object.

```

1 from GeneralModel import GeneralModel
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import sys

```

↖ The third way to import a module
 } The second way to import a module
 ← The first way to import a module

Figure A2.12. Three ways to import a module. Codes are from `test_p2.py`.

The Module Operator

A module is a file containing reusable variables and functions. Unlike a class, which enables the instantiation of multiple objects and modification of attributes after creation, a module is a static storage of the reusable attributes. Similar to namespaces, attributes in one module file are not visible to other files. To make them visible, we need to load the

module file first before using the variables or functions in the file. Figure A2.12 shows three ways to load a module file using the `import` keyword. One way is to load it directly without assigning a simple name to this module. We can retrieve the variables or functions in the module via its default name. The other way is to load the module and assign a new name to it. In the example of Figure


```

(a)
1 import numpy as np
2 K=np.zeros(49).reshape([7, 7]) # create an array with 49 zeros and convert it to 7*7 matrix
3 np.multiply(K,1/864) # multiple each element in K and 1/864
4 np.matmul(A,B) # return the multiplied matrix product of two arrays A and B
5 np.linspace(0,10*365,num = 10) # return 10 evenly spaced numbers over the interval [0,10*365]
6 np.abs(-18) # return the absolute value, 18
7 np.ndarray # an array object
8 np.linalg.inv(matrix_AK) # return the inverse of matrix_AK
9 np.sum(A) # return the sum of elements in matrix A

(b)
1 import pandas as pd
2 df=pd.DataFrame(Xc) # return a tabular data (df) from Xc
3 df.Columns=['X1','X2','X3'] # assign labeled column axis
4 df['X1'] = carbon_storage # arithmetic operations align on both row and column
   labels.
5 df.to_csv(filename, index=False) # write to a csv file without row names

(c)
1 from scipy.integrate import solve_ivp
2 solve_ivp(func,t,y0,t_eval=times,vectorized = True)
3 # Solve a system of ordinary differential equations given an initial value, e.g.
4 # dy/dt=func(t,y) and y(t0)=y0
5 # t is the interval (t0,tf) where the solver starts with t=t0 and ends t=tf
6 # t_eval is times at which to store the computed solution according to t
7 # vectorized indicates whether func is implemented in a vectorized fashion or not

(d)
1 import matplotlib.pyplot as plt
2 fig = plt.figure(12, figsize=(14, 7))
3 # create a figure.12 is its identifier.figsize sets its width and height in inches
4 plt.subplots_adjust() # specify the subplot layout
5 ax = plt.subplot(nrow, ncols, index)
6 # add subplot to a current figure at the specified grid
7 ax.plot(x,y) # plot y versus x as lines and/or markers
8 plt.xlabel('year') # add x labels
9 plt.ylabel('pool') # add y labels
10 plt.savefig(fileName) # save the figure to a file

```

Figure A2.13. Code examples in `GeneralModel.py` and `test_p2.py` using (a) Numpy module, (b) Pandas module, (c) Scipy module, and (d) Matplotlib module.

A2.12, we import the Numpy module and rename it as `np`. Then we can call functions via `np.funcName` such as `np.zeros(49)`. The third way is to load specific variables or functions from a module file, which saves memory in runtime. Unlike the first two ways, these variables or functions loaded can be used without the module name.

In the practice chapters of this book, we will use Numpy, Pandas, Scipy, and Matplotlib modules. Numpy offers comprehensive mathematical functions working on arrays. Pandas provides functionality to manipulate tabular data as a two-dimensional data structure. Scipy is based on Numpy and solves scientific and mathematical problems. Matplotlib contains functionality for plotting data. Figure A2.13 shows some common functions from these modules that are useful for the practice chapters of this book.

SUMMARY

If you have read this appendix carefully, you should now have a basic knowledge of Python programming including variables, if-else conditional control, for-loop, while-loop, list, function definition, class definition, and loading module. This knowledge is sufficient to perform the programming tasks of the practice chapters of this book. If you wish to learn more about Python programming, you can refer to the learning resources referenced below.

SUGGESTED READING

- <https://wiki.python.org/moin/BeginnersGuide>
- <http://pythontutor.com>

- <https://thepythonguru.com>
- <https://pymbook.readthedocs.io/en/latest/>
- <https://docs.python-guide.org>
- <https://www.w3schools.com/python/>

QUIZZES

1. What is an annotation?
2. What is an operator?
3. Can one function call another function?
4. How would you express an if-then sentence in Python?



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

APPENDIX 3

CarboTrain User Guide

Yuan Gao

Northern Arizona University, Flagstaff, USA

This appendix chapter provides a guide for the software CarboTrain, which is short for Carbon cycle modeling Training course, and is tailored for use in the training course *New Advances in Land Carbon Cycle Modeling*. Modeling studies usually require extensive techniques in programming. However, the main goal of the training course is to acquaint the trainees with advances in modeling land carbon dynamics. CarboTrain is designed to help trainees to reach their learning goals without getting bogged down in programming, in which they may have very different levels of skill. The software implements all the exercises in Units 2–10 of the training course, and of this book.

INTRODUCTION

CarboTrain has a user interface as shown in Figure A3.1, and it can run on computers running the Windows or macOS operating system.

This software was developed with Python 3.7.9 and PyQt5. In order to run the software properly, some other software systems have to be pre-installed. The following section provides a step-by-step guide on how to install and use CarboTrain. Please be aware that CarboTrain is not compatible with the CPU from Apple Silicon.

DOWNLOAD CARBOTRAIN

We have a docker image available on docker hub <https://hub.docker.com/r/gaoyuan199325/carbotrain>. If you have docker installed you are ready to go. We recommend using the docker image because we have all the essential software installed, so you don't need to create any virtual environments or struggle installing the software

on your computer. If you choose to use the docker image, you may skip this part and jump to “Uses of CarboTrain” further down in this appendix.

If you want to run CarboTrain without docker, you may download and install it with essential software on your computer. First, please download the software from: <https://www2.nau.edu/luo-lab/download/CarboTrain.zip>.

PREREQUISITE SOFTWARE

Three pre-installed software packages are required to run CarboTrain:

1. Python 3.7.9 and relevant packages
2. Fortran compiler
3. R

Since different operating systems may have different ways to install all the software needed, we will show how to install these software systems with different operating systems.

INSTALLATION ON WINDOWS

Install Python 3.7.9

Download Python 3.7.9 from <https://www.python.org/downloads/release/python-379/> (“**Windows x86-64 executable installer**”) and install it on your computer. When installing Python, check “**Add Python 3.7 to PATH**” as shown in Figure A3.2. After installing Python, open the CMD window to see whether it is installed correctly. To open the CMD window, type in “CMD” in the search bar of your computer, as shown in Figure A3.3. Follow the steps in Figure A3.4 to check whether you have installed Python successfully.

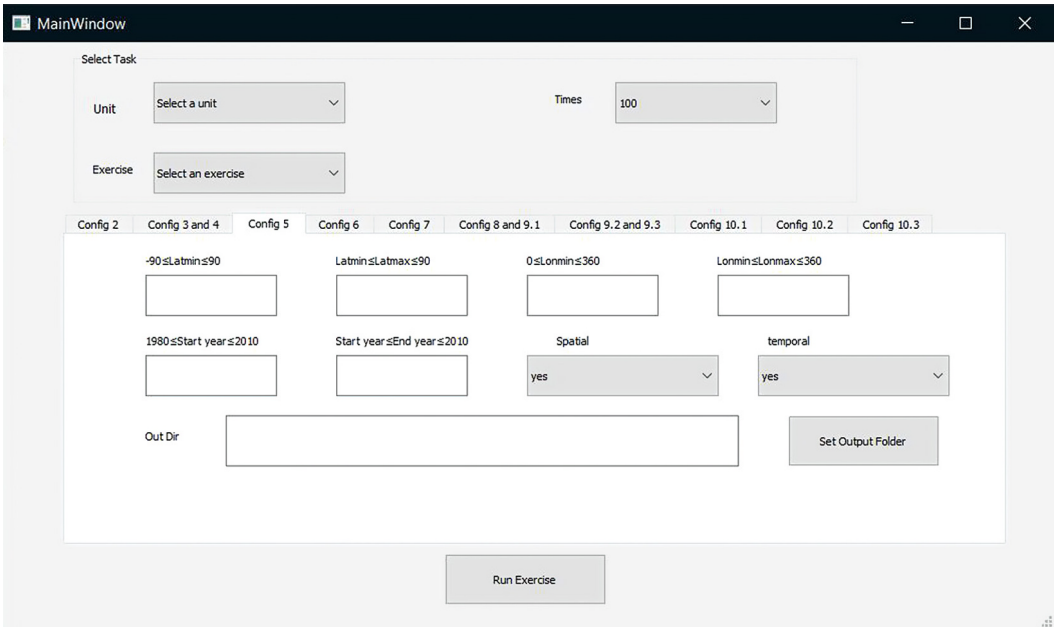


Figure A3.1. The Graphical User Interface (GUI) of CarboTrain.



Figure A3.2. Steps to install Python on your computer.

Fortran Compiler

Go to the website <https://sourceforge.net/projects/mingw/> and download the software as shown in step 1 in Figure A3.5. Then, step 2 is to run the downloaded file by double-clicking it. Next, you need to click the “Install” button to start the installation. Finally, you need click “Continue” several times to finish the installation.

Once the installation is finished, another window will be popped out automatically as shown in Figure A3.6.

Check each box in Figure A3.6 and then select “Mark for Installation”. Then go to “Installation and click “Update Catalogue” as shown in Figure A3.7 to confirm the changes. You then need to click “Review Changes” in the pop-up window and then click “Apply” in the following pop-up window to install all the packages needed.

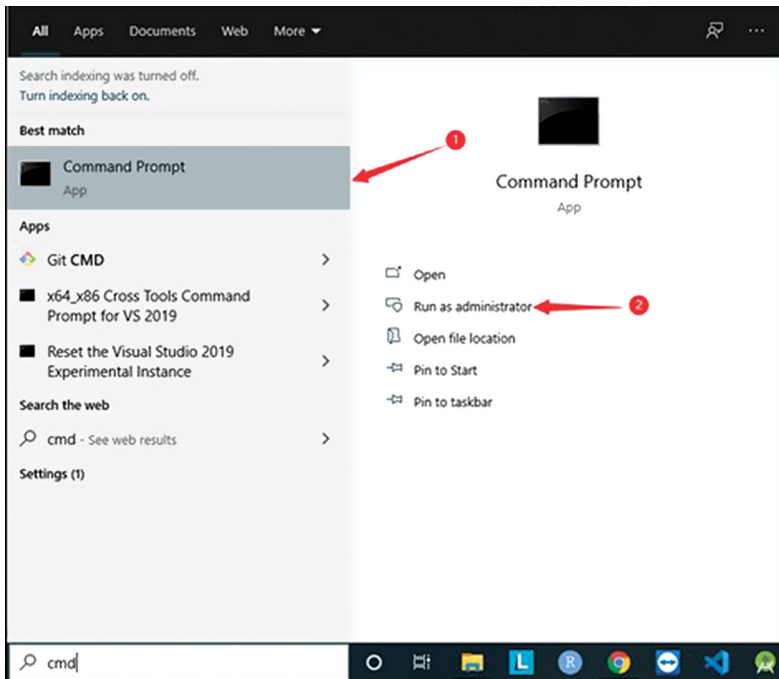


Figure A3.3. Steps to open a CMD window.



Figure A3.4. Steps to check whether Python 3.7.9 is installed.

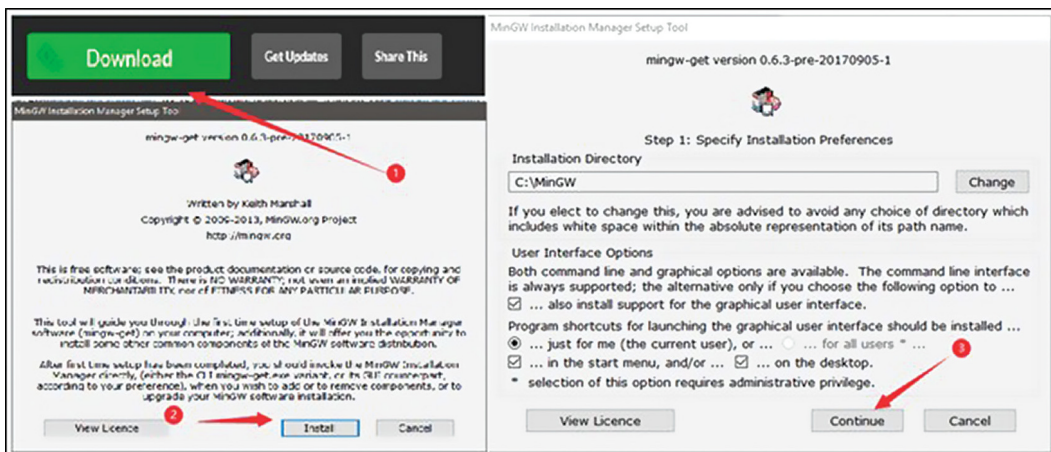


Figure A3.5. Steps to install MinGW.

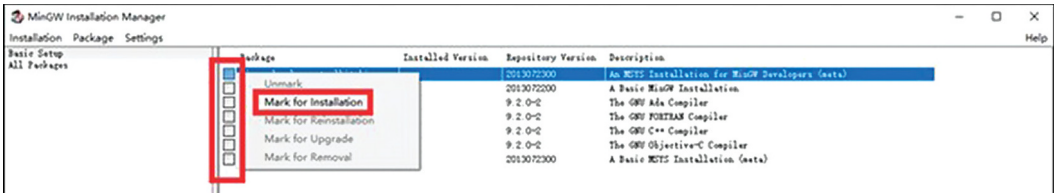


Figure A3.6. MinGW.

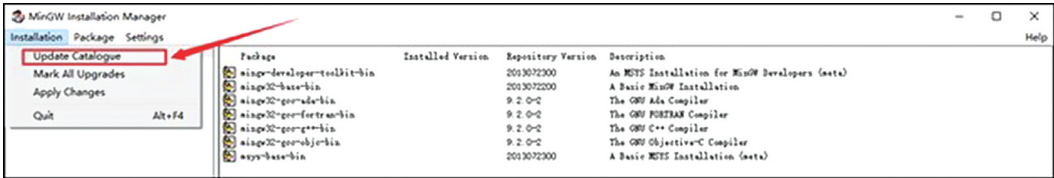


Figure A3.7. Update catalogue.

When you have finished the installation of all the packages, you need to set the environment variables in your computer. To do this, search for “edit the system environment variables” in the search bar of your computer, and then click “Edit the system environment variables” and the “System Properties” window will show up. Go to the “Advanced” menu and then click “Environment Variables...” as shown in Figure A3.8.

Follow the steps in Figure A3.8 to open the environment variable setting window as in Figure A3.9. Follow the steps in the figure to add a new path. First select “path” and then click “Edit...”, step 1 and 2, respectively in the left panel of Figure A3.9. In the right panel of the dialog shown in Figure A3.9, click “New” and then add “C:\MinGW\bin” to finish the setting of the path.

Finally, open a CMD window by following the steps shown in Figure A3.3 to check whether the Fortran compiler has been installed correctly by following the steps in Figure A3.10. When you type “gfortran”, it will show an error warning. This is because no input files were specified as an argument to the gfortran command. The Fortran compiler has been successfully installed.

Install R 3.6.3

Download R 3.6.3 from <https://cran.r-project.org/bin/windows/base/old/3.6.3/>, and install it in your computer. Once finished, set the environment variables for R following the two steps in Figure A3.11.

After setting the environment variables, you can now open a CMD window by following the

steps shown in Figure A3.3 and follow the steps in Figure A3.12 to see whether you have installed R 3.6.3 correctly.

After the installation, you need to install the Python packages and R packages needed in the training course and then compile the Fortran code. Locate the folder of CarboTrain, and copy the path of that folder as shown in Figure A3.13.

Once you have copied the path, open a CMD window by following the steps shown in Figure A3.3 and type in the following commands to install the Python packages and R packages:

1. rmdir /s/q C:\Users\YOUR_USER_NAME\Documents\R\win-library\3.6\
2. cd path_you_copied #Press Enter to confirm
3. pip3 install pyproj-2.6.1.post1-cp37-cp37m-win_amd64.whl #Press Enter to confirm
4. pip3 install basemap-1.2.2-cp37-cp37m-win_amd64.whl #Press Enter to confirm
5. pip3 install -r requirements.txt #Press Enter to confirm
6. Rscript Rinstall_packages_win.R #Press Enter to confirm

In order to run the TECO model properly, you also need to compile the source code. Go to the TECO source code folder under CarboTrain→Source_code→TECO_2.3 and copy the full path as we just did in the last step. Once you have copied the path, open a CMD window by following the steps shown in Figure A3.3 and type in the following command to compile the source code. You may ignore any warnings that come up.

```
1. gfortran -oTECO_2.3.exe    #Press Enter to confirm
TECO_2.3.f90
```

Windows users are now ready to run the practice sessions for each unit, and may skip to the section Uses of CarboTrain, below. We will now provide instructions on how to install CarboTrain on a macOS computer.

INSTALLATION ON MACOS

Install Python 3.7.9

Download Python 3.7.9 from <https://www.python.org/downloads/release/python-379/> (select “macOS 64-bit installer”) and install it in your MacBook. Once you have finished, open a terminal to test the installation following the two steps in Figure A3.14.

Once you have opened a Terminal, follow the steps in Figure A3.15 to check whether you have installed Python correctly.

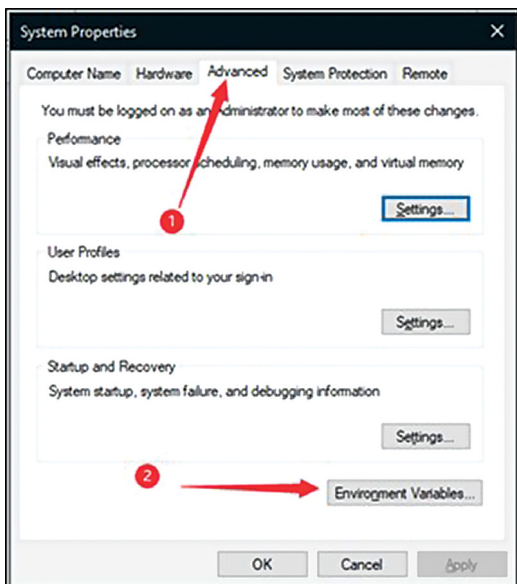


Figure A3.8. Steps to set the environment variables under “System Properties”.

Fortran Compiler

Download “gcc-x.x-bin.tar.gz” and “gfortran-x.x-bin.tar.gz” from <http://hpc.sourceforge.net>, where x.x indicates the different versions. Then type in the following commands:

1. `cd & cd Downloads` #Press Enter to confirm
2. `gunzip gcc-x.x-bin.tar.gz` #Press Enter to confirm
3. `gunzip gfortran-x.x-bin.tar.gz` #Press Enter to confirm
4. `sudo tar -xvf gcc-x.x-bin.tar -C /` #Press Enter to confirm
5. `sudo tar -xvf gfortran-x.x-bin.tar -C /` #Press Enter to confirm

You may be asked to type in your password to continue. Once finished, you may open a Terminal to check Figure A3.16.

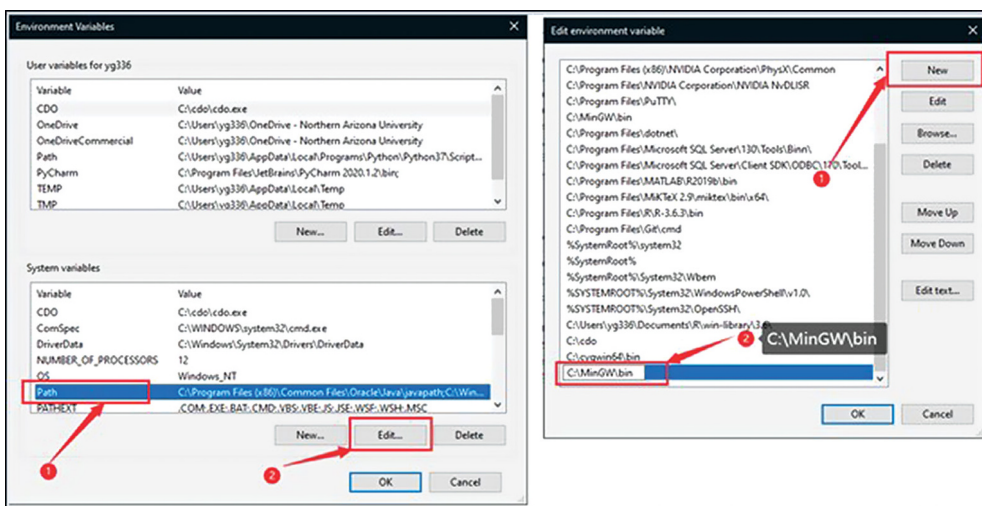


Figure A3.9. Steps to add a new path for MinGW.

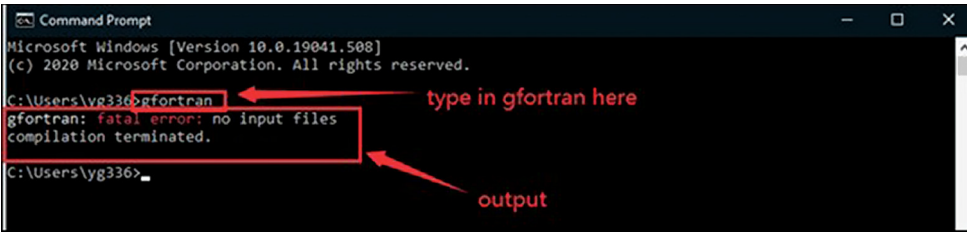


Figure A3.10. Test Fortran compiler installation.

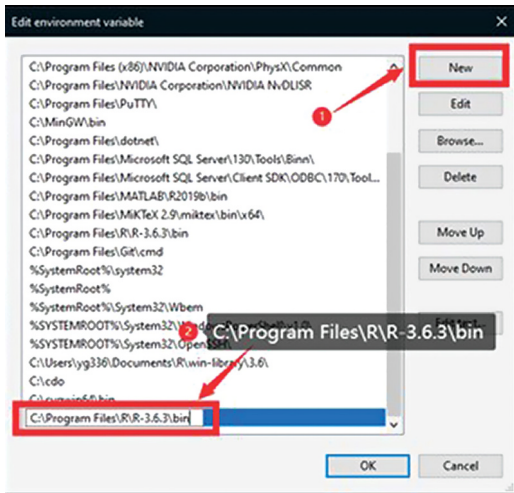


Figure A3.11. Set environment variables for R.

Install R 4.0.5

Download R 4.0.5 from <https://cran.r-project.org/bin/macosx/> (file name is R-4.0.5.pkg) and install it in your computer. After installation of R, run the following commands to create the soft links:

-
1. `ln -s /Library/Frameworks/R.framework/Resources/R /usr/bin/local/R` #Press Enter to confirm
 2. `ln -s /Library/Frameworks/R.framework/Resources/Rscript /usr/bin/local/Rscript` #Press Enter to confirm
-

You also need to open a Terminal to check whether the installation is successful (Figure A3.17).

Once you have installed the three software systems required, you need to install some Python

and R packages and compile the source code of the TECO model. To do this, copy the CarboTrain folder to your Desktop, open a Terminal and type in the following commands to install the packages:

-
1. `cd && cd Desktop/CarboTrain/` #Press Enter to confirm
 2. `/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install.sh)"` #Press Enter to confirm and then Enter the password and press Enter to confirm the installation
 3. `brew install geos` #Press Enter to confirm
 4. `brew install proj` #Press Enter to confirm
 5. `pip3.7 install basemap-1.2.2rel.tar` #Press Enter to confirm
 6. `pip3.7 install -r requirements.txt` #Press Enter to confirm
 7. `Rscript Rinstall_packages.R` #Press Enter to confirm
-

Next, you can compile the source code of TECO as follows:

-
1. `cd && cd Desktop/CarboTrain/Source_code/TECO_2.3` #Press Enter to confirm
 2. `gfortran -oTECO_2.3.exe TECO_2.3.f90` #Press Enter to confirm
-

When you see the information shown in Figure A3.18, click “Install” to install the tool. After installation, run the commands:

-
1. `cd && cd Desktop/CarboTrain/Source_code/TECO_2.3` #Press Enter to confirm
-

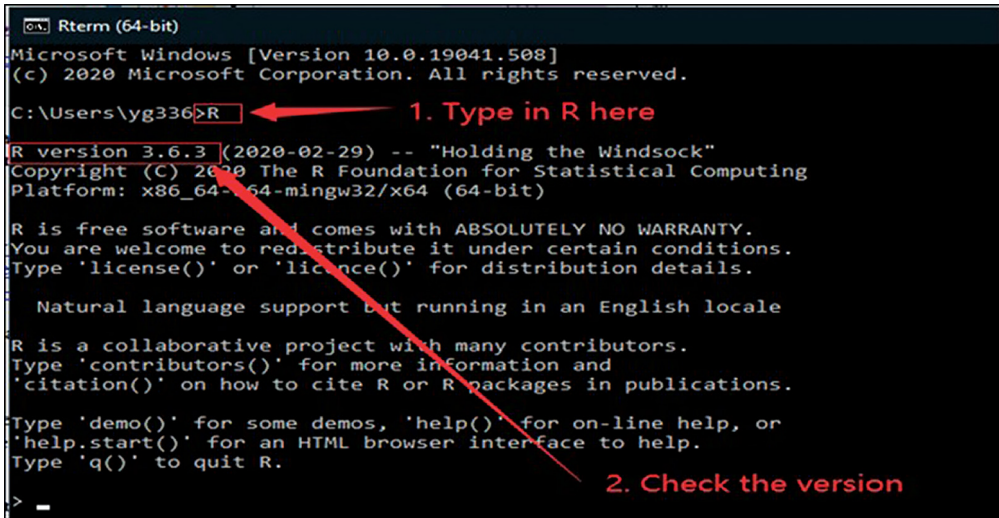


Figure A3.12. Steps to check whether R 3.6.3 is installed.

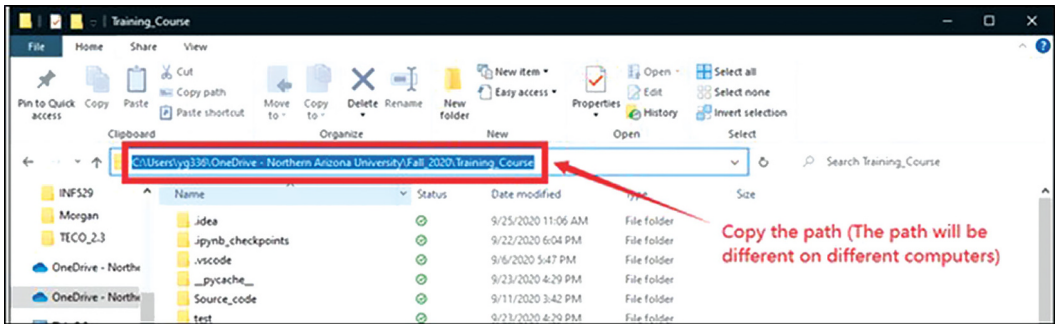


Figure A3.13. Locating the path of CarboTrain.

```

2. gfortran -o TECO_2.3.exe #Press Enter to confirm
   TECO_2.3.f90
  
```

You now have all the required software installed and are ready for the practices in the training course and in the chapters of this book.

USES OF CARBOTRAIN

CarboTrain will be used for the practices in Units 2 to 10. Units 2 to 4 make use of the matrix version of the TECO ecosystem model and Unit 5 is about traceability. In Unit 6, you will practice data assimilation (DA) with a simple version of TECO. In Units 7 to 9, you will use an intermediately complex version of the TECO model to perform data assimilation. Unit 10 is the practice for deep

learning with a process model, which is called PROcess-based deep learning and DATA-driven modeling (PRODA). The instructions for each practice are described in detail in the corresponding unit. Here we describe the general steps of how to use the software.

To use CarboTrain, you first need to launch it. In Windows, copy the path of CarboTrain, locate to the path you copied, and run the software in a CMD window as below:

```

1. cd path_you_copied #Press Enter to confirm
2. python main.py #Press Enter to confirm
  
```

In macOS, use the following commands to run the software:



Figure A3.14. The location of Terminal in macOS.

```

gaoyuan — Python — 80x24
Last login: Fri Oct 2 21:46:53 on console
gaoyuans-Mac:~ gaoyuan$ python3
Python 3.7.9 (v3.7.9:13c94747c7, Aug 15 2020, 01:48:08)
[Clang 6.0 (clang-600.0.57)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>>
  
```

Annotations: A red arrow labeled '1' points to the `python3` command. A red arrow labeled '2' points to the prompt `>>>`. A callout box says 'Type in python3' and another says 'Check the version'.

Figure A3.15. Steps to check Python version.

```

gaoyuan -- -bash -- 80x24
Last login: Fri Oct 2 22:21:08 on ttys000
gaoyuans-Mac:~ gaoyuan$ gfortran
gfortran: fatal error: no input files
compilation terminated.
gaoyuans-Mac:~ gaoyuan$
  
```

A red arrow points to the error message: `gfortran: fatal error: no input files compilation terminated.`

Figure A3.16. Check Fortran compiler installation.

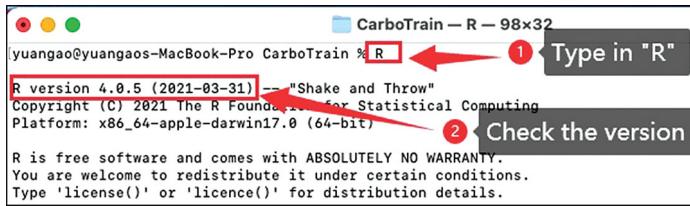


Figure A3.17 Steps to check R version.

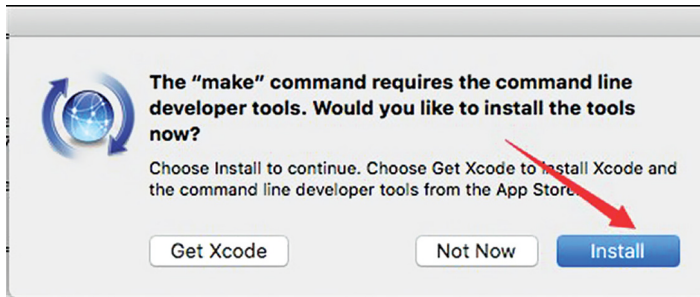


Figure A3.18. Installation of the command line developer tools.

-
1. `cd && cd Desktop/ CarboTrain/` #Press Enter to confirm
 2. `export QT_MAC_WANTS_LAYER=1` #Press Enter to confirm
 3. `Python3.7 main.py` #Press Enter to confirm
-

Once you have launched the software, you will see a GUI similar to Figure A3.1. The GUI of the software consists of two parts: exercise selection, and exercise configuration, as shown in Figure A3.19.

Each tab shown in the "Exercise configuration" part contains one or a set of exercises, as shown in Table A3.1.

The general steps for each practice are: (1) select a unit, (2) select an exercise, (3) configure the exercise you just selected, and (4) run the exercise. For example, in order to run Exercise 1 in Unit 10, you need to go through the steps shown in Figure A3.20. The first two steps are for selecting an exercise and step 3 is for configuring the exercise. The configuration required may be different depending on the details of the exercise. For this example, configuration involves selecting the output folder. In some exercises, you need to modify the source code, change the settings or customize the exercise according to your own questions. Once you have finished exercise selection and

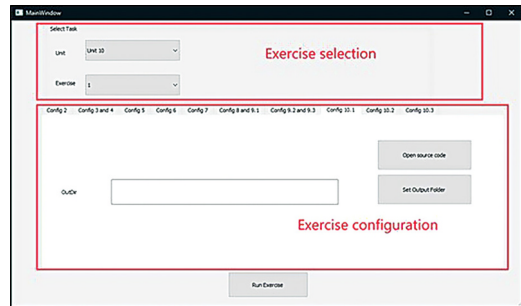


Figure A3.19. The CarboTrain GUI showing the two parts.

TABLE A3.1
Exercises in different tabs

Config	Exercise(s)
2	Exercise 2 of unit 2
3 and 4	Exercises of units 3 and 4
5	Exercises of unit 5
6	Exercises of unit 6
7	Exercises of unit 7
8 and 9.1	Exercises of unit 8 and exercise 1 of unit 9
9.2 and 9.3	Exercises 2 and 3 of unit 9
10.1	Exercise 1 of unit 10
10.2	Exercise 2 of unit 10
10.3	Exercise 3 of unit 10

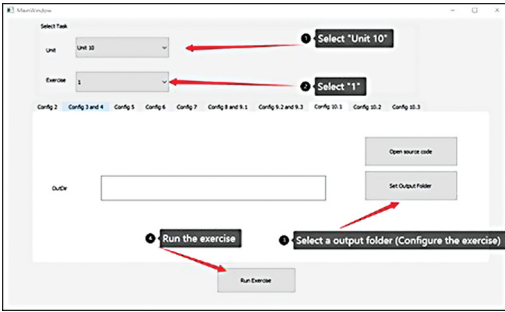


Figure A3.20. Steps to run Exercise 1 in Unit 10.

configuration, it is ready for step 4: clicking “Run Exercise” to start the exercise.

When you have clicked the “Run Exercise” button in each exercise, a pop-up window will show up with the message “Task submitted!” and you need to click “OK” before you can run the exercise.

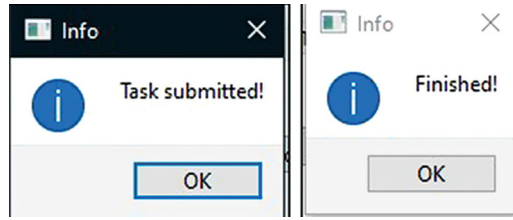


Figure A3.21. These pop-up windows appear when starting the running of an exercise, and when it is finished.

Once an exercise is done, a pop-up window with the message “Finished!” will inform you that the exercise has been completed with no error (Figure A3.21). If any errors occur, please double-check the steps you went through to run the exercise. If you still cannot figure out the cause of the error, please ask instructors.

REFERENCES

- Ahlström, Anders, Guy Schurgers, and Benjamin Smith. 2017. "The large influence of climate model bias on terrestrial carbon cycle simulations." *Environmental Research Letters* 12 (1):014004.
- Ahlström, Anders, Jianyang Xia, Almut Arneth, Yiqi Luo, and Benjamin Smith. 2015. "Importance of vegetation dynamics for future terrestrial carbon cycling." *Environmental Research Letters* 10 (5):054019.
- Akaike, Hirotugu. 1974. "A new look at the statistical model identification." *IEEE Transactions on Automatic Control* 19 (6):716–723.
- Allaire, J J, and Francois Chollet. 2020. Keras: R Interface to 'Keras'. R package version 2.2. 0.
- Anderson, David H. 2013. *Compartmental modeling and tracer kinetics*. Vol. 50: Springer Science & Business Media.
- Anderson, David H, and Towanna Roller. 1991. "Equilibrium points for nonlinear compartmental models." *Mathematical Biosciences* 103 (2):159–201.
- Andren, Olof, and Keith Paustian. 1987. "Barley straw decomposition in the field: A comparison of models." *Ecology* 68 (5):1190–1200.
- Arora, Vivek K, Anna Katavouta, Richard G Williams, Chris D Jones, Victor Brovkin, Pierre Friedlingstein, Jörg Schwinger, Laurent Bopp, Olivier Boucher, and Patricia Cadule. 2020. "Carbon–concentration and carbon–climate feedbacks in CMIP6 models and their comparison to CMIP5 models." *Biogeosciences* 17 (16):4173–4222.
- Bastin, Georges, and V Guffens. 2006. "Congestion control in compartmental network systems." *Systems & Control Letters* 55 (8):689–696.
- Bauer, Peter, Alan Thorpe, and Gilbert Brunet. 2015. "The quiet revolution of numerical weather prediction." *Nature* 525 (7567):47–55.
- Bishop, Christopher M. 2006. "Pattern recognition." *Machine Learning* 128 (9).
- Blackard, J A, M V Finco, E H Helmer, G R Holden, M L Hoppus, D M Jacobs, A J Lister, G G Moisen, M D Nelson, R Riemann, B Ruefenacht, D Salajanu, D L Weyermann, K C Winterberger, T J Brandeis, R L Czaplowski, R E McRoberts, P L Patterson, R P Tymcio. 2008. "Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information." *Remote Sensing of Environment*, 112(4): 1658–1677. <https://doi.org/10.1016/j.rse.2007.08.021>
- Blagodatsky, Sergey, Evgenia Blagodatskaya, Tatyana Yuyukina, and Yakov Kuzyakov. 2010. "Model of apparent and real priming effects: linking microbial activity with soil organic matter decomposition." *Soil Biology and Biochemistry* 42 (8):1275–1283.
- Bolin, Bert. 1981. "Steady state and response characteristics of a simple model of the carbon cycle." *Carbon Cycle Modelling* 16: 315–331.
- Bolin, Bert, and Henning Rodhe. 1973. "A note on the concepts of age distribution and transit time in natural reservoirs." *Tellus* 25 (1):58–62.
- Bolker, Benjamin M., Stephen W. Pacala, and William J Parton Jr. 1998. "Linear analysis of soil decomposition: Insights from the century model." *Ecological Applications* 8 (2):425–439.
- Box, George EP. 1979. "Robustness in the strategy of scientific model building." In *Robustness in statistics*, 201–236. Elsevier.
- Burrows, S M, M Maltrud, X Yang, Q Zhu, N Jeffery, X Shi, D Ricciuto, S Wang, G Bisht, and J Tang. 2020. "The DOE E3SM v1. 1 biogeochemistry configuration: Description and simulated

- ecosystem-climate responses to historical changes in forcing.” *Journal of Advances in Modeling Earth Systems* 12 (9):e2019MS001766.
- Cai, Andong, Guopeng Liang, Xubo Zhang, Wenju Zhang, Ling Li, Yichao Rui, Minggang Xu, and Yiqi Luo. 2018. “Long-term straw decomposition in agro-ecosystems described by a unified three-exponentiation equation with thermal time.” *Science of the Total Environment* 636:699–708.
- Canham, D Charles. 2003. *Models in ecosystem science*: Princeton University Press.
- Carvalho, Alexandre, José A Langa, and James Robinson. 2012. *Attractors for infinite-dimensional non-autonomous dynamical systems*. Vol. 182: Springer Science & Business Media.
- Ciais, Phillipe, Christopher Sabine, Govindasamy Bala, Laurent Bopp, Victor Brovkin, Josep Canadell, Abha Chhabra, Ruth DeFries, James Galloway, Martin Heimann, Christopher Jones, Corinne Le Quéré, Ranga B Myneni, Shilong Piao, Peter Thornton. 2014. “Carbon and other biogeochemical cycles.” In *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.
- Collier, Nathan, Forrest M Hoffman, David M Lawrence, Gretchen Keppel-Aleks, Charles D Koven, William J Riley, Mingquan Mu, and James T Randerson. 2018. “The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation.” *Journal of Advances in Modeling Earth Systems* 10 (11):2731–2754.
- Craiu, Radu V., and Jeffrey S. Rosenthal. 2014. “Bayesian computation via markov chain monte carlo.” *Annual Review of Statistics and Its Application* 1:179–201.
- Cui, Erqian, Kun Huang, Muhammad Altaf Arain, Joshua B Fisher, Deborah N Huntzinger, Akihiko Ito, Yiqi Luo, Atul K Jain, Jiafu Mao, and Anna M Michalak. 2019. “Vegetation functional properties determine uncertainty of simulated ecosystem productivity: A traceability analysis in the East Asian monsoon region.” *Global Biogeochemical Cycles* 33 (6):668–689.
- Davidson, Eric A., Ivan A. Janssens, and Yiqi Luo. 2006. “On the variability of respiration in terrestrial ecosystems: moving beyond Q10.” *Global Change Biology* 12 (2):154–164.
- De Kauwe, M G, M I Disney, T Quaife, P Lewis, and M Williams 2011. “An assessment of the MODIS Collection 5 leaf area index product for a region of mixed coniferous foTSrest.” *Remote Sensing of Environment* 115(2): 767–780.
- Doomaar, J F 1979. Organic matter characteristics of undisturbed and cultivated chernozemic and solonchic A horizons. *Canadian Journal of Soil Science* 59: 349–356.
- Du, Zhenggang, Ensheng Weng, Lifeng Jiang, Yiqi Luo, Jianyang Xia, and Xuhui Zhou. 2018. “Carbon–nitrogen coupling under three schemes of model representation: A traceability analysis.” *Geoscientific Model Development* 11 (11):4399–4416.
- Evensen, Geir. 2009. *Data assimilation: The ensemble Kalman filter*: Springer Science & Business Media.
- Fisher, Rosie A, Charles D Koven, William RL Anderegg, Bradley O Christoffersen, Michael C Dietze, Caroline E Farrior, Jennifer A Holm, George C Hurtt, Ryan G Knox, and Peter J Lawrence. 2018. “Vegetation demographics in Earth System Models: A review of progress and priorities.” *Global Change Biology* 24 (1):35–54.
- Forrester, Jay Wright. 1961. *Industrial dynamics*. Cambridge, Massachusetts: MIT Press.
- Forster, Piers, Venkatachalam Ramaswamy, Paulo Artaxo, Terje Berntsen, Richard Betts, David W Fahey, James Haywood, Judith Lean, David C Lowe, and Gunnar Myhre. 2007. “Changes in atmospheric constituents and in radiative forcing Chapter 2.” In *Climate change 2007. The physical science basis*.
- Friedlingstein, Pierre, Peter Cox, Richard Betts, Laurent Bopp, Werner von Bloh, Victor Brovkin, Patricia Cadule, Scott Doney, Michael Eby, and Inez Fung. 2006. “Climate–carbon cycle feedback analysis: results from the C4MIP model intercomparison.” *Journal of Climate* 19 (14):3337–3353.
- Friedlingstein, Pierre, Malte Meinshausen, Vivek K Arora, Chris D Jones, Alessandro Anav, Spencer K Liddicoat, and Reto Knutti. 2014. “Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks.” *Journal of Climate* 27 (2):511–526.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2014. *Bayesian data analysis*. Vol. 2. Boca Raton, FL, CRC Press.
- Gelman, Andrew, Gareth O. Roberts, and Walter R. Gilks. 1996. “Efficient Metropolis jumping rules.” *Bayesian Statistics* 5 (599–608):42.
- Giglio, L, J T Randerson, G R van der Werf, P S Kasibhatla, G J Collatz, D C Morton, and R S DeFries 2010. “Assessing variability and long-term trends in burned area by merging multiple satellite fire products.” *Biogeosciences* 7(3): 1171–1186. <https://doi.org/10.5194/bg-7-1171-2010>
- Gill, Allison L., Marc-André Giasson, Rieka Yu, and Adrien C. Finzi. 2017. “Deep peat warming increases surface methane and carbon dioxide emissions in a black spruce-dominated ombrotrophic bog.” *Global Change Biology* 23 (12):5398–5411.

- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*: MIT Press.
- Guckenheimer, John, and Philip Holmes. 2013. *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*. Vol. 42: Springer Science & Business Media.
- Guo, F, RS Yost, NV Hue, CI Evensen, and JA Silva. 2000. "Changes in phosphorus fractions in soils under intensive plant growth." *Soil Science Society of America Journal* 64 (5):1681–1689.
- Haario, Heikki, Eero Saksman, and Johanna Tamminen. 2001. "An adaptive Metropolis algorithm." *Bernoulli* 7 (2):223–242.
- Haddix, Michelle, Alain Plante, Richard Conant, Johan Six, J Megan Steinweg, Kim Magrini-Blair, Rhae Drijber, Sherri Morris, and Eldor Paul. 2011. "The role of soil characteristics on temperature sensitivity of soil organic matter." *Soil Science Society of America Journal* 75 (1):56–68.
- Hanson, PJ, AL Gill, X Xu, JR Phillips, DJ Weston, RK Kolka, JS Riggs, and LA Hook. 2016. "Intermediate-scale community-level flux of CO₂ and CH₄ in a Minnesota peatland: putting the SPRUCE project in a global context." *Biogeochemistry* 129 (3):255–272.
- Hanson, PJ, JR Phillips, JS Riggs, and WR Nettles. 2017b. *SPRUCE Large-Collar in Situ CO₂ and CH₄ Flux Data for the SPRUCE Experimental Plots: Whole-Ecosystem-Warming*. Oak Ridge, TN, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy. <https://doi.org/10.3334/CDIAC/spruce.034>
- Paul J Hanson, Natalie A Griffiths, Colleen M Iversen, Richard J Norby, Stephen D Sebestyen, Jana R Phillips, Jeffrey P Chanton, Randall K Kolka, Avni Malhotra, Keith C Oleheiser, Jeffrey M Warren, Xiaoying Shi, Xiaojuan Yang, Jiafu Mao, Daniel M Ricciuto 2020. "Rapid net carbon loss from a whole-ecosystem warmed peatland." *AGU Advances* 1(3): e2020AV000163.
- Hanson, Paul J, Jeffery S Riggs, W Robert Nettles, Jana R Phillips, Misha B Krassovski, Leslie A Hook, Lianhong Gu, Andrew D Richardson, Donald M Aubrecht, and Daniel M Ricciuto. 2017a. "Attaining whole-ecosystem warming using air and deep-soil heating methods with an elevated CO₂ atmosphere." *Biogeosciences* 14 (4):861–883.
- Hararuk, Oleksandra, Matthew J Smith, and Yiqi Luo. 2015. "Microbial models with data-driven parameters predict stronger soil carbon responses to climate change." *Global Change Biology* 21 (6):2439–2453.
- Hararuk, Oleksandra, Jianyang Xia, and Yiqi Luo. 2014. "Evaluation and improvement of a global land model against soil carbon data using a Bayesian Markov chain Monte Carlo method." *Journal of Geophysical Research: Biogeosciences* 119 (3):403–417.
- Hastie, Tibshirani, and Robert Tibshirani. 2009 & Friedman, J. (2008). *The elements of statistical learning; data mining, inference and prediction*, Second edition. Springer, Springer Science+Business Media, New York, NY.
- Hastings, W Keith. 1970. "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika* 57:97–109.
- Hengl, Tomislav, Jorge Mendes de Jesus, Gerard BM Heuvelink, Maria Ruiperez Gonzalez, Milan Kilibarda, Aleksandar Blagotić, Wei Shangguan, Marvin N Wright, Xiaoyuan Geng, and Bernhard Bauer-Marschallinger. 2017. "SoilGrids250m: Global gridded soil information based on machine learning." *PLoS One* 12 (2):e0169748.
- Horgan, John. 2016. "Bayes Theorem: What's the big deal." *Scientific American*, January, 4.
- Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. "Unbiased recursive partitioning: A conditional inference framework." *Journal of Computational and Graphical Statistics* 15 (3):651–674.
- Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2015. "Ctree: Conditional inference trees." *The Comprehensive R Archive Network* 8. 1–34.
- Hou, Enqing, Xiang Tan, Marijke Heenan, and Dazhi Wen. 2018. "A global dataset of plant available and unavailable phosphorus in natural soils derived by Hedley method." *Scientific Data* 5 (1):1–13.
- Hou, Enqing, Xingjie Lu, Lifen Jiang, Dazhi Wen, and Yiqi Luo. 2019. "Quantifying soil phosphorus dynamics: A data assimilation approach." *Journal of Geophysical Research: Biogeosciences* 124 (7):2159–2173.
- Huang, Y Y, D Zhu, P Ciaias, B Guenet, Y Huang, D S Goll, M Guimberteau, A Jornet-Puig, X J Lu, and Y Q Luo. 2018a. "Matrix-based sensitivity assessment of soil organic carbon storage: A case study from the ORCHIDEE-MICT Model." *Journal of Advances in Modeling Earth Systems* 10 (8):1790–1808. <https://doi.org/10.1029/2017ms001237>
- Huang, Yuanyuan, Jiang Jiang, Shuang Ma, Daniel Ricciuto, Paul J Hanson, and Yiqi Luo. 2017. "Soil thermal dynamics, snow cover, and frozen depth under five temperature treatments in an ombrotrophic bog: Constrained forecast with data assimilation." *Journal of Geophysical Research: Biogeosciences* 122 (8):2046–2063.
- Huang, Yuanyuan, Mark Stacy, Jiang Jiang, Nilutpal Sundi, Shuang Ma, Volodymyr Saruta, Chang Gyo Jung, Zheng Shi, Jianyang Xia, and Paul J Hanson. 2019. "Realized ecological forecast through an interactive Ecological Platform for Assimilating

- Data (EcoPAD, v1. 0) into models." *Geoscientific Model Development* 12 (3):1119–1137.
- Huang, Yuanyuan, Xingjie Lu, Zheng Shi, David Lawrence, Charles D Koven, Jianyang Xia, Zhenggang Du, Erik Kluzek, and Yiqi Luo. 2018b. "Matrix approach to land carbon cycle modeling: A case study with the Community Land Model." *Global Change Biology* 24 (3):1394–1404.
- Hugelius, G, J G Bockheim, P Camill, B Elberling, G Grosse, J W Harden, K Johnson, T Jorgenson, C D Koven, P Kuhry, G Michaelson, U Mishra, J Palmtag, C-L Ping, J O'Donnell, L Schirmermeister, EA G Schuur, Y Sheng, L C Smith, J Strauss, Z Yu. 2013. "A new data set for estimating organic carbon storage to 3 m depth in soils of the northern circumpolar permafrost region." *Earth System Science Data* 5(2): 393–402. <https://doi.org/10.5194/essd-5-393-2013>
- Jacquez, John A, and Carl P Simon. 1993. "Qualitative theory of compartmental systems." *Siam Review* 35 (1):43–79.
- Jiang, Jiang, Yuanyuan Huang, Shuang Ma, Mark Stacy, Zheng Shi, Daniel M Ricciuto, Paul J Hanson, and Yiqi Luo. 2018. "Forecasting responses of a northern peatland carbon cycle to elevated CO₂ and a gradient of experimental warming." *Journal of Geophysical Research: Biogeosciences* 123 (3):1057–1071.
- Jiang, Lifan, Zheng Shi, Jianyang Xia, Junyi Liang, Xingjie Lu, Ying Wang, and Yiqi Luo. 2017. "Transient traceability analysis of land carbon storage dynamics: Procedures and its application to two forest ecosystems." *Journal of Advances in Modeling Earth Systems*, 2822–2835.
- Jones, Chris, Eddy Robertson, Vivek Arora, Pierre Friedlingstein, Elena Shevliakova, Laurent Bopp, Victor Brovkin, Tomohiro Hajima, Etsushi Kato, and Michio Kawamiya. 2013. "Twenty-first-century compatible CO₂ emissions and airborne fraction simulated by CMIP5 earth system models under four representative concentration pathways." *Journal of Climate* 26 (13):4398–4413.
- Jones, Thouis R, Anne E Carpenter, Michael R Lamprecht, Jason Moffat, Serena J Silver, Jennifer K Grenier, Adam B Castoreno, Ulrike S Eggert, David E Root, and Polina Golland. 2009. "Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning." *Proceedings of the National Academy of Sciences* 106 (6):1826–1831.
- Keenan, Trevor F, Eric A Davidson, J William Munger, and Andrew D Richardson. 2013. "Rate my data: Quantifying the value of ecological data for the development of models of the terrestrial carbon cycle." *Ecological Applications* 23 (1):273–286.
- Kloeden, Peter E, and Martin Rasmussen. 2011. *Nonautonomous dynamical systems*: American Mathematical Soc.
- Koven, C D, W J Riley, Z M Subin, J Y Tang, M S Torn, W D Collins, G B Bonan, D M Lawrence, and S C Swenson. 2013. "The effect of vertically resolved soil biogeochemistry and alternate soil C and N models on C dynamics of CLM4." *Biogeosciences* 10 (11):7109.
- Kuzyakov, Yakov. 2010. "Priming effects: interactions between living and dead organic matter." *Soil Biology and Biochemistry* 42 (9):1363–1371.
- Lardy, Romain, Gianni Bellocchi, and J-F Soussana. 2011. "A new method to determine soil organic carbon equilibrium." *Environmental Modelling & Software* 26 (12):1759–1763.
- Lawrence, David M, Rosie A Fisher, Charles D Koven, Keith W Oleson, Sean C Swenson, Gordon Bonan, Nathan Collier, Bardan Ghimire, Leo van Kampenhou, and Daniel Kennedy. 2019. "The Community Land Model version 5: Description of new features, benchmarking, and impact of forcing uncertainty." *Journal of Advances in Modeling Earth Systems* 11 (12):4245–4287.
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." *Nature* 521 (7553):436.
- LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86 (11):2278–2324.
- Lee, Hung-yi. 2016. "Deep Learning Tutorial." https://speech.ee.ntu.edu.tw/~tlkagk/slide/Tutorial_HYLee_Deep.pdf.
- LeVeque, Randall J. 2007. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*, SIAM.
- Lewis, John M, Sivaramakrishnan Lakshminarayanan, and Sudarshan Dhall. 2006. *Dynamic data assimilation: A least squares approach*. Vol. 13: Cambridge University Press.
- Liang, Junyi, Dejun Li, Zheng Shi, James M Tiedje, Jizhong Zhou, Edward AG Schuur, Konstantinos T Konstantinidis, and Yiqi Luo. 2015. "Methods for estimating temperature sensitivity of soil organic matter based on incubation data: A comparative evaluation." *Soil Biology and Biochemistry* 80:127–135.
- Liang, Junyi, Jianyang Xia, Zheng Shi, Lifan Jiang, Shuang Ma, Xingjie Lu, Marguerite Mauritz, Susan M Natali, Elaine Pegoraro, and Christopher Ryan Penton. 2018. "Biotic responses buffer warming-induced soil organic carbon loss in Arctic tundra." *Global Change Biology* 24 (10):4946–4959.
- Lu, Dan, Daniel Ricciuto, Miroslav Stoyanov, and Lianhong Gu. 2018a. "Calibration of the E3SM land model using surrogate-based global

- optimization." *Journal of Advances in Modeling Earth Systems* 10 (6):1337–1356.
- Lu, Xingjie, Ying-Ping Wang, Yiqi Luo, and Lifan Jiang. 2018b. "Ecosystem carbon transit versus turnover times in response to climate warming and rising atmospheric CO₂ concentration." *Biogeosciences* 15 (21):6559–6572.
- Lu, Xingjie, Zhenggang Du, Yuanyuan Huang, David Lawrence, Erik Kluzek, Nathan Collier, Danica Lombardozi, Negin Sobhani, Edward AG Schuur, and Yiqi Luo. 2020. "Full Implementation of Matrix Approach to Biogeochemistry Module of CLM5." *Journal of Advances in Modeling Earth Systems* 12 (11):e2020MS002105.
- Luo, Yiqi, Anders Ahlström, Steven D Allison, Niels H Batjes, Victor Brovkin, Nuno Carvalhais, Adrian Chappell, Philippe Ciais, Eric A Davidson, and Adrien Finzi. 2016. "Toward more realistic projections of soil carbon dynamics by Earth system models." *Global Biogeochemical Cycles* 30 (1):40–56.
- Luo, Yiqi, Trevor F Keenan, and Matthew Smith. 2015. "Predictability of the terrestrial carbon cycle." *Global Change Biology* 21 (5):1737–1751.
- Luo, Yiqi, Lianhai Wu, Jeffrey A Andrews, Luther White, Roser Matamala, Karina VR Schäfer, and William H Schlesinger. 2001. "Elevated CO₂ differentiates ecosystem carbon processes: Deconvolution analysis of Duke Forest FACE data." *Ecological Monographs* 71 (3):357–376.
- Luo, Yiqi, Kiona Ogle, Colin Tucker, Shenfeng Fei, Chao Gao, Shannon LaDeau, James S Clark, and David S Schimel. 2011. "Ecological forecasting and data assimilation in a data-rich era." *Ecological Applications* 21 (5):1429–1442.
- Luo, Yiqi, James T Randerson, P Friedlingstein, K Hibbard, F Hoffman, D Huntzinger, CD Jones, C Koven, D Lawrence, and DJ Li. 2012. "A framework for benchmarking land models." *Biogeosciences* 9 (10):3857–3874.
- Luo, Yiqi and James F. Reynolds 1999. "Validity of extrapolating field CO₂ experiments to predict carbon sequestration in natural ecosystems." *Ecology* 80(5): 1568–1583.
- Luo, Yiqi, and Edward AG Schuur. 2020. "Model parameterization to represent processes at unresolved scales and changing properties of evolving systems." *Global Change Biology* 26 (3):1109–1117.
- Luo, Yiqi, Zheng Shi, Xingjie Lu, Jianyang Xia, Junyi Liang, Jiang Jiang, Ying Wang, Matthew J Smith, Lifan Jiang, and Anders Ahlström. 2017. "Transient dynamics of terrestrial carbon storage: mathematical foundation and its applications." *Biogeosciences* 14 (1):145.
- Luo, Yiqi, BO Su, William S Currie, Jeffrey S Dukes, Adrien Finzi, Ueli Hartwig, Bruce Hungate, Ross E McMurtrie, RAM Oren, and William J Parton. 2004. "Progressive nitrogen limitation of ecosystem responses to rising atmospheric carbon dioxide." *Bioscience* 54 (8):731–739.
- Luo, Yiqi, and Ensheng Weng. 2011. "Dynamic disequilibrium of the terrestrial carbon cycle under global change." *Trends in Ecology & Evolution* 26 (2):96–104.
- Luo, Yiqi, Ensheng Weng, Xiaowen Wu, Chao Gao, Xuhui Zhou, and Li Zhang. 2009. "Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models." *Ecological Applications* 19 (3):571–574.
- Luo, Yiqi, Luther W White, Josep G Canadell, Evan H DeLucia, David S Ellsworth, Adrien Finzi, John Lichter, and William H Schlesinger. 2003. "Sustainability of terrestrial carbon sequestration: A case study in Duke Forest with inversion approach." *Global Biogeochemical Cycles* 17 (1):1021. <https://doi.org/10.1029/2002GB001923>
- Ma, Shuang, Jiang Jiang, Yuanyuan Huang, Zheng Shi, Rachel M Wilson, Daniel Ricciuto, Stephen D Sebestyen, Paul J Hanson, and Yiqi Luo. 2017. "Data-constrained projections of methane fluxes in a northern Minnesota peatland in response to elevated CO₂ and warming." *Journal of Geophysical Research: Biogeosciences* 122 (11):2841–2861.
- Malhotra, Avni, Deanne J Brice, Joanne Childs, Jake D Graham, Erik A Hobbie, Holly Vander Stel, Sarah C Feron, Paul J Hanson, and Colleen M Iversen. 2020. "Peatland warming strongly increases fine-root growth." *Proceedings of the National Academy of Sciences* 117 (30):17627–17634.
- Matis, James H, Bernard C Patten, and Gary C White. 1979. *Compartmental analysis of ecosystem models*. Vol. 10: International Cooperative Publishing House.
- Medlyn, Belinda E, Sönke Zaehle, Martin G De Kauwe, Anthony P Walker, Michael C Dietze, Paul J Hanson, Thomas Hickler, Atul K Jain, Yiqi Luo, and William Parton. 2015. "Using ecosystem experiments to improve vegetation models." *Nature Climate Change* 5 (6):528–534.
- Meng, L, P G M Hess, N M Mahowald, J B Yavitt, W J Riley, Z M Subin, ... D R Fuka 2012. "Sensitivity of wetland methane emissions to model assumptions: Application and model testing against site observations." *Biogeosciences* 9(7): 2793–2819. <https://doi.org/10.5194/bg-9-2793-2012>
- Metzler, Holger, Markus Müller, and Carlos A Sierra. 2018. "Transit-time and age distributions for nonlinear time-dependent compartmental

- systems." *Proceedings of the National Academy of Sciences* 115 (6):1150–1155.
- Metzler, Holger, and Carlos A Sierra. 2018. "Linear autonomous compartmental models as continuous-time Markov chains: Transit-time and age distributions." *Mathematical Geosciences* 50 (1):1–34.
- Mullholland, Robert J, and Marvin S Keener. 1974. "Analysis of linear compartment models for ecosystems." *Journal of Theoretical Biology* 44 (1):105–116.
- Müller, Markus, and Carlos A Sierra. 2017. "Application of input to state stability to reservoir models." *Theoretical Ecology* 10 (4):451–475.
- Murphy, Kevin P. 2012. *Machine learning: a probabilistic perspective*: MIT Press.
- Myhre, G, D Shindell, F-M Bréon, W Collins, J Fuglested, J Huang, D Koch, J-F Lamarque, D Lee, B Mendoza, T Nakajima, A Robock, G Stephens, T Takemura and H Zhang. 2013. "Anthropogenic and natural radiative forcing." In: *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*, edited by: TF Stocker, D Qin, G-K Plattner, M Tignor, SK Allen, J Boschung, A Nauels, Y Xia, V Bex and PM Midgley (eds.]. Cambridge, United Kingdom and New York, NY, USA, Cambridge University Press.
- Norby, Richard J, Joanne Childs, Paul J Hanson, and Jeffrey M Warren. 2019. "Rapid loss of an ecosystem engineer: Sphagnum decline in an experimentally warmed bog." *Ecology and Evolution* 9 (22):12571–12585.
- Odum, H.T. 1971. *Environment, Power, and Society*. Wiley-Interscience, New York.
- Odum, Eugene Pleasants, and Gary W Barrett. 1971. *Fundamentals of ecology*. Vol. 3: Saunders Philadelphia.
- Oreskes, Naomi. 2003. "The role of quantitative models in science." In *Models in ecosystem science*, edited by: CD Canham, JJ Cole, and WK Lauenroth:13–31. Princeton, NJ, Princetaon, University Press.
- Oreskes, Naomi, Kristin Shrader-Frechette, and Kenneth Belitz. 1994. "Verification, validation, and confirmation of numerical models in the earth sciences." *Science* 263 (5147):641–646.
- Parton, W J, J W B Stewart, and C V Cole. 1988. "Dynamics of C, N, P and S in grassland soils – a model." *Biogeochemistry* 5 (1):109–131. <https://doi.org/10.1007/bf02180320>
- Philippe, Ciais, Christopher Sabine, Govindasamy Bala, Laurent Bopp, Victor Brovkin, Josep Canadell, Abha Chhabra, Ruth DeFries, James Galloway, Martin Heimann, Christopher Jones, Corinne Le Quéré, Ranga B Myneni, Shilong Piao, Peter Thornton. 2014. "Carbon and other biogeochemical cycles." In *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.
- Platt, John R. 1964. Strong inference. *Science* 146(3642): 347–353.
- Poplin, R, A V Varadarajan, K Blumer, Y Liu, M McConnell, G Corrado, L Peng, and D Webster. 2017. "Predicting cardiovascular risk factors from retinal fundus photographs using deep learning. arXiv 2017." *arXiv preprint arXiv:1708.09843*.
- Qu, Yang, Shamil Maksyutov, and Qianlai Zhuang. 2018. "An efficient method for accelerating the spin-up process for process-based biogeochemistry models." *Biogeosciences* 15 (13):3967–3973.
- Rafique, Rashid, Jianyang Xia, Oleksandra Hararuk, Ghassem R Asrar, Guoyong Leng, Yingping Wang, and Yiqi Luo. 2016. "Divergent predictions of carbon storage between two global land models: Attribution of the causes through traceability analysis." *Earth System Dynamics (Online)* 7 (PNNL-SA-115823).
- Rafique, Rashid, Jianyang Xia, Oleksandra Hararuk, Guoyong Leng, Ghassem Asrar, and Yiqi Luo. 2017. "Comparing the performance of three land models in global C cycle simulations: A detailed structural analysis." *Land Degradation Development* 28 (2):524–533.
- Rasmussen, Martin. 2007. *Attractivity and bifurcation for nonautonomous dynamical systems*: Springer.
- Rasmussen, Martin, Alan Hastings, Matthew J Smith, Folashade B Agosto, Benito M Chen-Charpentier, Forrest M Hoffman, Jiang Jiang, Katherine EO Todd-Brown, Ying Wang, and Ying-Ping Wang. 2016. "Transit times and mean ages for nonautonomous and autonomous compartmental systems." *Journal of Mathematical Biology* 73 (6–7):1379–1398.
- Reichstein, Markus, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, and Nuno Carvalhais. 2019. "Deep learning and process understanding for data-driven Earth system science." *Nature* 566 (7743):195.
- Rey, ANA, and Paul Jarvis. 2006. "Modelling the effect of temperature on carbon mineralization rates across a network of European forest sites (FORCAST)." *Global Change Biology* 12 (10):1894–1908.
- Ricciuto, Daniel, Khachik Sargsyan, and Peter Thornton. 2018. "The impact of parametric uncertainties on biogeochemistry in the E3SM land model." *Journal of Advances in Modeling Earth Systems* 10 (2):297–319.
- Richardson, Andrew D, Koen Hufkens, Thomas Milliman, Donald M Aubrecht, Morgan E Furze, Bijan Seyednasrollah, Misha B Krassovski, John M Latimer, W Robert Nettles, and Ryan R Heiderman. 2018. "Ecosystem warming extends vegetation

- activity but heightens vulnerability to cold temperatures." *Nature* 560 (7718):368–371.
- Richardson, Andrew D, Mathew Williams, David Y Hollinger, David JP Moore, D Bryan Dail, Eric A Davidson, Neal A Scott, Robert S Evans, Holly Hughes, and John T Lee. 2010. "Estimating parameters of a forest ecosystem C model with measurements of stocks and fluxes as joint constraints." *Oecologia* 164 (1):25–40.
- Riley, W J, Z M Subin, D M Lawrence, S C Swenson, M S Torn, L Meng, N M Mahowald, and P Hess. 2011. "Barriers to predicting changes in global terrestrial methane fluxes: analyses using CLM4Me, a methane biogeochemistry model integrated in CESM." *Biogeosciences* 8 (7):1925–1953.
- Russell, Bryan C, Antonio Torralba, Kevin P Murphy, and William T Freeman. 2008. "LabelMe: A database and web-based tool for image annotation." *International Journal of Computer Vision* 77 (1–3):157–173.
- Rykiel Jr, Edward J. 1996. "Testing ecological models: the meaning of validation." *Ecological Modelling* 90 (3):229–244.
- Saunois, Marielle, Ann R Stavert, Ben Poulter, Philippe Bousquet, Josep G Canadell, Robert B Jackson, Peter A Raymond, Edward J Dlugokencky, Sander Houweling, and Prabir K Patra. 2020. "The global methane budget 2000–2017." *Earth System Science Data* 12 (3):1561–1623.
- Schädel, Christina, Edward AG Schuur, Rosvel Bracho, Bo Elberling, Christian Knoblauch, Hanna Lee, Yiqi Luo, Gaius R Shaver, and Merritt R Turetsky. 2014. "Circumpolar assessment of permafrost C quality and its vulnerability over time using long-term incubation data." *Global Change Biology* 20 (2):641–652.
- Schädel, Christina, Yiqi Luo, R David Evans, Shenfeng Fei, and Sean M Schaeffer. 2013. "Separating soil CO₂ efflux into C-pool-specific decay rates via inverse analysis of soil incubation data." *Oecologia* 171 (3):721–732.
- Scheffer, Marten, Steve Carpenter, Jonathan A Foley, Carl Folke, and Brian Walker. 2001. "Catastrophic shifts in ecosystems." *Nature* 413 (6856):591–596.
- Segers, R. 1998. "Methane production and methane consumption: A review of processes underlying wetland methane fluxes." *Biogeochemistry* 41 (1): 23–51. <http://www.jstor.org/stable/1469307>
- Sellers, P J, D A Randall, G J Collatz, J A Berry, C B Field, D A Dazlich, C Zhang, G D Collelo, and L Bounoua. 1996. "A revised land surface parameterization (SiB2) for atmospheric GCMs. Part I: Model formulation." *Journal of Climate* 9 (4):676–705.
- Shannon, Claude Elwood. 1948. "A mathematical theory of communication." *The Bell System Technical Journal* 27 (3):379–423.
- Shi, Xiaoying, Daniel M Ricciuto, Peter E Thornton, Xiaofeng Xu, Fengming Yuan, Richard J Norby, Anthony P Walker, Jeffrey M Warren, Jiafu Mao, and Paul J Hanson. 2021. "Extending a land-surface model with Sphagnum moss to simulate responses of a northern temperate bog to whole ecosystem warming and elevated CO₂." *Biogeosciences* 18 (2):467–486.
- Shi, Xiaoying, Peter E Thornton, Daniel M Ricciuto, Paul J Hanson, Jiafu Mao, Stephen D Sebestyen, Natalie A Griffiths, and Gautam Bisht. 2015. "Representing northern peatland microtopography and hydrology within the Community Land Model." *Biogeosciences* 12 (21):6463–6477.
- Shi, Zheng, Sean Crowell, Yiqi Luo, and Berrien Moore. 2018. "Model structures amplify uncertainty in predicted soil carbon responses to climate change." *Nature Communications* 9 (1):2171.
- Shi, Zheng, Yuanhe Yang, Xuhui Zhou, Ensheng Weng, Adrien C Finzi, and Yiqi Luo. 2016. "Inverse analysis of coupled carbon–nitrogen cycles against multiple datasets at ambient and elevated CO₂." *Journal of Plant Ecology* 9 (3):285–295.
- Sierra, Carlos A, Verónica Ceballos-Núñez, Holger Metzler, and Markus Müller. 2018. "Representing and understanding the carbon cycle using the theory of compartmental dynamical systems." *Journal of Advances in Modeling Earth Systems* 10 (8):1729–1734.
- Sierra, Carlos A, and Markus Müller. 2015. "A general mathematical framework for representing soil organic matter dynamics." *Ecological Monographs* 85 (4):505–524.
- Sierra, Carlos A, Markus Müller, Holger Metzler, Stefano Manzoni, and Susan E Trumbore. 2017. "The muddle of ages, turnover, transit, and residence times in the carbon cycle." *Global Change Biology* 23 (5):1763–1773.
- Sitch, Stephen, Pierre Friedlingstein, Nicolas Gruber, Steve D Jones, Guillermo Murray-Tortarolo, Anders Ahlström, Scott C Doney, H Graven, Christoph Heinze, and Chris Huntingford. 2015. "Recent trends and drivers of regional sources and sinks of carbon dioxide." *Biogeosciences* 12 (3):653–679.
- Sontag, Eduardo D. 2013. *Mathematical control theory: deterministic finite dimensional systems*. Vol. 6: Springer Science & Business Media.
- Spall, James C. 2005. *Introduction to stochastic search and optimization: Estimation, simulation, and control*. Vol. 65: John Wiley & Sons.

- Stanford, George, and SJ Smith. 1972. "Nitrogen mineralization potentials of soils." *Soil Science Society of America Journal* 36 (3):465–472.
- Strogatz, Steven H. 1994. "Nonlinear dynamics and chaos: With applications to physics, Biology, Chemistry and Engineering" *Computers in Physics* 8,532. Addison-Wesley, Reading, MA.
- Sun, Y, D S Goll, J F Chang, P Ciais, B Guenet, J Helfenstein, Y Y Huang, R Lauerwald, F Maignan, V Naipal, Y L Wang, H Yang, and H C Zhang. 2021. "Global evaluation of the nutrient-enabled version of the land surface model ORCHIDEE-CNP v1.2 (r5986)." *Geoscientific Model Development* 14 (4):1987–2010. <https://doi.org/10.5194/gmd-14-1987-2021>.
- Tang, JY, and W J Riley. 2013. "A total quasi-steady-state formulation of substrate uptake kinetics in complex networks and an example application to microbial litter decomposition." *Biogeosciences* 10 (12):8329–8351.
- Tang, Jinyun, and William J Riley. 2015. "Weaker soil carbon–climate feedbacks resulting from microbial and abiotic interactions." *Nature Climate Change* 5 (1):56–60.
- Tao, Feng, Zhenghu Zhou, Yuanyuan Huang, Qianyu Li, Xingjie Lu, Shuang Ma, Xiaomeng Huang, Yishuang Liang, Gustaf Hugelius, Lifan Jiang, Russell Doughty, Zehao Ren, and Yiqi Luo. 2020. "Deep learning optimizes data-driven representation of soil organic carbon in earth system model over the conterminous United States." *Frontiers in Big Data* 3 (17). <https://doi.org/10.3389/fdata.2020.00017>.
- Taussky, Olga. 1949. "A recurring theorem on determinants." *The American Mathematical Monthly* 56 (10P1):672–676.
- Thornton, Peter E, Beverley E Law, Henry L Gholz, Kenneth L Clark, Eva Falge, David S Ellsworth, Allen H Goldstein, Russell K Monson, David Hollinger, and Michael Falk. 2002. "Modeling and measuring the effects of disturbance history and climate on carbon and water budgets in evergreen needleleaf forests." *Agricultural and Forest Meteorology* 113 (1–4):185–222.
- Thornton, Peter E, and Nan A Rosenbloom. 2005. "Ecosystem model spin-up: Estimating steady state conditions in a coupled terrestrial carbon and nitrogen cycle model." *Ecological Modelling* 189 (1–2):25–48.
- Todd-Brown, K E O, J T Randerson, W M Post, F M Hoffman, C Tarnocai, E A G Schuur, and S D Allison. 2013. "Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations." *Biogeosciences* 10 (3):1717–1736.
- Torn, Margaret S, Susan E Trumbore, Oliver A Chadwick, Peter M Vitousek, and David M Hendricks. 1997. "Mineral control of soil organic carbon storage and turnover." *Nature* 389 (6647):170.
- Troy Baisden, W, and Ronald Amundson. 2003. "An analytical approach to ecosystem biogeochemistry modeling." *Ecological Applications* 13 (3):649–663.
- Tuomi, Mikko, T Thum, H Järvinen, S Fronzek, B Berg, M Harmon, JA Trofymow, S Sevanto, and J Liski. 2009. "Leaf litter decomposition—Estimates of global variability based on Yasso07 model." *Ecological Modelling* 220 (23):3362–3371.
- Turner, Monica G. 2018. "Yellowstone Rebounded from an Epic 1988 Fire—That May Be Harder in Future." *Scientific American*. August 28, 2018.
- VanderJagt, Dorothy J, Marti Morales, Tom D Thacher, Marco Diaz, and Robert H Glew. 2001. "Bioelectrical impedance analysis of the body composition of nigerian children with calcium-deficiency rickets." *Journal of Tropical Pediatrics* 47 (2):92–97.
- Venables, William N, and Brian D Ripley. 2013. *Modern applied statistics with S-PLUS: Fourth edition* Springer Science & Business Media.
- Walter, B 1998. Development of a Process-Based Model to Derive Methane Emissions from Natural Wetlands for Climate Studies. PhD Thesis, University of Hamburg, Hamburg. Cite as: <http://hdl.handle.net/21.11116/0000-0009-E32C-6>
- Walter, Bernadette P, and Martin Heimann. 2000. "A process-based, climate-sensitive model to derive methane emissions from natural wetlands: Application to five wetland sites, sensitivity to model parameters, and climate." *Global Biogeochemical Cycles* 14 (3):745–765.
- Wang, Y P, B C Chen, William R Wieder, M Leite, Belinda E Medlyn, Martin Rasmussen, Matthew J Smith, Folashade B Augusto, Forrest Hoffman, and Y Q Luo. 2014. "Oscillatory behavior of two nonlinear microbial models of soil carbon decomposition." *Biogeosciences* 11:1817–1831.
- Wang, YP, J Jiang, Benito M Chen-Charpentier, Folashade B Augusto, Alan Hastings, Forrest M Hoffman, Martin Rasmussen, Matthew J Smith, Katherine EO Todd-Brown, and Y Wang. 2016. "Responses of two nonlinear microbial models to warming and increased carbon input." *Biogeosciences* 13: 887–902.

- Wang, Ying Ping, Eva Kowalczyk, Ray Leuning, Gab Abramowitz, Michael R Raupach, Bernard Pak, Eva van Gorsel, and Ashok Luhar. 2011. "Diagnosing errors in a land surface model (CABLE) in the time and frequency domains." *Journal of Geophysical Research: Biogeosciences* 116 (G1).
- Wania, R, I Ross, and I C Prentice. 2010. "Implementation and evaluation of a new methane model within a dynamic global vegetation model: LPJ-WHyMe v1. 3.1." *Geoscientific Model Development* 3 (2):565–584.
- Wasserman, Larry. 2004. *All of statistics: a concise course in statistical inference*. Vol. 26:Springer.
- Weng, Ensheng, and Yiqi Luo. 2008. "Soil hydrological properties regulate grassland ecosystem responses to multifactor global change: A modeling analysis." *Journal of Geophysical Research: Biogeosciences* 113 (G3).
- Weng, Ensheng, and Yiqi Luo. 2011. "Relative information contributions of model vs. data to short- and long-term forecasts of forest carbon dynamics." *Ecological Applications* 21 (5):1490–1505.
- Wieder, W R, A S Grandy, C M Kallenbach, P G Taylor, and G B Bonan. 2015. "Representing life in the Earth system with soil microbial functional traits in the MIMICS model." *Geoscientific Model Development Discussions* 8 (2).
- Wieder, William R, Gordon B Bonan, and Steven D Allison. 2013. "Global soil carbon projections are improved by modelling microbial processes." *Nature Climate Change* 3 (10):909–912.
- Wikipedia. "Sensitivity analysis." https://en.wikipedia.org/w/index.php?title=Sensitivity_analysis&oldid=985820253.
- Wikipedia. "Variance-based sensitivity analysis." https://en.wikipedia.org/w/index.php?title=Variance-based_sensitivity_analysis&oldid=942429004
- Wilson, R M, A M Hopple, M M Tfaily, S D Sebestyen, C W Schadt, L Pfeifer-Meister, Cassandra Medvedeff, KJ McFarlane, J E Kostka, and M Kolton. 2016. "Stability of peatland carbon to rising temperatures." *Nature Communications* 7 (1):1–10.
- Wu, Zhendong, Gustaf Hugelius, Yiqi Luo, Benjamin Smith, Jianyang Xia, Rasmus Fensholt, Veiko Lehsten, and Anders Ahlström. 2019. "Approaching the potential of model-data comparisons of global land carbon storage." *Scientific Reports* 9 (1):3367.
- Wutzler, T, and M Reichstein. 2008. "Colimitation of decomposition by substrate and decomposers—a comparison of model formulations." *Biogeosciences* 5 (3):749–759.
- Xia, J Y, Y Q Luo, Y-P Wang, ES Weng, and O Hararuk. 2012. "A semi-analytical solution to accelerate spin-up of a coupled carbon and nitrogen land model to steady state." *Geoscientific Model Development* 5 (5):1259–1271.
- Xia, Jianyang, Yiqi Luo, Ying-Ping Wang, and Oleksandra Hararuk. 2013. "Traceable components of terrestrial carbon storage capacity in biogeochemical models." *Global Change Biology* 19 (7):2104–2116.
- Xiao, Han, Kashif Rasul, and Roland Vollgraf. 2017. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms." arXiv preprint arXiv:1708.07747.
- Xu, Tao, Luther White, Dafeng Hui, and Yiqi Luo. 2006. "Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction." *Global Biogeochemical Cycles* 20 (2).
- Xu, Xia, Zheng Shi, Dejun Li, Ana Rey, Honghua Ruan, Joseph M Craine, Junyi Liang, Jizhong Zhou, and Yiqi Luo. 2016. "Soil properties control decomposition of soil organic carbon: Results from data-assimilation analysis." *Geoderma* 262:235–242.
- Yang, Yuanhe, Yiqi Luo, and Adrien C Finzi. 2011. "Carbon and nitrogen dynamics during forest stand development: A global synthesis." *New Phytologist* 190 (4):977–989.
- Zhang, Deqiang, Dafeng Hui, Yiqi Luo, and Guoyi Zhou. 2008. "Rates of litter decomposition in terrestrial ecosystems: Global patterns and controlling factors." *Journal of Plant Ecology* 1 (2):85–93.
- Zhang, Y, C Li, C C Trettin, H Li, and G Sun 2002. "An integrated model of soil, hydrology, and vegetation for carbon dynamics in wetland ecosystems." *Global Biogeochemical Cycles* 16(4): 9-1–9-17. <https://doi.org/10.1029/2001GB001838>
- Zhou, Jian, Jianyang Xia, Ning Wei, Yufu Liu, Chenyu Bian, Yuqi Bai, and Yiqi Luo. 2021. "A traceability analysis system for model evaluation on land carbon dynamics: Design and applications." *Ecological Processes* 10 (1):1–14.
- Zhou, Sha, Junyi Liang, Xingjie Lu, Qianyu Li, Lifeng Jiang, Yao Zhang, Christopher R Schwalm, Joshua B Fisher, Jerry Tjiputra, and Stephen Sitch. 2018. "Sources of uncertainty in modeled land carbon storage within and across three MIPs: Diagnosis with three new techniques." *Journal of Climate* 31 (7):2833–2851.
- Zhu, Q, J Liu, C Peng, H Chen, X Fang, H Jiang, ... X Zhou 2014. "Modelling methane emissions from

natural wetlands by development and application of the TRIPLEX-GHG model." *Geoscientific Model Development* 7(3): 981–999. <https://doi.org/10.5194/gmd-7-981-2014>

Zhuang, Qianlai, Jerry M Melillo, David W Kicklighter, Ronald G Prinn, A David McGuire, Paul A Steudler,

Benjamin S Felzer, and Shaomin Hu. 2004. "Methane fluxes between terrestrial ecosystems and the atmosphere at northern high latitudes during the past century: A retrospective analysis with a process-based biogeochemistry model." *Global Biogeochemical Cycles* 18 (3).

INDEX

(Page numbers in *italic* indicate figures, Page numbers in **bold** indicate tables)

- 1-3-5 scheme of diagnostics, 73, 74; *see also* unified diagnostic system for uncertainty analysis
- 3-pool model
 - data assimilation (soil incubation data), 192, 193, 195
 - nonautonomous ODE system solver and stability analysis, 107–108, 109, 110

A

Akaike information criterion (AIC), 260

B

- Bayesian statistics
 - CARDAMOM approach (data assimilation), 231
 - Ecological Platform for Assimilating Data (EcoPAD), 296
 - and Markov Chain Monte Carlo (MCMC), 181–187
 - quizzes, 187
 - suggested reading, 187
 - training course example, 182, 183–184
- Beijing Climate Centre (BCC), 76
- benchmarking, 20, 73, 157–162
 - aspects of land models to be evaluated, 158–159
 - CLM versions and future improvements, 160–162
 - Community Land Model version 4 (CLM4), 160–162, 161
 - Community Land Model version 4.5 (CLM4.5), 160–162, 161
 - Community Land Model version 5 (CLM5), 160–162, 161
 - International Land Model Benchmarking (ILAMB), 158, 159, 160, 161, 162
 - metrics, 160
 - quizzes, 162
 - reference datasets, 159–160, **159**
 - root-mean-square-error (RMSE), 160
 - schematic of framework, 158
 - suggested reading, 162
 - validation of model data, 157
 - variability in model simulations, 160

- Biome-BGC model, 117
- Brazil, CARDAMOM approach (data assimilation), 232–234, 233, 234
- butterfly effect, 115

C

- CABLE
 - semi-analytical spin-up (SASU) method, 115, 117, 118, 119–120, 119, 120, 129
 - traceability analysis, 141–143, 142, 163, 165
 - unified diagnostic system for uncertainty analysis, 75, 75, 78
- Canada (CAN) model, 76
- carbon balance equations, 25–27, 37
- carbon flow diagrams, 23–25, 24, 25
- carbon residence time
 - traceability analysis, 169, 169
 - transient traceability framework, 154, 154, 155, 156
- carbon sequestration, 45
- carbon vertical transfers, 9
- CarboTrain, 66–69
 - assimilation of carbon flux measurements, 279–280, 280
 - assimilation of carbon pool and flux measurements, 281–282, 282
 - assimilation of carbon pool measurements, 277–279, 278
- CMD window (Windows), 355, 356
- Community Land Model version 5 (CLM5), 331–332, 332
- data assimilation, 197–205
- docker, 353
- download instructions, 353
- Ecological Platform for Assimilating Data (EcoPAD), 303–306, 304, 305
- Edit source code, 68
- exercises, 204–206, 205, 206, 361–362, **361**, 362
- Fortran Compiler (MacOS), 359
- Fortran Compiler (Windows), 354–356, 355, 356, 357, 358

- GUI, 354, 361
- information contents of model and data, 273–283
- install Python 3.7.9 (MacOS), 358, 360
- install Python 3.7.9 (Windows), 353–354, 354, 355
- install R 3.6.3 (Windows), 356, 358, 359
- install R 4.0.5 (MacOS), 358–359, 361
- launch (MacOS), 361
- launch (Windows), 361
- MacOS installation, 357–361
- model intercomparison projects (MIPs), 163
- modifying default source code, 67–68
- New Advances in Land Carbon Cycle Modeling* (training course), 353
- Open solutions, 69
- real data assimilation, 283
- software (prerequisite), 353
- SPRUCE, 303–306, 305
- TECO model, 356–358, 361
- traceability analysis, 163
- user guide, 353–362
- uses of, 359–362
- Windows installation, 353–358
- without data assimilation, 274–277, 275
- CARDAMOM approach (data assimilation), 225–235
- analysis for Brazil, 232–234, 233, 234
- Bayesian statistics, 231
- carbon cycle data assimilation system (CCDAS), 235
- Copernicus LAI time series, 232, 234
- Data Assimilation Linked Ecosystem Carbon (DALEC), 229–230, 229, 234
- data-model integration, 227–228
- data sparsity, 227
- ecological and dynamic constraints (EDCs), 231
- eddy covariance (EC), 228
- example framework C cycle diagnostics, 228–234, 230
- innovations of approach, 232
- key challenges and opportunities, 234–235
- Markov Chain Monte Carlo (MCMC), 230, 231
- model complexity, 226–227
- model error, 227
- over-fitting, 227
- overview, 225–226, 228–229
- photosynthetic process studies, 227
- pixel, 232
- probabilistic approach, 234
- process rates, 226
- quizzes, 235
- schematic of framework, 230
- soil organic matter (SOM), 230
- Soil-Plant-Atmosphere (SPA) model, 229
- spatial resolutions, 235
- suggested reading, 235
- CellProfiler Analyst system, 311
- CENTURY, 27, 27
 - matrix models (developing), 65–66
 - matrix phosphorus model and data assimilation, 88
 - practice exercise, 31–32, 32
- CLM3.5 (Community Land Model version 3.5)
 - PRODA, 320
 - unified diagnostic system for uncertainty analysis, 75, 75, 78
- CLM4 (Community Land Model version 4)
 - benchmarking, 160–162, 161
 - global soil carbon models (data-constrained uncertainty analysis), 264
- CLM4.5 (Community Land Model version 4.5), 8, 29, 30
 - benchmarking, 160–162, 161
 - global soil carbon models (data-constrained uncertainty analysis), 266, 270, 270
 - matrix models (developing), 39
 - pool number, 38
 - soil layers, 53
 - transient traceability framework, 149
- CLM5 (Community Land Model version 5), 9
 - benchmarking, 160–162, 161
 - CarboTrain, 331–332, 332
 - coupled carbon-nitrogen matrix models, 47–54, 48, 49
 - deep learning parameterization, 329–336, 330
 - diagnostic variables in matrix models, 96–97
 - global validation for C and N simulations, 55–56, 55
 - Markov Chain Monte Carlo (MCMC), 329
 - and neural networks, 330–331
 - practice, 329–336, 330
 - PRODA, 320–321, 321, 323, **324**, 325, 327, 328, 329, 334–336
 - semi-analytical spin-up (SASU) method, 129
 - SOC optimized distribution exercise, 334–336, 335
 - soil layers, 53–54
 - unified diagnostic system for uncertainty analysis, 74–75
 - vertical distribution of SOC across U.S., 334–336, 335
- CMIP5 (Coupled Model Intercomparison Project Phase 5), 73
 - Harmonized World Soil Database (HWSD), 319
 - PRODA, 319
 - transient traceability framework, 147, 153
- CMIP6 (Coupled Model Intercomparison Project Phase 6), 139
 - traceability analysis, 146, 163, 167–169, 168, 169
- coarse woody debris (CWD), 54, 117
- Community Earth System Model Biogeochemical module (CESM GBC), 76, 269
- Community Land Model version 3.5 (CLM3.5)
 - PRODA, 320
 - unified diagnostic system for uncertainty analysis, 75, 75, 78
- Community Land Model version 4 (CLM4)
 - benchmarking, 160–162, 161
 - global soil carbon models (data-constrained uncertainty analysis), 264
- Community Land Model version 4.5 (CLM4.5), 8, 29, 30
 - benchmarking, 160–162, 161
 - global soil carbon models (data-constrained uncertainty analysis), 266, 270, 270
 - matrix models (developing), 39
 - pool number, 38
 - soil layers, 53
 - transient traceability framework, 149
- Community Land Model version 5 (CLM5), 9
 - benchmarking, 160–162, 161
 - CarboTrain, 331–332, 332
 - coupled carbon-nitrogen matrix models, 47–54, 48, 49
 - deep learning parameterization, 329–336, 330
 - diagnostic variables in matrix models, 96–97
 - global validation for C and N simulations, 55–56, 55
 - Markov Chain Monte Carlo (MCMC), 329
 - and neural networks, 330–331
 - practice, 329–336, 330
 - PRODA, 320–321, 321, 323, **324**, 325, 327–329, 334–336
 - semi-analytical spin-up (SASU) method, 129

- SOC optimized distribution exercise, 334–336, 335
 - soil layers, 53–54
 - unified diagnostic system for uncertainty analysis, 74–75
 - vertical distribution of SOC across U.S., 334–336, 335
 - compartmental dynamic system, 16, 57–64
 - autonomous systems (properties and long-term behavior), 60–62
 - autonomous vs. nonautonomous systems, 59–60
 - classification, 59–64, **59**
 - control theory concepts, 64
 - definition, 58–59
 - linear systems, 60–63
 - linear vs. nonlinear systems, 60
 - mass balance of single compartment, 58
 - matrix representation, 57
 - nonautonomous systems, 62–64
 - nonlinear systems, 61, 62, 63–64
 - quizzes, 64
 - stability analysis near equilibria, 61
 - suggested reading, 64; *see also* time characteristics of compartmental systems
 - conversion of mass law, 31
 - Copernicus LAI time series, 232, 234
 - coupled carbon-nitrogen matrix models, 45–56
 - application of matrix representation, 47
 - biological N fixation, 48
 - case study, 47
 - Community Land Model version 5 (CLM5), 47–54, 48, 49, 55–56
 - Duke Forest Free-Air CO₂ Enrichment (FACE), 47
 - global validation of CLM5, 55–56
 - litterfall, 48
 - matrix representation, 45–47
 - metrological datasets, 56
 - nitrogen transfers, 45
 - quizzes, 56
 - suggested reading, 56
 - Terrestrial Ecosystem (TECO) model, 45–47, 46
 - Coupled Model Intercomparison Project Phase 5 (CMIP5), 73
 - Harmonized World Soil Database (HWSD), 319
 - PRODA, 319
 - transient traceability framework, 147, 153
 - Coupled Model Intercomparison Project Phase 6 (CMIP6), 139
 - traceability analysis, 146, 163, 167–169, 168, 169
 - CWD (coarse woody debris), 54, 117
- D**
- DALEC (Data Assimilation Linked Ecosystem Carbon), 229–230, 229
 - data assimilation, 173–180
 - Bayesian statistics, 181–183, 183
 - CarboTrain, 197–205
 - choose a model, 216–218
 - convergence of MCMC results, 186–187
 - cost function, 177, 185, 186, 192–193, 200–201, 201, 213, 218
 - defining objective, 197, 216
 - deviance information criterion (DIC), 180
 - Ecological Platform for Assimilating Data (EcoPAD), 174, 295–296
 - ecosystem carbon sequestration, 175, 175
 - elements for realistic model, 174
 - estimate parameters, 203, 220–221
 - exercises, 204–206, 205, 206
 - flux data, 178
 - Free-Air CO₂ Enrichment Model Data Synthesis (FACE-MDS), 174–175
 - Gelman-Rubin (G-R) diagnostic method, 178, 186
 - Harmonized World Soil Database (HWSD), 175
 - Markov Chain Monte Carlo (MCMC), 177, 181, 183–187, 184, 218–220, 220
 - Metropolis-Hastings algorithm, 177, 183–185, 201–203
 - Michaelis-Menten model, 179–180
 - model choice, 198–200
 - need for, 174–176
 - optimization method, 201–203, 218–220
 - overview, 173–174
 - ‘perfect candidate’, 183–184
 - practice, 197–205
 - predict height from arm length, 176–177
 - prediction, 203–204, 204, 221
 - preparing data, 198, 216
 - priming, 179
 - prior knowledge, 183
 - process-based land carbon models, 174
 - quizzes, 180, 187
 - scientific values, 178–180, 179
 - seven step procedure, 176–178, 176, 197–205, **198**, 198, 216–222
 - suggested reading, 180, 187, 206
 - Terrestrial Ecosystem (TECO) model, 177, 178, 197, 198–200, 199
 - training course example, 182, 183–184
 - Data Assimilation Linked Ecosystem Carbon (DALEC), 229–230, 229, 234; *see also* CARDAMOM approach (data assimilation)
 - data assimilation (soil incubation data), 189–196
 - 3-pool model, 192, 193, 195
 - application of, 191–196, **192**, **193**, 193–196
 - cost function, 192–193
 - Markov Chain Monte Carlo (MCMC), 193
 - maximum likelihood estimates (MLEs), 194
 - quizzes, 196
 - soil carbon models, 190–191, 191
 - soil incubation experiments, 189–190
 - suggested reading, 196
 - data points, 79
 - data set types (parameters and predictions), 245–253
 - correlation between posterior parameter values, 252
 - entropy, 247, 248
 - information contents of model and data, 246, 248–251, 250
 - litterfall data, 249, 250
 - model equifinality, 251–252, 251
 - model error (inherent), 246
 - model predictions, 252
 - model uncertainty, 245
 - parameter uncertainty, 250
 - prediction of land carbon dynamics after data assimilation, 252–253
 - quizzes, 253
 - Shannon information index, 246–248, 253
 - short- and long-term information, 248, 249
 - soil organic matter (SOM), 246, 247
 - suggested reading, 253
 - Terrestrial Ecosystem (TECO) model, 246

- deviance information criterion (DIC), 180
- DGVMs (dynamic global vegetation models), 18
- diagnostic variables in matrix models
 - biosphere-atmosphere feedback, 95
 - carbon storage capacity and potential, 97–98
 - Community Land Model version 5 (CLM5), 96–97
 - exercises, 96–98
 - mathematical foundation, 95–97
 - practice, 95–99
 - suggested reading, 99
 - Terrestrial Ecosystem (TECO) model, 95, 96–97
 - uncertainty, 95
- DIC (deviance information criterion), 180
- donor pool, 4
- donor-pool-dominant transfer, 4–6
- DroughtNet, 293
- Duke Forest Free-Air CO₂ Enrichment (FACE), 21, 47, 76, 77
 - data assimilation, 174–175
 - transient traceability framework, 148–152
- dynamic global vegetation models (DGVMs), 18

E

- Earth system models (ESMs), 16
 - land surface models (LSMs), 17–18
 - machine learning and neural networks, 315–317, 316
 - PRODA, 319–320
- ecological forecasting, 287–292
 - data availability to constrain forecast, 290
 - disturbance events, 288, 290
 - Ecological Platform for Assimilating Data (EcoPAD), 287, 290–292, 291
 - eddy-flux networks, 290, 292
 - Long Term Ecological Research (LTER), 290, 292
 - models and predictability of terrestrial carbon cycle, 288–290, **289**
 - quizzes, 292
 - real-time sensors, 291
 - SPRUCE, 290–292
 - suggested reading, 292
 - weather forecasting, 287–288
 - workflow system to facilitate, 290–292
- Ecological Platform for Assimilating Data (EcoPAD), 174, 287, 290–292, 291, 293–300
 - accessing and working with, 302–303
 - Bayesian statistics, 296
 - CarboTrain, 303–306, 304, 305
 - Celery (job queue), 296, 297
 - custom workflow web portal, 303
 - data assimilation, 295–296
 - datasets, 294–295, 297, 301–302
 - docker (virtualization platform), 297–298
 - DroughtNet, 293
 - Eco-PAD-SPRUCE portal, 303
 - exercises, 303–305, 304, 305
 - file transfer protocol, 302
 - flux observations, 302
 - FLUXNET, 293, 302
 - forecasting uncertainty exercise, 305–306
 - general structure, 294–298
 - GitHub, 302
 - gross primary production (GPP), 298–300, 299
 - Markov Chain Monte Carlo (MCMC), 296, 298
 - model-experiment (ModEx) system, 294, 295
 - National Ecological Observatory Network (NEON), 293

- photosynthetically active radiation (PAR), 302
- posterior parameters exercise, 303–305, 305
- practice, 301–306
- probability density functions, 296
- quizzes, 300
- Representational State Transfer (RESTful) API, 297, 298
- scientific workflow, 296–297, 296
- SPRUCE, 298–300, 299, 301–306
- structured result storage, 298
- suggested reading, 300
- Terrestrial Ecosystem (TECO) model, 295, 302
- user-model interactions, 300
- web request-response, 294
- why do we need, 293–294
- Energy Exascale Earth System Model (E3SM), SPRUCE, 211–213
- ESMs (Earth system models), 16
 - land surface models (LSMs), 17–18
 - machine learning and neural networks, 315–317, 316
 - PRODA, 319–320
- Euler method, 26

F

- FACE (Duke Forest Free-Air CO₂ Enrichment), 21, 47, 76, 77
 - data assimilation, 174–175
 - transient traceability framework, 148–152
- Fick's Law, 216–218
- Fine Root Ecology Database (FRED), 212
- fire events, 9, 11
- flow diagrams and balance equations, 23–30
 - bank account example, 23, 25
 - carbon balance equation, 25–27
 - carbon balance equations, 28
 - carbon flow diagrams, 23–25, 24, 25
 - CENTURY, 27, 27, 31–32, 32
 - Community Land Model version 4.5 (CLM4.5), 29, 30
 - litter pools, 27
 - litterfall, 24, 26
 - one-pool carbon model, 25
 - ORCHIDEE-MICT, 28, 29, 30
 - practice, 31–34, 32, **33**
 - quizzes, 30
 - ReSOM model, 33–34, **33**
 - respiration, 24, 25
 - stock and flow, 23, 31
 - suggested reading, 30
 - Terrestrial Ecosystem (TECO) model, 27, 27
- two-pool carbon model, 26
- flux data
 - data assimilation, 178
 - Peatland methane study (data assimilation), 215
- FLUXNET, 293, 302
- forest fires, 3–4, 6, 11
- FRED (Fine Root Ecology Database), 212
- Free-Air CO₂ Enrichment Model Data Synthesis (FACE-MDS), 20, 47, 76, 77
 - data assimilation, 174–175
- FUN module, 53
- Fundamentals of Ecology (Odum and Odum), 16

G

- Gelman-Rubin (G-R) diagnostic method, 91, 186
 - data assimilation, 178

- PRODA, 322, 323
- general systems theory, 16
- global soil carbon models (data-constrained uncertainty analysis), 263–271
- alternative model structures, 264–267
 - Community Land Model version 4 (CLM4), 264
 - Community Land Model version 4.5 (CLM4.5), 266, 270, 270
 - datasets and data-model fusion, 267–268
 - Harmonized World Soil Database (HWSD), 264, 267
 - Markov Chain Monte Carlo (MCMC), 263, 267
 - microbial growth efficiency (MGE), 267
 - Microbial-Mineral Carbon Stabilization (MIMICS), 266–268, 270, 270
 - Northern Circumpolar Soil Carbon Database (NCSCD), 264, 267
 - permafrost soils, 267
 - posterior distribution of model parameters, 268–269, 268
 - quizzes, 271
 - representative concentration pathway 8.5 (RCP8.5), 269–270, 269
 - sensitivity to initial conditions and model parameters, 270–271
 - soil C decomposition models, 264, 264
 - suggested reading, 271
- GPP (gross primary production), 75–76
- Ecological Platform for Assimilating Data (EcoPAD), 298–300, 299
- Great Lakes, PRODA, 323
- Great Plains, PRODA, 324
- gross primary production (GPP), 75–76
- Ecological Platform for Assimilating Data (EcoPAD), 298–300, 299
- ## H
- Harmonized World Soil Database (HWSD), 175, 263–264, 267
- CMIP(5), 319
 - PRODA, 323
- Harvard Forest, 77
- transient traceability framework, 148–152
- Henry's Law, 218
- Houtouwan, 3, 4, 11
- HWSD (Harmonized World Soil Database), 175, 264, 267
- CMIP(5), 319
 - PRODA, 323
- ## I
- ILAMB (International Land Model Benchmarking), 158–160
- benchmarking, 160, 162
- image classification, 311, 314–315, 315
- industrial revolution, 115–116
- information contents of model and data, 246, 248–251, 250, 271
- assimilation of carbon flux measurements, 279–280, 280
 - assimilation of carbon pool and flux measurements, 281–282, 282
 - assimilation of carbon pool measurements, 277–279, 278
 - carbon flux predictions, 283
 - CarboTrain, 273–283
 - model uncertainty, 273
 - posterior distribution of model parameters, 276
 - practice, 273–283
 - real data assimilation, 283
 - without data assimilation, 274–277, 275; *see also* data set types (parameters and predictions)
- input-to-state stability (ISS), 64
- International Land Model Benchmarking (ILAMB), 158–160
- benchmarking, 160, 162
- IPCC climate projections, 18, 19, 23, 139
- ISS (input-to-state stability), 64
- ## J
- Jacobian matrix, 62
- ## L
- LabelMe, 312
- LAI (Leaf Area Index), 218, 228
- lake carbon dynamics (controlling mechanisms), 255–262
- Akaike information criterion (AIC), 260
 - epilimnetic C dynamics, 256–262, 257, 261
 - hypothesis (gravity-driven density currents), 258
 - hypothesis (hypolimnetic water), 258–259
 - hypothesis (photooxidation of DOC), 258
 - hypothesis (varying DOC lability), 257, 260
 - Long lake, 256
 - Markov Chain Monte Carlo (MCMC), 255
 - Metropolis-Hastings algorithm, 259, 260
 - model-based hypothesis testing, 256
 - model calibration and selection, 259–260
 - models and data-model fusion, 255–256, 261
 - overfitting, 256, 259
 - quizzes, 262
 - suggested reading, 262
 - thermocline, 256
- land surface models (LSMs), Earth system models (ESMs), 17–18
- landscapes, vegetation take over, 3, 6, 11
- Leaf Area Index (LAI), 218, 228
- leaf pools
- carbon balance, 6
 - carbon from photosynthesis, 6, 7
- litter decay constant, 5
- litter decomposition, 5
- litter pools, 4
- flow diagrams and balance equations, 27
 - matrix equation, 39
 - number, 38
- litterfall
- coupled carbon-nitrogen matrix models, 48
 - data, 249, 250
 - flow diagrams and balance equations, 24, 26
 - rate, 4
 - recipient pool, 4, 5
- Long Term Ecological Research (LTER), 290, 292
- LUNA module, 53
- ## M
- machine learning and neural networks, 309–317
- activation function, 330
 - baseline accuracy, 311

cell image classification, 311
 CellProfiler Analyst system, 311
 correlation between outputs, 316–317
 cross-validation of parameters of ESMs, 315–317, 316
 email spam filtering, 311
 epoch number, 331
 estimate of variance, 311
 generalization goal, 310
 hyper-parameters, 312, 313, 331–332
 image classification, 311, 314–315, 315
 K-fold cross-validation, 310–311, 310, 313, 315–317
 loss function, 330
 MNIST/Fashion-MNIST datasets, 309, 310, 315
 optimizers, 331
 under/overfitting, 312–314, 313, 333
 overview, 309–311, 309
 practice, 329–336
 PRODA vs. data assimilation alone, 334–336
 quizzes, 317
 retinal photographs, 312
 suggested reading, 317
 train dataset, 310, 313, 315
 training target, 330
 tuning for better performance, 333, 333
 validation error (U shape), 314

Markov Chain Monte Carlo (MCMC), 124
 and Bayesian statistics, 181–187
 CARDAMOM approach (data assimilation), 230, 231
 Community Land Model version 5 (CLM5), 329
 data assimilation, 177, 183–187, 218–220, 220
 data assimilation (soil incubation data), 193
 Ecological Platform for Assimilating Data (EcoPAD), 296, 298
 global soil carbon models (data-constrained uncertainty analysis), 263, 267
 lake carbon dynamics (controlling mechanisms), 255
 Metropolis-Hastings algorithm, 183–185, 184
 PRODA, 320, 322, 323
 SPRUCE, 213

matrix algebra, 337–342
 eigenvalues and eigenvectors, 341–342
 linear system, 341
 matrix equations, 340–341
 matrix multiplication, 338–339
 matrix operations, 338
 motivations, 337–338
 quizzes, 339, 341, 342
 Strang's Linear Algebra lecture (MIT), 337
 suggested reading, 342

matrix approach (model representation), 6–10
 pool-and-flux, 6, 38

matrix models (developing), 37–43
 CENTURY, 65–66
 coding and running, 66–69, 67–68, 69
 Community Land Model version 4.5 (CLM4.5), 39
 deriving the matrix equation, 39–42, 40
 litter pools, 39
 matrix version of carbon balance equation, 37–39
 ORCHIDEE-MICT, 39–42, 39, 41–42
 phosphorus, 89–91, 90
 practice, 65–69
 Python, 66–69
 quizzes, 43
 suggested reading, 42

Terrestrial Ecosystem (TECO) model, 38, 39
 matrix phosphorus model and data assimilation, 87–94
 balance equations, 90–91
 CENTURY, 88
 construction of model, 89–90, 90
 data assimilation example, 89–91, 90
 data selection and description, 89
 matrix approach, 88–89, 93
 model validation and data assimilation, 91
 phosphorus, 87–88
 quizzes, 94
 soil dynamics models, 88
 soil P and other ecosystem properties, 93, 94
 soil P dynamics quantified, 91–93, 92–93, 92
 spin-up, 88, 89
 suggested reading, 94
 supercomputer use, 80, 93

MCMC (Markov Chain Monte Carlo), 124
 and Bayesian statistics, 181–187
 CARDAMOM approach (data assimilation), 230, 231
 Community Land Model version 5 (CLM5), 329
 data assimilation, 177, 183–187, 218–220, 220
 data assimilation (soil incubation data), 193
 Ecological Platform for Assimilating Data (EcoPAD), 296, 298
 global soil carbon models (data-constrained uncertainty analysis), 263, 267
 lake carbon dynamics (controlling mechanisms), 255
 Metropolis-Hastings algorithm, 183–185, 184
 PRODA, 320, 322, 323
 SPRUCE, 213

methane, 215; *see also* Peatland methane study (data assimilation)

metrological datasets, 56

Metropolis-Hastings algorithm, 91
 data assimilation, 177, 183–185, 184, 201–203
 lake carbon dynamics (controlling mechanisms), 259, 260
 Markov Chain Monte Carlo (MCMC), 183–185, 184

MGE (microbial growth efficiency), 267

Michaelis-Menten model, 179–180, 266

microbial growth efficiency (MGE), 267

Microbial-Mineral Carbon Stabilization (MIMICS), 266–268, 270, 270

MIPs (model intercomparison projects), 73, 139
 CarboTrain, 163
 transient traceability framework, 147, 148, 152–156

model equifinality, data set types (parameters and predictions), 251–252, 251

model error, inherent, 20, 246

model intercomparison projects (MIPs), 73, 139
 CarboTrain, 163
 transient traceability framework, 147, 148, 152–156

model predictions, 252

model simulations, 79

modeling (introduction), 13–21
 common everyday models, 14
 modeling workflow, 18–21
 models in research, 14–15, 15
 modes of application, 15
 predicting forward in time, 15
 quizzes, 21
 suggested reading, 21
 system dynamics, 16

- types of land carbon cycle models, 16–18, **18**
- ways of using models, 15–16
 - what is a model, 13
- modeling workflow, 18–21
 - calibrate the model, 20
 - choose a model, 19
 - data assimilation, 197–205
 - design the model experiment, 21
 - initialization/spin-up, 21
 - seven step procedure, 176–178, 176, 197–205, **198**, 198, 216–222
 - software application, 19
 - specify the question/hypothesis, 18
 - validate the model, 20–21
 - verify the model works, 19–20
- Monte Carlo sampling, 80
- Multiscale Synthesis and Terrestrial Model Intercomparison Project (MstTMP), 116

N

- National Ecological Observatory Network (NEON), 293
- NBP (net biome production), 20
- NCSCD (Northern Circumpolar Soil Carbon Database), 264, 267
- NEE (net ecosystem exchange), 232
- NEON (National Ecological Observatory Network), 293
- net biome production (NBP), 20
- net ecosystem exchange (NEE), 232
- net primary productivity (NPP), 73, 75–76, 77, 96, 119
 - PRODA, 322
 - traceability analysis, 142, 142, 165
 - transient traceability framework, 148, 151, 151, 152, 153, 153, 154, 155, 156
- NimBioS workshop (2012), 10, 12
- nitrogen cycle, 17
- nitrogen transfers
 - biological N fixation, 48
 - coupled carbon-nitrogen matrix models, 45
- nonautonomous ODE system solver and stability analysis, 103–113
 - 3-pool model, 107–108, 109, 110
 - analytical solution, 104–112
 - case study, 111–112
 - first-order non-homogeneous scalar equation, 104
 - global attractor, 109–111
 - homogeneous nonautonomous ODEs system, 105
 - n-pool model, 106–107
 - non-homogeneous nonautonomous ODEs system, 105–106
 - one-pool carbon model, 104–105
 - quizzes, 112–113
 - stability, 108–111
 - suggested reading, 112
- nonautonomous systems, 10, 77
- nonlinear microbial models, 10
- Northern Circumpolar Soil Carbon Database (NCSCD), 264, 267
- NPP (net primary productivity), 73, 75–76, 77, 96, 119
 - PRODA, 322
 - traceability analysis, 142, 142, 165
 - transient traceability framework, 148, 151, 151, 152, 153, 153, 154–156
- numerical simulation models, 14
- numerical weather prediction (NWP), 14–15, 288

O

- observational error, 227
- ODEs (ordinary differential equations), 103; *see also*
 - nonautonomous ODE system solver and stability analysis
- one-pool carbon model, 25
 - nonautonomous ODE system solver and stability analysis, 104–105
- ORCHIDEE, 28
 - semi-analytical spin-up (SASU) method, 129
- ORCHIDEE-CNP, 25
- ORCHIDEE-MICT, 28, 29, 30
 - matrix models (developing), 39–42, 39, 41–42
 - pool number, 38
 - sensitivity analysis with matrix equations, 80, **81–82**
 - vertical soil layers, 42
- ordinary differential equations (ODEs), 103; *see also*
 - nonautonomous ODE system solver and stability analysis
- overfitting, 227
 - lake carbon dynamics (controlling mechanisms), 256, 259
 - machine learning and neural networks, 312–314, 313, 333

P

- PAR (photosynthetically active radiation), 302
- Peatland methane study (data assimilation), 215–223
 - flux data, 215
 - Markov Chain Monte Carlo (MCMC), 218–220, 220
 - quizzes, 223
 - seven step procedure, 216–222
 - soil temperature, 221
 - suggested reading, 223
 - Terrestrial Ecosystem (TECO) model, 216–222, 217, **219**, 220–222
 - uncertainty in methane modeling, 215–216
- PFT (plant functional types), 232
- phosphorus
 - genetic information carriers, 87
 - matrix model and data assimilation, 87–88
- photosynthetic products, 4
- photosynthetically active radiation (PAR), 302
- plant functional types (PFT), 232
- practice
 - Community Land Model version 5 (CLM5), 329–336, 330
 - data assimilation, 197–205
 - diagnostic variables in matrix models, 95–99
 - Ecological Platform for Assimilating Data (EcoPAD), 301–306
 - flow diagrams and balance equations, 31–34, 32, **33**
 - information contents of model and data, 273–283
 - machine learning and neural networks, 329–336
 - matrix models (developing), 65–69
 - semi-analytical spin-up (SASU) method, 129–135
 - SPRUCE, 237–241, **237**
 - traceability analysis, 163–170
- process-based land carbon models, 80
- PROcess-guided deep learning and DAta-driven modeling (PRODA), 319–328
 - big data, 320

- Community Land Model version 3.5 (CLM3.5), 320
- Community Land Model version 5 (CLM5), 320–321, 321, 323, **324**, 325, 327–329, 334–336
- Coupled Model Intercomparison Project Phase 5 (CMIP5), 319
- deep learning model, 322–323
- Earth system models (ESMs), 319–320
- Gelman–Rubin (G-R) diagnostic method, 322, 323
- Great Lakes, 323
- Great Plains, 324
- Harmonized World Soil Database (HWSD), 323
- Markov Chain Monte Carlo (MCMC), 320, 322, 323
- model representation of SOC across U.S. sites, 323
- net primary productivity (NPP), 322
- process-based model, 320–322
- quizzes, 328
- reference SOC data products, 323
- SOC distribution (realistic representations), 327–328, 327
- soil carbon and site-level data assimilation, 322
- soil layer numbers, 321
- soil layer thickness, 321
- spatial distribution of SOC across U.S., 323–324, **324**, 324
- suggested reading, 328
- vertical distribution of SOC across U.S., 325–327, 325, 326, 334–336, 335
- workflow of PRODA, 320–323

Python

- advanced variables and operators, 346–350
- Boolean values, 344
- CarboTrain, 66–69, 67–68, 69, 197–205
- class operator, 348–349, 349
- code block, 344, 344, 348
- data assimilation, 197–205
- first program, 343–344
- function operator, 347–348, 347
- introduction to programming, 343–351
- keras package, 331
- list variable, 346–347, 347
- module operator, 349–350, 350
- namespaces, 349
- quizzes, 351
- shell window, 343
- suggested reading, 350–351
- syntax, 343, 347, 348
- variables and operators, 344–346
- workflow of function calls, 348, 348

Q

quizzes

- benchmarking, 162
- compartmental dynamic system, 64
- coupled carbon-nitrogen matrix models, 56
- data assimilation, 180
- data assimilation (soil incubation data), 196
- data set types (parameters and predictions), 253
- ecological forecasting, 292
- Ecological Platform for Assimilating Data (EcoPAD), 300
- flow diagrams and balance equations, 30
- global soil carbon models (data-constrained uncertainty analysis), 271
- lake carbon dynamics (controlling mechanisms), 262
- machine learning and neural networks, 317

- matrix algebra, 339, 341, 342
- matrix models (developing), 43
- matrix phosphorus model and data assimilation, 94
- modeling (introduction), 21
- nonautonomous ODE system solver and stability analysis, 112–113
- Python, 351
- semi-analytical spin-up (SASU) method, 121
- sensitivity analysis with matrix equations, 85
- SPRUCE, 214
- theoretical foundations, 12
- time characteristics of compartmental systems, 127
- traceability analysis, 146
- transient traceability framework, 156
- unified diagnostic system for uncertainty analysis, 78

S

- SASU *see* semi-analytical spin-up (SASU) method
- scientific hypotheses, 14
- secondary succession, 4, 6
- semi-analytical spin-up (SASU) method, 115–121
 - accelerated decomposition (AD), 116–117
 - accelerated spin-up (ASU), 117
 - CABLE, 115, 117, 118, 119–120, 119, 120, 129
 - Community Land Model version 5 (CLM5), 129
 - computational efficiency, 120–121, 129
 - exercises, 130–135
 - industrial revolution, 115–116
 - mathematical foundation, 117–119
 - native dynamics spin-up (ND), 116, 118, 129, 131
 - nonlinear systems, 132–135
 - ORCHIDEE, 129
 - passive soil turnover rate/soil carbon, 134
 - practice, 129–135
 - quizzes, 121
 - soil dynamics models, 89
 - spin-up, 115–117
 - steady state, 116
 - suggested reading, 121
 - supercomputer use, 116, 120–121
 - Terrestrial Ecosystem (TECO) model, 129–135, **130**, **133**
- sensitivity analysis with matrix equations, 79–85
 - one-at-a-time sensitivity analysis, 82–85, 83, 84
 - ORCHIDEE-MICT, 80, **81–82**
 - quizzes, 85
 - sensitivity analysis, 79–80
 - Sobol sensitivity analysis, 80–82, 83
 - spatial pattern, 85
 - suggested reading, 85
 - supercomputer use, 80
 - variance-based approach, 80
- Shannon information index, 246–248, 253
- SiB2 (Simple Biosphere Model), 17
- Silver Springs (Florida), 16–17, 17
- Simple Biosphere Model (SiB2), 17
- Sobol sensitivity analysis, 80–82, 83
- SOC (soil organic carbon), 5
 - distribution across U.S. PRODA model, 323–324, **324**, 324, 325–327, 325, 326, 334–336, 335
 - three-pool models, 6
- software application, modeling workflow, 19
- soil carbon cycling, 4
- soil dynamics models

- matrix phosphorus model and data assimilation, 88
 - semi-analytical spin-up (SASU) method, 89
 - spin-up, 88, 89
 - soil incubation, 5
 - soil incubation experiments, data assimilation (soil incubation data), 189–190, 190
 - soil layers, coupled carbon-nitrogen matrix models, 53–54
 - soil organic carbon (SOC), 5
 - distribution across U.S. PRODA model, 323–324, **324**, 324, 325–327, 325, 326, 334–336, 335
 - three-pool models, 6
 - soil organic matter (SOM), 4
 - CARDAMOM approach (data assimilation), 230
 - data set types (parameters and predictions), 246, 247
 - modeling workflow, 19
 - single-pool, 19
 - soil phosphorus, terrestrial carbon dynamics, 87
 - Soil-Plant-Atmosphere (SPA) model, 229
 - soil respiration, 24, 25
 - Soil-Vegetation-Atmosphere Transfer (SVAT) schemes, 17
 - SOM (soil organic matter), 4
 - CARDAMOM approach (data assimilation), 230
 - data set types (parameters and predictions), 246, 247
 - modeling workflow, 19
 - single-pool, 19
 - SPA (Soil-Plant-Atmosphere) model, 229
 - SPRUCE (Spruce and Peatland Response Under Changing Environments)
 - CarboTrain, 237–241
 - cost function, 213
 - data assimilation with TECO, 239–240
 - ecological forecasting, 290, 291, 292
 - Ecological Platform for Assimilating Data (EcoPAD), 298–300, 299, 301–306
 - ELM-SPRUCE, 211–213
 - enclosure, 210
 - enclosure design, 210–211
 - Energy Exascale Earth System Model (E3SM), 211–213
 - hummocks, 210
 - key science questions, 211
 - Markov Chain Monte Carlo (MCMC), 213
 - model data integration, 209–214
 - model validation and uncertainty quantification, 212–214
 - modeling for the experiment, 211–212
 - overview, 209–210
 - parameter sensitivity of model output, 238–239
 - practice, 237–241, **237**
 - quizzes, 214
 - site description, 210–211, 214
 - Sphagnum productivity, 213–214
 - suggested reading, 214
 - Terrestrial Ecosystem (TECO) model, 212, 239–240, 240
 - United States Department of Energy (DOE), 211
 - suggested reading
 - benchmarking, 162
 - compartmental dynamic system, 64
 - coupled carbon-nitrogen matrix models, 56
 - data assimilation, 180, 206
 - data assimilation (soil incubation data), 196
 - data set types (parameters and predictions), 253
 - diagnostic variables in matrix models, 99
 - ecological forecasting, 292
 - Ecological Platform for Assimilating Data (EcoPAD), 300
 - flow diagrams and balance equations, 30
 - global soil carbon models (data-constrained uncertainty analysis), 271
 - lake carbon dynamics (controlling mechanisms), 262
 - machine learning and neural networks, 317
 - matrix models (developing), 42
 - matrix phosphorus model and data assimilation, 94
 - modeling (introduction), 21
 - nonautonomous ODE system solver and stability analysis, 112
 - Python, 350–351
 - semi-analytical spin-up (SASU) method, 121
 - sensitivity analysis with matrix equations, 85
 - SPRUCE, 214
 - theoretical foundations, 12
 - time characteristics of compartmental systems, 127
 - traceability analysis, 146
 - transient traceability framework, 156
 - unified diagnostic system for uncertainty analysis, 78
 - supercomputer use, 80, 93
 - semi-analytical spin-up (SASU) method, 116, 120–121, 129
 - SVAT (Soil-Vegetation-Atmosphere Transfer) schemes, 17
 - systems theory, 16
- T**
- TECO (Terrestrial Ecosystem) model *see* Terrestrial Ecosystem (TECO) model
 - terrestrial biosphere models (TBMs), 18
 - terrestrial carbon cycle matrix model, 7
 - Terrestrial Ecosystem (TECO) model, 7–9, 17, 27, 27
 - accelerated spin-up (ASU), 117
 - coupled carbon-nitrogen matrix models, 45–47, 46
 - data assimilation, 177, 178, 197, 198–200, 199
 - data set types (parameters and predictions), 246
 - diagnostic variables in matrix models, 95, 96–97
 - Ecological Platform for Assimilating Data (EcoPAD), 295, 302
 - matrix models (developing), 38, 39
 - non-zero elements, 47
 - Peatland methane study (data assimilation), 216–222, 217, 219, 220–222
 - semi-analytical spin-up (SASU) method, 129–135, **130**, **133**
 - SPRUCE, 212, 239–240, 240
 - TECO-CH₄, 216, 217, 218, **219**, 220, 221
 - traceability analysis, 144–145, 145
 - transient traceability framework, 149
 - zero elements, 47
 - theoretical foundations, 3–12
 - donor-pool-dominant transfer, 4–6
 - dynamic disequilibrium of land carbon cycle, 11–12
 - fundamental properties, 6
 - matrix approach (model representation), 6–10
 - paradox of matrix equation and nonautonomous systems, 10–11
 - predictability and carbon cycle, 11
 - predictability land carbon cycle, 11
 - quizzes, 12
 - suggested reading, 12
 - three-pool models, soil organic carbon (SOC), 6
 - time characteristics of compartmental systems, 123–127
 - age and transit time, 123, 124–126, 124
 - age distributions, 126
 - nonlinear systems, 125

- phase-type distribution, 125
 - pool age, 123, 124
 - quizzes, 127
 - stochastic representation, 126
 - suggested reading, 127
 - transit time distribution, 126
 - traceability analysis, 78
 - authentic traceability analysis, 164–170, 165
 - CABLE, 141–143, 142, 163, 165
 - carbon cycle variations (understanding), 141–142
 - carbon residence time, 169, 169
 - CarboTrain, 163, 164–170
 - case study, 141–146
 - Coupled Model Intercomparison Project Phase 6 (CMIP6), 146, 163, 167–169, 168, 169
 - exercises, 164–170
 - external forcing (evaluating), 143–144, 143
 - framework design and key components, 140–141, 141
 - identification of uncertainty sources, 139, 141–146
 - intermodel comparisons, 142–143
 - LPJ-GUESS model, 143–144
 - model intercomparison projects (MIPs), 139
 - net primary productivity (NPP), 142, 142, 165
 - new processes in models, 144–145
 - overview, 139–146
 - pace of model evaluation, 145–146
 - post-MIP traceability analysis, 167–170, 167
 - practice, 163–170
 - quizzes, 146
 - suggested reading, 146
 - terrestrial carbon storage (spatial distribution), 166
 - Terrestrial Ecosystem (TECO) model, 144–145, 145
 - TraceME, 145–146
 - TraceME
 - traceability analysis, 145–146
 - transient traceability framework, 156
 - transient dynamical equation, 11, 12
 - transient traceability framework, 147–156
 - carbon residence time, 154, 154, 155, 156
 - carbon storage variance decomposition, 155
 - Community Land Model version 4.5 (CLM4.5), 149
 - Coupled Model Intercomparison Project Phase 5 (CMIP5), 147, 153
 - Duke Forest Free-Air CO₂ Enrichment (FACE), 148–152
 - Harvard Forest, 148–152
 - land carbon storage dynamics, 148
 - model intercomparison projects (MIPs), 147, 148, 152–156
 - net primary productivity (NPP), 148, 151, 151, 152, 153, 153, 154–156
 - quizzes, 156
 - suggested reading, 156
 - Terrestrial Ecosystem (TECO) model, 149
 - TraceME, 156
 - Trends in Net Land-Atmosphere Carbon Exchange (TRENDY), 148, 153
 - Trends in Net Land-Atmosphere Carbon Exchange (TRENDY), 148
 - transient traceability framework, 153
 - two-pool carbon model, 26
- ## U
- underfitting, machine learning and neural networks, 312–314, 313
 - unified diagnostic system for uncertainty analysis, 73–78
 - 1-3-5 scheme of diagnostics, 73, 74
 - benchmarking, 73
 - CABLE, 75, 75, 78
 - Community Land Model version 3.5 (CLM3.5), 75, 75, 78
 - Community Land Model version 5 (CLM5), 74–75
 - diagnostic variables in matrix models, 95
 - one formula, 74–75
 - quizzes, 78
 - suggested reading, 78
 - three-dimensional space to describe model outputs, 75–78, 76, 77
 - traceability analysis, 78
 - uncertainty in land carbon cycle modeling, 73–74
 - United States
 - Silver Springs (Florida), 16–17, 17
 - SOC distribution, 323–324, **324**, 324, 325–327, 325, 326, 334–336, 335
- ## V
- vegetation carbon dynamics, 9
- ## W
- weather forecasting, 287–288
 - numerical weather prediction (NWP), 288
 - wood pool, 8
- ## Y
- Yellowstone fires, 3–4, 11