https://ntrs.nasa.gov/search.jsp?R=20170005725 2020-05-09T22:45:51+00:00Z

DOE/SC-0186 | doi:10.2172/1330803

U.S. DEPARTMENT OF ENERGY | Office of Science

2016 International Land Model Benchmarking (ILAMB) Workshop Report

Recommended Citation

Hoffman, F. M., C. D. Koven, G. Keppel-Aleks, D. M. Lawrence, W. J. Riley, J. T. Randerson, A. Ahlström, G. Abramowitz, D. D. Baldocchi, M. J. Best, B. Bond-Lamberty, M. G. De Kauwe, A. S. Denning, A. Desai, V. Eyring, J. B. Fisher, R. A. Fisher, P. J. Gleckler, M. Huang, G. Hugelius, A. K. Jain, N. Y. Kiang, H. Kim, R. D. Koster, S. V. Kumar, H. Li, Y. Luo, J. Mao, N. G. McDowell, U. Mishra, P. R. Moorcroft, G. S. H. Pau, D. M. Ricciuto, K. Schaefer, C. R. Schwalm, S. P. Serbin, E. Shevliakova, A. G. Slater, J. Tang, M. Williams, J. Xia, C. Xu, R. Joseph, and D. Koch (2017), *International Land Model Benchmarking (ILAMB) 2016 Workshop Report*, DOE/SC-0186, U.S. Department of Energy, Office of Science, Germantown, Maryland, USA, doi:10.2172/1330803.

2016 International Land Model Benchmarking (ILAMB) Workshop Report

Report of an international workshop
held in Washington, DC, USA, May 16–18, 2016

Supported by

U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research

Organizers

Renu Joseph
Regional and Global Climate Modeling

Dorothy Koch
Earth System Modeling

Workshop Co-Chairs

Forrest M. Hoffman
Oak Ridge National Laboratory

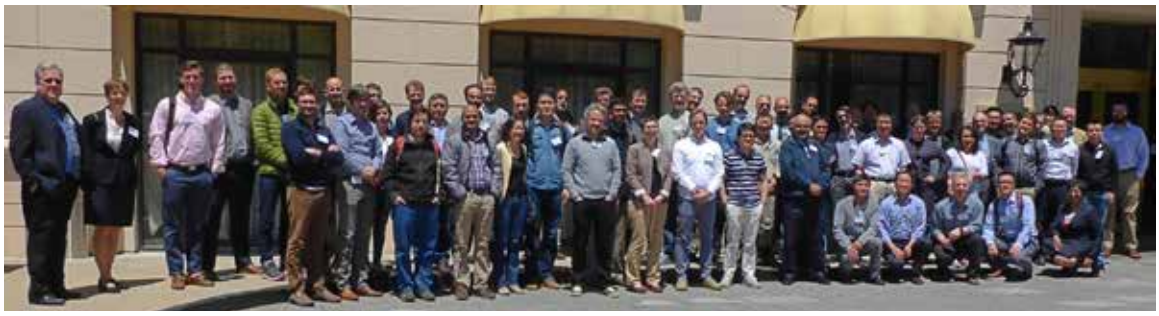
William J. Riley
Lawrence Berkeley National Laboratory

James T. Randerson
University of California at Irvine

Gretchen Keppel-Aleks
University of Michigan at Ann Arbor

David M. Lawrence
National Center for Atmospheric Research

U.S. DEPARTMENT OF ENERGY | Office of Science



2016 International Land Model Benchmarking (ILAMB) Workshop, May 16–18, 2016, Washington, DC, USA

Lead Authors

Forrest M. Hoffman
Charles D. Koven
Gretchen Keppel-Aleks
David M. Lawrence
William J. Riley
James T. Randerson

Contributing Authors

Anders Ahlström
Gabriel Abramowitz
Dennis D. Baldocchi
Martin J. Best
Ben Bond-Lamberty
Martin G. De Kauwe
A. Scott Denning
Ankur Desai
Veronika Eyring
Joshua B. Fisher
Rosie A. Fisher
Peter J. Gleckler
Maoyi Huang
Gustaf Hugelius
Atul K. Jain
Nancy Y. Kiang
Hyungjun Kim
Randal D. Koster
Sujay V. Kumar
Hongyi Li
Yiqi Luo
Jiafu Mao
Nathan G. McDowell
Umakant Mishra
Paul R. Moorcroft
George S. H. Pau
Daniel M. Ricciuto
Kevin Schaefer
Christopher R. Schwalm
Shawn P. Serbin
Elena Shevliakova
Andrew G. Slater
Jinyun Tang
Mathew Williams
Jianyang Xia
Chonggang Xu

Executive Summary

As earth system models (ESMs) become increasingly complex, there is a growing need for comprehensive and multi-faceted evaluation of model projections. To advance understanding of terrestrial biogeochemical processes and their interactions with hydrology and climate under conditions of increasing atmospheric carbon dioxide, new analysis methods are required that use observations to constrain model predictions, inform model development, and identify needed measurements and field experiments. Better representations of biogeochemistry–climate feedbacks and ecosystem processes in these models are essential for reducing the acknowledged substantial uncertainties in 21st century climate change projections.

Building upon past model evaluation studies, the goals of the International Land Model Benchmarking (ILAMB) project are to:

1. Develop internationally accepted benchmarks for land model performance by drawing upon international expertise and collaboration
2. Promote the use of these benchmarks by the international community for model intercomparison
3. Strengthen linkages among experimental, remote sensing, and climate modeling communities in the design of new model tests and new measurement programs
4. Support the design and development of open source benchmarking tools.

The second ILAMB Workshop in the United States was convened on May 16 to 18, 2016, in Washington, District of Columbia, USA. Sponsored by the U.S. Department of Energy's (DOE's) Office of Biological and Environmental Research, the workshop was convened by the Biogeochemistry–Climate Feedbacks Scientific Focus Area (BGC Feedbacks SFA) project. Overarching goals of the workshop were to engage the international research community in defining scientific priorities for (1) design of new metrics, (2) improvement of model development and workflow practices, (3) Coupled Model Intercomparison Project (CMIP) evaluation, and to learn about new observational data sets and measurement campaigns.

The workshop drew more than 60 on-site participants, and between 20 and 30 individuals—including students and postdocs—attended online at any time during the plenary sessions. Participants were from Australia, Canada, China, Germany, Japan, Netherlands, Sweden, United Kingdom, and the United States and represented 10 different major modeling centers. Plenary presentations focused on model benchmarking, emergent constraints, evaluation metrics, uncertainty quantification, and field experiment and measurement networks.

Outcomes of the 2016 ILAMB Workshop

This 2016 ILAMB Workshop Report provides a synopsis of the current state of the science and highlights challenges and opportunities for benchmarking, model development, and field and laboratory measurements needed to advance climate science. The main text of the report provides a synthesis of the ideas, concepts, and scientific priorities presented and discussed at the workshop. The appendix of the report consists of topical white papers that summarize invited presentations, describe breakout group proceedings, and offer recommendations. In addition, the white papers identify critical gaps and opportunities in measurement programs, offer new approaches for model evaluation, and point out synergies among research teams and tools being constructed to support model development, parameter estimation, and model–data integration.

As depicted in the schematic figure below, the topical white papers within the categories of Major Processes and Integrating and Cross-cutting Themes were synthesized with those on the needs of Model Intercomparison Projects (MIPs) to produce a set of next generation Benchmarking Challenges and Priorities resulting from the workshop. Moreover, Benchmarking Approaches for addressing these challenges were identified and Enabling Capabilities needed to facilitate next generation benchmarking and model development were distilled from the white papers. Addressing these challenges will advance climate science by enabling process understanding, quantifying feedbacks, reducing uncertainties, and improving model projections.

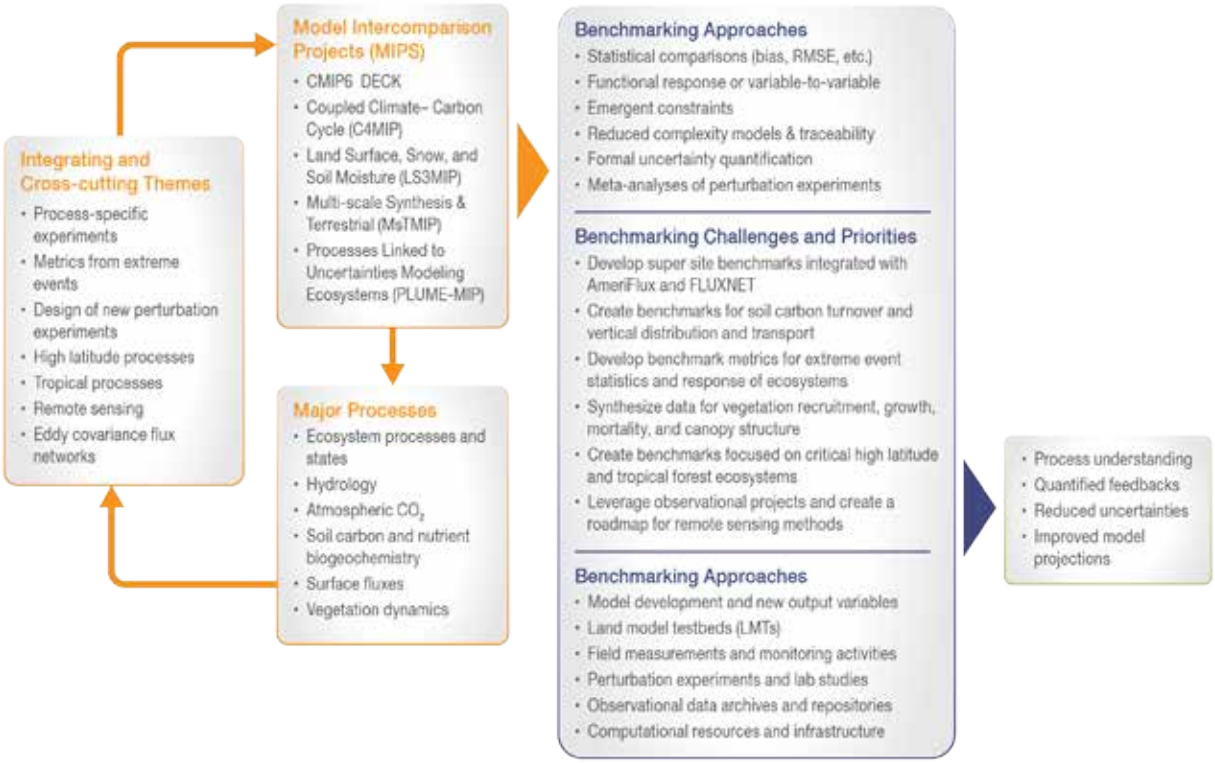
Benchmarking Tools

Model evaluation and benchmarking tools currently employed by international modeling centers were assessed at the workshop. Features of current benchmarking tools—including the Protocol for the Analysis for Land Surface models (PALS), the Program for Climate Model Diagnosis and Intercomparison (PCMDI) Metrics Package (PMP), the Earth System Model Evaluation Tool (ESMValTool), and the Land surface Verification Toolkit (LVT)—were reviewed, and the new ILAMB benchmarking systems were described and demonstrated.

The ILAMB version 1 (v1) and ILAMB version 2 (v2) benchmarking systems compare model results with best-available observational data products, focusing on atmospheric CO_2, surface fluxes, hydrology, soil carbon and nutrient biogeochemistry, ecosystem processes and states, and vegetation dynamics. ILAMBv2 is expected to become an integral part of the workflow for model frameworks, including the Accelerated Climate Modeling for Energy (ACME) model and the Community Earth System Model (CESM). Moreover, ILAMBv2 will contribute model analysis and evaluation capabilities to phase 6 of the Coupled Model Intercomparison Project (CMIP6) and future model and model–data intercomparison projects.

Benchmarking Challenges and Priorities

A variety of statistical approaches have been adopted to evaluate model accuracy through comparison with observations, including calculations of bias, root-mean-square error (RMSE), phase, amplitude, spatial distribution, Taylor diagrams and scores, functional relationship metrics, and perturbation and sensitivity tests. While many of these statistical measures are not independent, each provides slightly different information about contemporary model performance with respect to observational data and about implications for future projections from ESMs.



However, developing metrics that make appropriate use of observational data remains a scientific challenge because of the spatial and temporal mismatch between models and measurements, poorly characterized uncertainties in observationally constrained data products, biases in reanalysis and forcing data, model simplifications, and structural and parametric uncertainties. A variety of benchmarking challenges and opportunities emerged from workshop breakout group meeting reports. Common themes included the following:

- Need for collocated measurements, particularly around a core set of AmeriFlux and FLUXNET sites with a sustained record of observations for repeated model testing
- Lack of quantified uncertainty information for observational data
- Utility of functional response metrics and variable-to-variable comparisons
- Value of metrics for future projections based on emergent constraints

- Unrealized opportunities for global observational data sets based on satellite remote sensing synthesized with ancillary databases, using new algorithms
- Importance of applying statistical and machine learning methods to upscaling sparse measurements from sites to regions to the globe
- Need for process-level benchmarks and metrics for extreme events
- Opportunities for collaboration with earth system model developers (e.g., ACME, CESM, and others)
- Opportunities for collaboration with important field and laboratory experiments and monitoring activities, including AmeriFlux and FLUXNET, Integrated Carbon Observation System (ICOS), Next Generation Ecosystem Experiments (NGEE) Arctic, Arctic–Boreal Vulnerability Experiment (ABoVE), Spruce and Peatland Responses Under Climatic and Environmental Change (SPRUCE) project, Critical Zone Observatories (CZOs), Long-Term Ecological Research (LTER) sites, National Ecological Observatory Network (NEON), NGEE Tropics, and Tropical Responses to Altered Climate Experiment (TRACE).

Recommendations for next-generation Benchmarking Challenges and Priorities included the following:

- Develop supersite benchmarks integrated with AmeriFlux and FLUXNET
- Create benchmarks for soil carbon turnover and vertical distribution and transport
- Develop benchmark metrics for extreme event statistics and response of ecosystems
- Synthesize data for vegetation recruitment, growth, mortality, and canopy structure
- Create benchmarks focused on critical high latitude and tropical forest ecosystems
- Leverage observational projects and create a roadmap for remote sensing methods.

Model Intercomparison Project (MIPs)

Model Intercomparison Project (MIPs) are important activities for assessing the coherence and reliability of ESMs. Ongoing and future MIPs focused on modeling terrestrial water, energy, and carbon cycles are particularly relevant to ILAMB. Benchmarking needs were evaluated for the CMIP6 historical and Diagnostic, Evaluation, and Characterization of Klima (DECK) experiments; the Coupled Climate–Carbon Cycle MIP (C^4MIP); the Land Surface, Snow and Soil Moisture MIP (LS3MIP); and the Land Use MIP (LUMIP). Opportunities for benchmarking model results from other MIPs were also considered.

Key recommendations that emerged on MIP benchmarking needs were the following:

- Develop methods to attribute emergent model behaviors such as carbon feedback parameters to specific processes through emergent constraint and traceability approaches
- Benchmark across coupling and complexity hierarchies—from offline land-only simulations to fully coupled ESMs—to attribute model biases and uncertainties to specific domains and identify feedbacks between domains
- Develop paired site data sets for benchmarking model representations of subgrid scale heterogeneity.

Benchmarking Approaches

New and existing Benchmarking Approaches were identified from the workshop. While traditional statistical comparisons with observations offer a great deal of information about model performance, metrics based on functional responses or variable-to-variable comparisons often suggest why models produce incorrect results. Benchmarking future projections can be accomplished through careful use of emergent constraints. Reduced complexity models and traceability frameworks are usefully applied to enable greater process understanding through more frequent and detailed testing with reduced computational costs. Formal uncertainty quantification (UQ) frameworks and methods, described in papers in the appendix, provide rigorous techniques for understanding model predictions. Finally, meta-analyses of perturbation experiments provide a new approach for constraining model predictions of ecosystem responses under controlled environmental change conditions.

Enabling Capabilities

To address the identified next generation Benchmarking Challenges and Priorities, certain Enabling Capabilities are needed. New model development focused on improving process representations is required, and additional model variables should be saved for comparison with data. A new Land Model Testbed (LMT) capability employing community benchmarks and supporting UQ frameworks would enable more rapid model development and verification, particularly for major ESM frameworks like ACME and CESM.

Additional field measurements and monitoring activities, as well as perturbation experiments and lab studies, could provide valuable observational data for constraining models. High priority measurement needs for developing benchmarks and improving ESMs include the following:

- Long-term energy, carbon, and water flux measurements at AmeriFlux and FLUXNET sites with standardized instrumentation and methods, and additional frequent or continuous ancillary *in situ* measurements of soil moisture, sap flow, tree height and diameter, litterfall, and soil nutrients
- High latitude and tundra soil core measurements of carbon and nutrient distributions, including isotopes and ice/water content, and observations of vegetation growth and expansion of woody vegetation
- Characterization of tropical ecosystem traits and canopy structure and chemistry; observations of tropical ecosystem responses to drought, increased temperatures, and elevated atmospheric CO_2; and measurements of nutrient cycling and hydrology in tropical forests, focusing on land–atmosphere interactions
- Remote sensing algorithms and processing infrastructure for generating data products useful for large-scale ecosystem characterization and monitoring, scaling up *in situ* measurements, and informing future measurement site selection.

Improved observational data archives (e.g., DOE Atmospheric Radiation Measurement (ARM) Climate Research Facility and Environmental System Science (ESS) archives, NASA Distributed Active Archive Centers (DAACs)) and repositories (e.g., Obs4MIPs) are needed that offer data discovery, server-side analysis, and advanced distribution capabilities. Finally, new computational resources and cyber infrastructure will be required to realize the promise of new benchmarking capabilities. This infrastructure needs to offer a balance between pure compute capacity (high core count) and throughput (e.g., cache size, memory size and bandwidth, and input/output bandwidth) to support *in situ* analysis and benchmarking, growing observational data sets, and multi-model comparisons.

Conclusions and Next Steps

The 2016 ILAMB Workshop was successful in bringing together the international community to identify scientific challenges and priorities for future research. The workshop demonstrated broad interest on the part of a vibrant community of scientists spanning many disciplines that are committed to reducing barriers for information flow between the measurement and modeling communities.

To effectively address the individual processes and cross-cutting themes discussed above, small, targeted working groups should be formed to research and publish supporting analyses. A top priority is supporting CMIP6 activities, where additional development of ILAMB functionality could yield powerful automated analyses and model intercomparison capabilities for such national and international assessment efforts.

Over the next decade, the community envisions the ILAMB system to serve as a core capability within a U.S. or international center that will provide a home to focused synthesis working groups, host MIP-related activities, and support expanded use of, and access to, ESMs by a broader cross section of scientists within disciplines of ecosystem ecology, biogeochemistry, and hydrology.

Contents

Executive Summary iii

1.0 Model Benchmarking Principles and Workshop Introduction 1

2.0 Benchmarking Tools 6
- 2.1 Evaluation and Benchmarking Tools 6
- 2.2 Other Model Evaluation Capabilities in Use at Modeling Centers 8
- 2.3 Synergies Across Different Benchmarking Activities 9

3.0 Current Status of the ILAMB Software Packages 11

4.0 Next Generation Benchmarking Challenges 14
- 4.1 Major Processes 15
  - 4.1.1 Carbon and Energy Fluxes 15
  - 4.1.2 Soil Carbon and Nutrient Biogeochemistry 17
  - 4.1.3 Hydrology 17
  - 4.1.4 Vegetation Dynamics and Biomass 18
- 4.2 Integrating and Cross-cutting Themes 19
  - 4.2.1 High Latitude Processes 19
  - 4.2.2 Tropical Processes 20
  - 4.2.3 Remote Sensing 21
  - 4.2.4 Process-specific and Perturbation Experiments 22

5.0 Model Intercomparison Projects (MIPs) 23
- 5.1 The Roles of Benchmarking in MIPs 23
- 5.2 Descriptions of MIPs and Their Benchmarking Needs 23
  - 5.2.1 CMIP6 Historical and DECK 23
  - 5.2.2 C^4MIP 23
  - 5.2.3 LS3MIP 24
  - 5.2.4 LUMIP 24
  - 5.2.5 MsTMIP 24
  - 5.2.6 PLUME-MIP 24
- 5.3 New Metrics, Approaches, and Model Output Requirements 25
- 5.4 Available Observations and Data Gaps 25
- 5.5 Expected Results from MIPs and ILAMB 26

6.0 Model Development and Evaluation Testbeds 27

7.0 Traceability and Uncertainty Quantification Frameworks 31
- 7.1 Traceability Framework 31
- 7.2 Scientific Driver for UQ of LSM 33
- 7.3 Observational Data Needs 35
- 7.4 Algorithm Needs 36
- 7.5 Computational, Visualization, and Data Analysis Needs 36

8.0 Computational Needs and Requirements 38

9.0 Conclusions and Next Steps 40
- 9.1 Workshop Conclusions 40
- 9.2 Long-term Vision for Model Benchmarking 42

Appendix A. Benchmarking Tools 44
A.1 PALS/PLUMBER 44
A.2 PCMDI Metrics Package (PMP) 46
A.3 ESMValTool 47
A.4 NASA Land Surface Verification Toolkit (LVT) 50
A.5 ABoVE Benchmarking System 51
Appendix B. Metrics for Major Processes 52
B.1 Ecosystem Processes and States 52
Specific Points and Recommendations 53
B.1.1 Scientific Challenges and Opportunities for Model Evaluation 54
B.1.2 New Metrics and Benchmarking Approaches 54
B.1.3 Observational Data Needs and Priorities 55
B.1.4 Model Development and Output Requirements 56
B.2 Hydrology 57
B.2.1 Scientific Challenges and Opportunities for Model Evaluation 57
Current State of Process Representations in Models 57
Existing Approaches for Assessing Model Performance 57
B.2.2 New Metrics and Benchmarking Approaches 58
New Metrics, Scores, and Functional Relationships 58
Current Best-available Data Sets for Specific New Metrics 58
B.2.3 Observational Data Needs 58
Gaps in Current Data Availability 58
New *in situ* or Remote Sensing Measurement Needs 59
Spatial and Temporal Extent and Resolution Requirements 59
Synthesis Activities Needs and Approaches 59
B.3 Atmospheric CO_2 59
B.3.1 Scientific Challenges and Opportunities for Model Evaluation 59
B.3.2 New Metrics and Benchmarking Approaches 60
B.3.3 Observational Data Needs 60
B.3.4 Model Development and Output Requirements 61
B.4 Soil Carbon and Nutrient Biogeochemistry 61
B.4.1 Introduction 61
B.4.2 Scientific Challenges and Opportunities for Model Evaluation 61
B.4.3 Observational Data, New Metrics, and Benchmarking Approaches 62
B.5 Surface Fluxes (Energy and Carbon) 63
B.5.1 Scientific Challenges and Opportunities for Model Evaluation 63
Specific Points and Recommendations 64
B.5.2 New Metrics and Benchmarking Approaches 64
B.5.3 Observational Data Needs 65
B.5.4 Model Development and Output Requirements 65
B.6 Vegetation dynamics 65
B.6.1 Scientific Challenges and Opportunities for Model Evaluation 65
B.6.2 Observational Data Needs 66
B.6.3 New Metrics and Benchmarking Approaches 66

Appendix C. Metrics for Integrating and Cross-cutting Themes 69
C.1 Process-specific Experiments 69
C.1.1 Scientific Challenges and Opportunities for Model Evaluation 69
C.1.2 New Metrics and Benchmarking Approaches 69
C.1.3 Experimental/Observational Data Needs 70
C.1.4 Model Development and Output Requirements 71
C.2 Metrics from Extreme Events 71
C.2.1 Scientific Challenges and Opportunities for Model Evaluation 71
C.2.2 New Metrics and Benchmarking Approaches and Observational Data Needs 72
C.3 Design of New Perturbation Experiments 75
C.3.1 Scientific Challenges and Opportunities for Model Evaluation 75
C.3.2 New Metrics and Benchmarking Approaches 77
C.3.3 Observational Data Needs 77
C.3.4 Model Development and Output Requirements 78
C.4 High Latitude Processes 78
C.4.1 Scientific Challenges and Opportunities for Model Evaluation 78
C.4.2 New Metrics and Benchmarking Approaches 79
C.4.3 Observational Data Needs 79
C.4.4 Model Development and Output Requirements 80
C.5 Tropical Processes 80
C.5.1 Scientific Challenges and Opportunities for Model Evaluation 80
C.5.2 New Metrics and Benchmarking Approaches 80
C.5.3 Observational Data Needs 81
C.5.4 Model Development and Output Requirements 81
C.6 Remote Sensing 82
C.6.1 Scientific Challenges and Opportunities for Model Evaluation 82
C.6.2 New Metrics and Benchmarking Approaches 83
C.6.3 Observational Data Needs 86
C.6.4 Potential Pitfalls and Misuse of Remote Sensing in Model Benchmarking 88
C.6.5 Model Development and Output Requirements 89
C.6.6 Computational Needs and Requirements 90
C.7 Roles for Flux Networks 90
C.7.1 FLUXNET: A Network of Eddy Covariance Flux Measurement Networks 90
C.7.2 Current and Future Roles of FLUXNET for Carbon Cycle Synthesis 91

Appendix D. Model Intercomparison Project (MIP) Benchmarking Needs and Evaluation Priorities 94
D.1 CMIP6 Historical and DECK 94
D.1.1 Scientific Challenges and Opportunities for Model Evaluation 94
D.1.2 New Metrics and Benchmarking Approaches 95
D.2 C^4MIP 95
D.2.1 Scientific Challenges and Opportunities for Model Evaluation 95
D.2.2 New Metrics and Benchmarking Approaches 96
D.2.3 Observational Data Needs 96
D.2.4 Model Development and Output Requirements 97

D.3 LS3MIP 97
D.4 LUMIP 99
D.4.1 Land-use Metrics 100
D.4.2 Land-only Versus Coupled Model Assessment 101
D.4.3 Subgrid Data Reporting and Analysis 101
D.5 MsTMIP 101
D.6 PLUME-MIP 102
Appendix E. Integration with Uncertainty Quantification Frameworks 104
E.1 An Uncertainty Quantification Framework Designed for Land Models 104
E.2 Use of Emulators in Uncertainty Quantification 107
E.3 Uncertainty Quantification in the ACME Land Model: Summary 108
E.4 The Predictive Ecosystem Analyzer (PEcAn): A Community Tool to Enable Land Model Synthesis, Evaluation, and Forecasting 110
Appendix F. ILAMB 2016 Workshop Materials 117
F.1 Agenda 117
F.2 Plenary Presentation Abstracts 120
F.3 List of On-site Participants 133
Appendix G. References 134
Appendix H. Acronyms and Abbreviations 157

1.0 Model Benchmarking Principles and Workshop Introduction

As Earth system models become increasingly complex, there is a growing need for comprehensive and multi-faceted evaluation of model projections. To advance understanding of biogeochemical processes and their interactions with hydrology and climate under conditions of increasing atmospheric carbon dioxide (Figure 1.1), new analysis methods are required that use observations to constrain model predictions, inform model development, and identify needed measurements and field experiments. Better representations of biogeochemistry–climate feedbacks and ecosystem processes in these models are essential for reducing uncertainties associated with projections of climate change during the remainder of the 21st century.

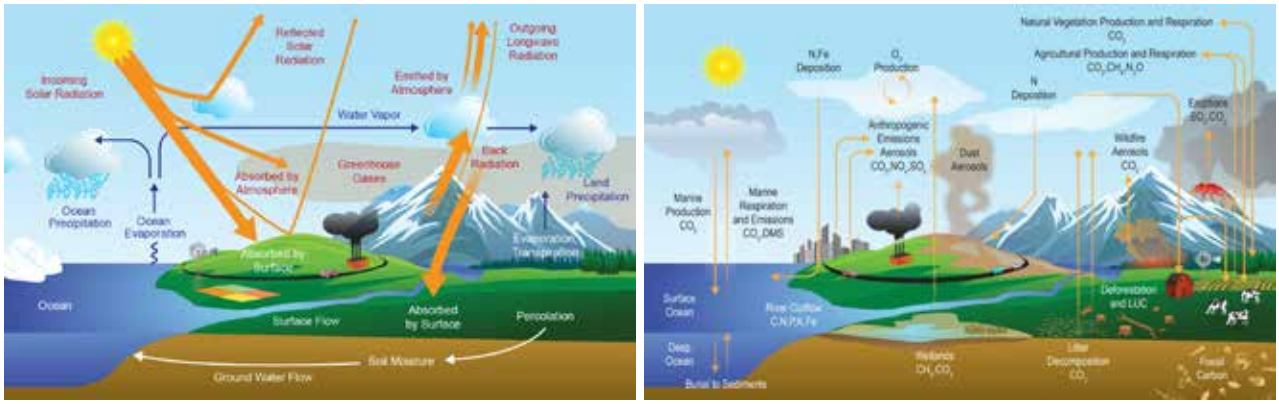


Figure 1.1. Today's advanced Earth system models must represent the interacting energy and radiation dynamics and water cycle processes (left) as well as the geochemical and biological processes that control global carbon and nutrient cycles (right) under conditions of increasing atmospheric carbon dioxide.

Building upon past model evaluation studies, the goals of the **International Land Model Benchmarking (ILAMB)** activity (Section 3; **https://www.ilamb.org/**) are the following:

1. Develop internationally accepted benchmarks for land model performance by drawing upon international expertise and collaboration.
2. Promote the use of these benchmarks by the international community for model intercomparison.
3. Strengthen linkages among experimental, remote sensing, and climate modeling communities in the design of new model tests and new measurement programs.
4. Support the design and development of a new, open source, benchmarking software system for use by the international community.

To further these goals and advance the development of benchmarking software tools for use by the international community, a diverse team of national laboratory and university researchers sponsored by the US Department of Energy is engaged in developing new diagnostic approaches and model benchmarking tools for evaluating Earth System Model (ESM) hydrological and biogeochemical process representations. Collaborating through the **Biogeochemistry–Climate Feedbacks Scientific Focus Area (BGC Feedbacks SFA)** project (**https://www.bgc-feedbacks.org/**), this team performs simulations, analyses, and benchmarking to identify model weaknesses that lead to model improvements and determine needed measurements that inform the design of future field campaigns (Figure 1.2). Research activities such as the BGC Feedbacks SFA play a critical role in the model–data experimentation (ModEx) enterprise for the US Department of Energy and other agencies by connecting field and laboratory data with models and producing syntheses, analysis methods, and open source tools that are made available to the larger international scientific community (Figure 1.3).

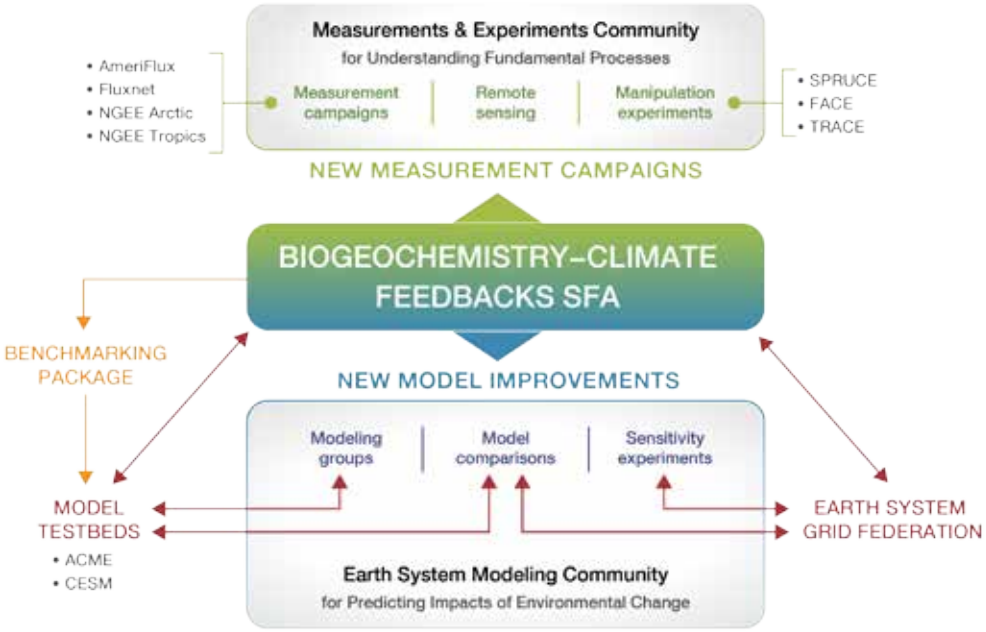


Figure 1.2. The Biogeochemistry–Climate Feedbacks Scientific Focus Area (SFA) uses best-available observational data sets to evaluate the fidelity of Earth system models. Open source benchmarking tools are produced to support model testbeds for Accelerated Climate Modeling for Energy (ACME) and Community Earth System Model (CESM) frameworks. The project identifies model gaps and weaknesses, informs new model development, and suggests new measurement and field campaigns.

The benchmarking system developed by the BGC Feedbacks SFA compares model results with best-available observational data products, focusing on atmospheric CO_2, surface fluxes, hydrology, soil carbon and nutrient biogeochemistry, ecosystem processes and states, and vegetation dynamics. The system is expected to become an integral part of model verification for future rapid model development cycles for the model frameworks from the **Accelerated Climate Modeling for Energy (ACME)** project and the **Community Earth System Model (CESM)**. Moreover, it will contribute model analysis and evaluation capabilities to phase 6 of the Coupled Model Intercomparison Project (CMIP6) and future model and model–data intercomparison experiments.

The second ILAMB Workshop in the United States was convened on May 16–18, 2016, in Washington, District of Columbia, USA. **The overarching goal of the workshop was to engage the international research community in defining the scientific priorities for the design of new metrics, the identification of model development and workflow practices, and CMIP6 evaluation needs, and to learn about new observational data sets and measurement campaigns**. The workshop drew more than 60 on-site participants and included attendees from Australia, Japan, China, Germany, Sweden, Netherlands, United Kingdom, and the United States. They represented 10 different major modeling centers. Approximately 90 individuals registered to participate remotely, and between 20 and 30 were online at any time during the plenary sessions, including students and postdocs from various universities and labs and invitees unable to travel from Canada, China, and elsewhere. The workshop agenda, presentation abstracts, and the participant list are contained in Appendix F. Plenary presentations focused on model benchmarking, emergent constraints, evaluation metrics, uncertainty quantification, and measurement networks.

SECOND ILAMB WORKSHOP IN THE U.S.

More than 5 years after the first ILAMB workshop in the United States in 2011, the 2016 ILAMB workshop, jointly sponsored by the U.S. Department of Energy's Regional & Global Climate Modeling (RGCM) and Earth System Modeling (ESM) Programs, was convened to:

- Highlight new techniques and metrics for model evaluation, including applications of the emergent constraints approach.
- Enable coordination among the Coupled Climate–Carbon Cycle Model Intercomparison Project; Land Surface, Snow, and Soil Moisture Model Intercomparison Project; and the Land Use Model Intercomparison Project.
- Increase awareness of new tools that will be available for model verification and benchmarking, drawing upon data streams from field experiments, remote sensing, *in situ* measurements, and synthesis activities.
- Increase the use and sharing of information and community tools for model evaluation and benchmarking.
- Design new metrics and evaluation approaches for integration into the next generation ILAMB system.
- Create new metrics that integrate across carbon, surface energy, hydrology, and land use disciplines.

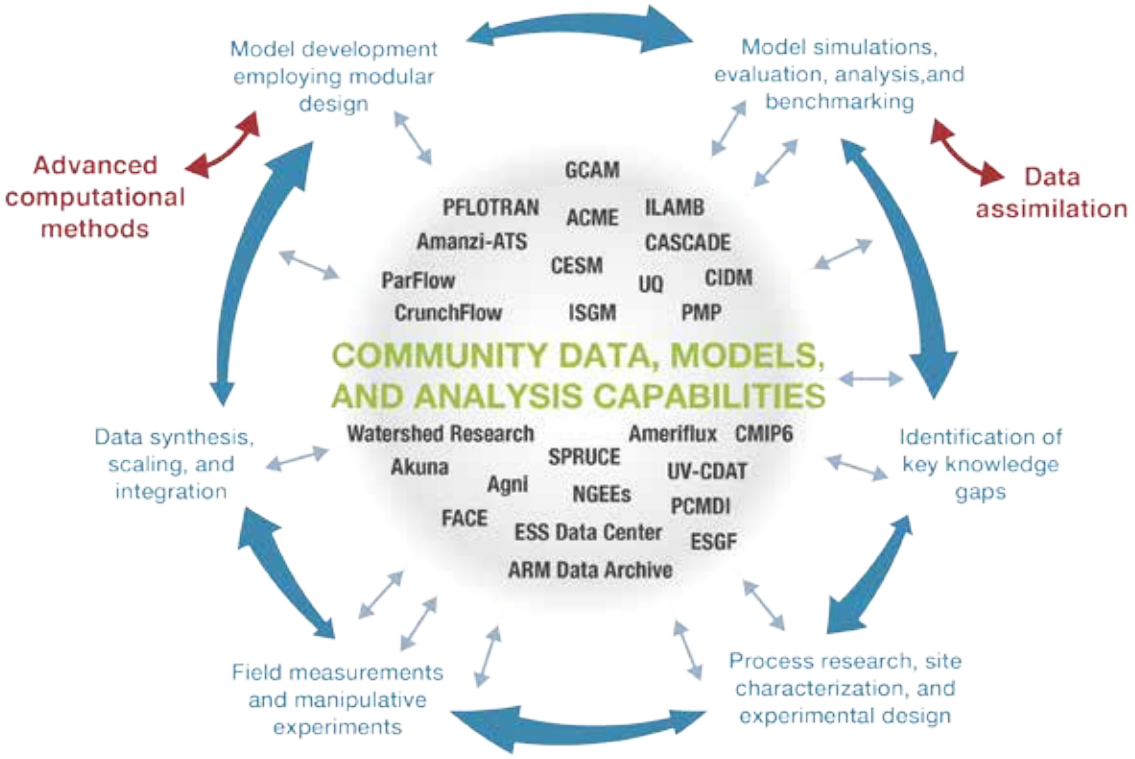


Figure 1.3. Model simulations and benchmarking play a critical role in the model–data experimentation (ModEx) enterprise outlined in this diagram. By identifying model weaknesses and knowledge gaps, benchmarking helps inform process research and experimental design, which generate data that drives new model development in a cyclic fashion. All of these steps both use and produce data, models, and analysis capabilities and tools that can be shared and used by the larger international research community.

The white papers in the Appendix of this report were authored through "crowdsourcing" for the widest possible engagement with researchers at the workshop, attending remotely, or with general interest in model evaluation. Breakout group co-leads and plenary presenters, listed as authors of the respective white papers, contributed additional effort to resolve comments and produce the combined draft form of the report. In addition to transmitting audio and slides over the Internet from all plenary sessions, workshop updates from various participants were provided to the community via social media (see sidebar on *The Cloud and Social Media at the ILAMB Workshop*). On the second and third afternoons of the workshop, ILAMBv2 software tutorials were webcast to increase outreach to students, postdocs, and early career scientists interested in land model benchmarking.

This report provides a synopsis of the current state of the science and highlights challenges and opportunities for benchmarking, model development, and field and laboratory measurements needed to advance climate science. The main text provides a synthesis of the ideas, concepts, and scientific priorities presented and discussed at the workshop. Section 4 highlights benchmarking priorities identified by the scientific community. Categorized as *Major Processes* (detailed in Appendix B) and *Integrating and Cross-cutting Themes* (detailed in Appendix C), these topics are listed below, and the process by which the corresponding white papers were synthesized for the main body of the report is summarized in Figure 1.4.

Major Processes

- ecosystem processes and states
- hydrology
- atmospheric CO_2
- soil carbon and nutrient biogeochemistry
- surface fluxes (energy and carbon)
- vegetation dynamics

Integrating and Cross-cutting Themes

- process-specific experiments
- metrics from extreme events
- design of new perturbation experiments
- high latitude processes
- tropical processes
- remote sensing
- eddy covariance flux networks

The Appendix of this report summarizes the invited presentations, describes breakout group proceedings and recommendations, and identifies critical gaps and opportunities in measurement programs, new approaches for model evaluation, and synergies among research teams and tools being constructed to support model development, parameter estimation, and model–data integration.

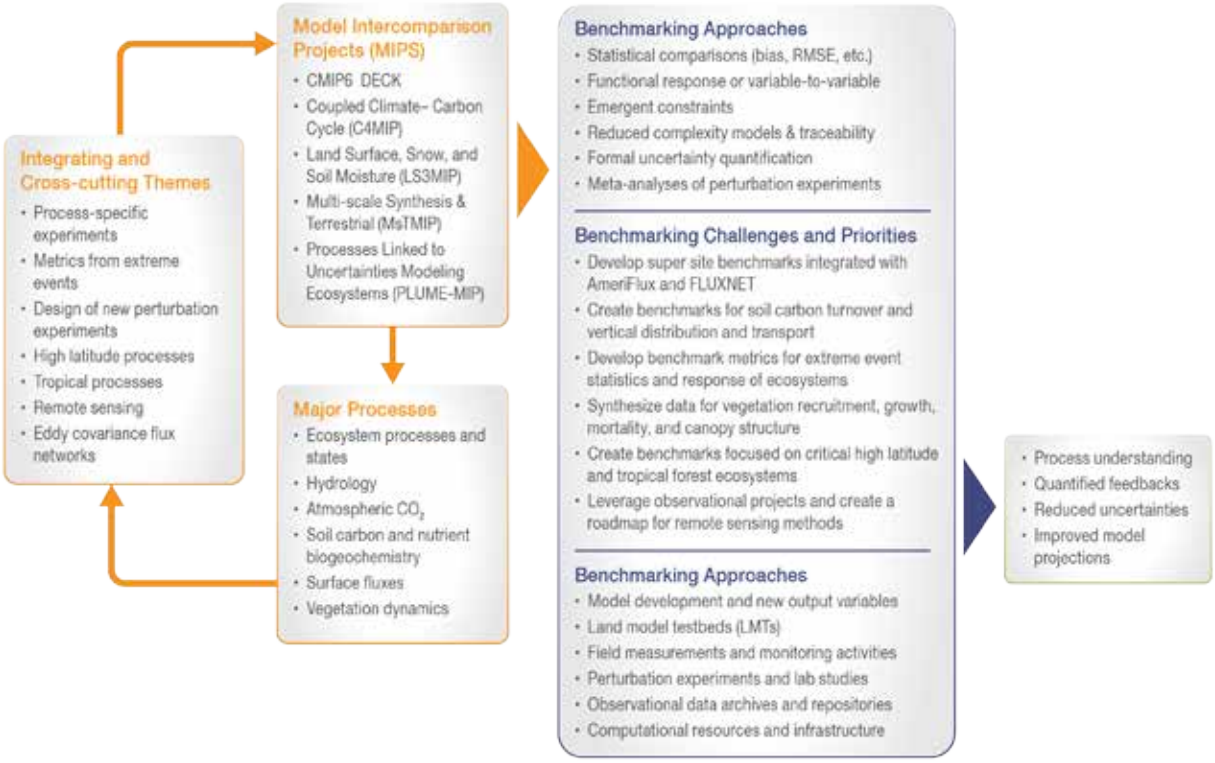


Figure 1.4. The topical white papers within the categories of *Major Processes* and *Integrating and Cross-cutting Themes* were synthesized with those on the needs of *Model Intercomparison Projects (MIPs)* to produce a set of next generation *Benchmarking Challenges and Priorities* resulting from the workshop. In addition, *Benchmarking Approaches* for addressing these challenges were identified and *Enabling Capabilities* needed to facilitate next generation benchmarking and model development were distilled from the white papers. Addressing these challenges will advance climate science by enabling process understanding, quantifying feedbacks, reducing uncertainties, and improving model predictions.

Section 2 describes a collection of existing land model evaluation or benchmarking tools and identifies other model evaluation capabilities currently employed in international climate modeling centers. Strengths and weaknesses of these existing approaches are considered in the discussion of potential synergies for future development across varied benchmarking packages. Section 3 presents an overview of the ILAMB Software Packages (ILAMBv1 and ILAMBv2) released to the community at the workshop. Section 4 focuses on next generation benchmarking challenges, identified by the international community, for confronting models. Suggestions for careful consideration of how best to employ measurements and observationally constrained data products are an important community contribution from the workshop. In some cases, the community would benefit from synthesis of existing data or from entirely new measurements. Section 5 describes future model intercomparison projects (MIPs), particularly those associated with the 6th phase of the Coupled Model Intercomparison Project (CMIP6), and discusses model evaluation needs, challenges, and opportunities expected in the future.

Section 6 describes a proposed land model development and evaluation testbed methodology and highlights specific metrics and datasets identified for evaluating new process parameterizations being developed for the ACME Land Model (ALM). Section 7 illustrates a mathematical methodology for evaluating structural components of carbon cycle models and describes approaches for integration of uncertainty quantification techniques into model benchmarking activities and tools. Section 8 presents computational needs and challenges for large scale climate data analytics, with a focus on model evaluation and benchmarking. Finally, Section 9 describes next steps for the scientific enterprise of model benchmarking through focused mini-workshops, use of extensible archives for data expressly designed for model comparison (e.g., obs4MIPs), and community research opportunities centered on science questions to be addressed by large MIPs. The Appendixes that follow these sections provide detailed descriptions of presentations, notes from meeting sessions, and citations to relevant research in support of the main body of the report.

THE CLOUD AND SOCIAL MEDIA AT THE ILAMB WORKSHOP

Conferencing technology, document crowdsourcing in the Cloud, and social media were all employed at the ILAMB Workshop to maximize community participation. Audio and slides from plenary sessions all three days were transmitted over the Internet through software called BlueJeans.

All slides and meeting notes were developed and edited by workshop participants using Google Slides and Google Docs, allowing local and remote attendees to contribute notes and comments for any plenary or breakout group session. Twitter was employed by many participants to make comments, post ideas, or ask questions during the workshop. A sampling of these tweets is shown here.

This workshop report was developed by crowdsourcing through the community using Google Docs, which enabled participants to continue contributing new ideas, figures, and references to relevant research right up until final production.

The use of technology even helped reduce gender, racial, and cultural imbalances among workshop participants since female caretakers could attend from their homes and researchers in foreign countries could attend without traveling long distances.

2.0 Benchmarking Tools

2.1 Evaluation and Benchmarking Tools

To prepare ILAMB Workshop participants for discussions of model evaluation and benchmarking, several of the leading benchmarking tools being employed by the research community were reviewed and presented by invited workshop speakers. These tools, some of which were designed specifically for evaluating land models and others for general applicability to Earth system models, are described here. The **Protocol for the Analysis for Land Surface models (PALS**; Abramowitz, 2012) is an online web application for the automated evaluation and benchmarking of land surface model (LSM) simulations. The **Program for Climate Model Diagnosis and Intercomparison (PCMDI) Metrics Package (PMP**; Gleckler et al., 2016) emphasizes summary statistics that gauge model errors across a range of spatial and temporal scales for the atmosphere, ocean, and sea ice, and is designed to deliver systematic benchmarking for Coupled Model Intercomparison Project (CMIP) Diagnostic, Evaluation and Characterization of Klima (DECK) simulations. The **Earth System Model Evaluation Tool (ESMValTool**; Eyring et al., 2016a) is a community effort to encourage open exchange of diagnostic source code and evaluation results through a standardized workflow framework. The **Land surface Verification Toolkit (LVT**; Kumar et al., 2012), originally designed to support NASA's Land Information System (LIS; Kumar et al., 2006), is an automated evaluation framework that incorporates a model–data fusion paradigm. The new **ILAMB packages** (Section 3), ILAMBv1 (Mu et al., 2016) and ILAMBv2 (Collier et al., 2016), are open source land model evaluation systems that operate on global-, regional-, and site-level data and provide a hierarchical scoring system to indicate model fidelity (Mu et al., in prep.).

The ESM community agrees that systematic model assessment should be a routine component of the model development process. Benchmarking systems should provide a mechanism for archiving of previous results in a manner that allows for ease of viewing later. For example, ILAMB facilitates the comparison of multiple models or model versions simultaneously (e.g., Figure 2.1). Scores for individual metrics can easily be compared to determine the tradeoffs resulting from model modifications. Likewise, the PALS system retains all datasets, analysis scripts, and results for efficient comparison across model versions. The second phase of PALS will introduce a distributed architecture in which analysis nodes are located at modeling centers to circumvent the need for repeated transfers of large files that may be a barrier to routine model evaluation.

Model evaluation tools should be designed to test the predictive power of a model under conditions of a changing climate. Given that direct model evaluation is possible only with contemporary observations, it is difficult to establish whether a model has predictive skill. However, within the ILAMB system, development of functional benchmarks to relate biogeochemical or biogeophysical responses to a physical driver will test whether a model can accurately simulate the relationship between a model variable and a physical driver across a range of driver values. When a model can reproduce functional relationships across a full range of present day climate regimes, for example, it may yield more robust responses to future change (e.g., Figure 2.2). While this approach is indirect, it moves beyond simple time series or spatial comparisons to probe relationships among variables and drivers or among variables and other variables. These relationships may then be useful for testing future predictions using emergent constraint approaches. A complementary approach for testing model predictivity, prototyped in the LVT, are metrics based on information

KEY RECOMMENDATIONS

- Well-established aspects of model assessment should be a routine component of the model development process that over time becomes increasingly comprehensive.
- Evaluation tools should include testing the predictive power of models under a changing climate.
- Benchmarking packages should span a wide range of spatial and temporal scales and extents.
- Integration of a diversity of evaluation tools into a common workflow framework could lead to new insights into climate processes and phenomena.
- Evaluation and benchmarking systems should be open source and freely distributed to leverage the work of many modeling teams and to minimize redundancy.
- Benchmarking tools should be integrated with data repositories that support standardized access through an applications programming interface.

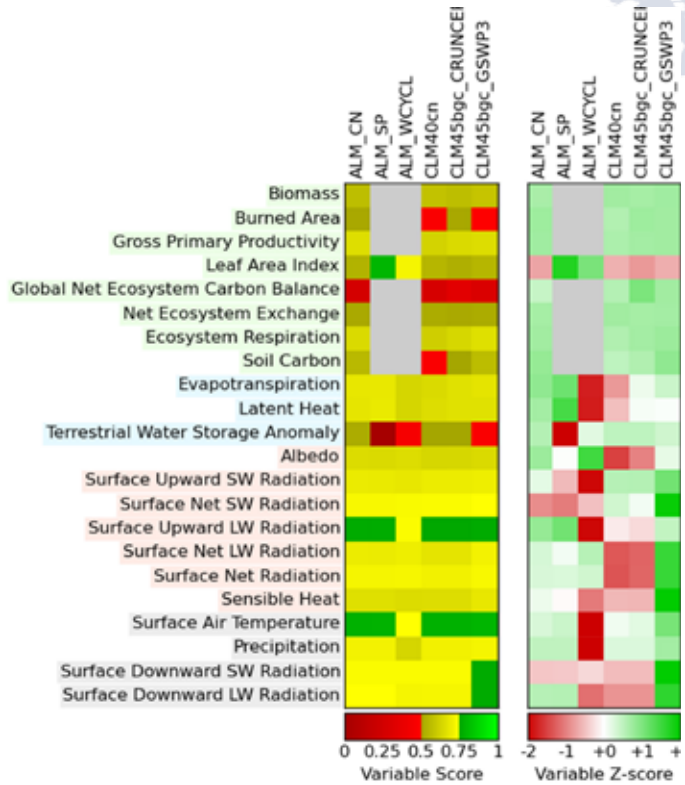


Figure 2.1. The ILAMBv2 package produces a summary graphic depicting model performance across a wide variety of variables, emphasizing absolute performance (left) as well as relative performance (right) with respect to comparisons with observations. This figure compares results from the ACME Land Model (ALM) run offline with carbon–nitrogen (CN) biogeochemistry (ALM_CN), run offline in satellite phenology (SP) mode (ALM_SP), and fully coupled in SP mode (ALM_WCYCL) with the Community Land Model (CLM) run offline for CLM-4.0 (CLM40cn), for CLM-4.5-BGC (CLM45bgc_CRUNCEP) and for CLM-4.5-BGC with Global Soil Wetness Project version 3 (GSWP3) forcing (CLM45bgc_GSWP3). (image to the right)

theory. By considering entropy or information content within model output, a package may be able to evaluate the robustness of model predictions to a different mean state.

Building a benchmarking system that spans spatial and temporal scales is crucial. Land surface processes are heterogeneous, but the climate impact of biogeochemical exchange with the atmosphere is global. ILAMB currently incorporates both global gridded observations and site-level time series and offers a scheme for scoring performance on both kinds of comparisons, representing both spatial and temporal aspects of model performance. The LVT system dynamically transforms model output to match the scale of observational constraints. The PALS system is presently limited to a set of flux tower sites with high data density in the temporal domain and ancillary observations.

Ultimately, linking ILAMB to existing model evaluation tools for physical climate will facilitate improved prediction in fully coupled Earth system models. The PCMDI Metrics Package and the ESMValTool are community tools designed to evaluate a set of outputs complementary to ILAMB, especially from non-terrestrial components of the Earth system. We see opportunities for linking with these packages because a lack of fidelity in the simulation of physical climate in biogeochemical hotspots, such as the Amazon, may induce a cascade of impacts across ecosystems, aerosols, atmospheric chemistry, and atmospheric dynamics. Routinely employing ILAMB or other diagnostics packages for analysis of the 6th phase of the Coupled Model Intercomparison Project (CMIP6) will facilitate sharing of process-level insights for more rapid and productive future model development and evaluation.

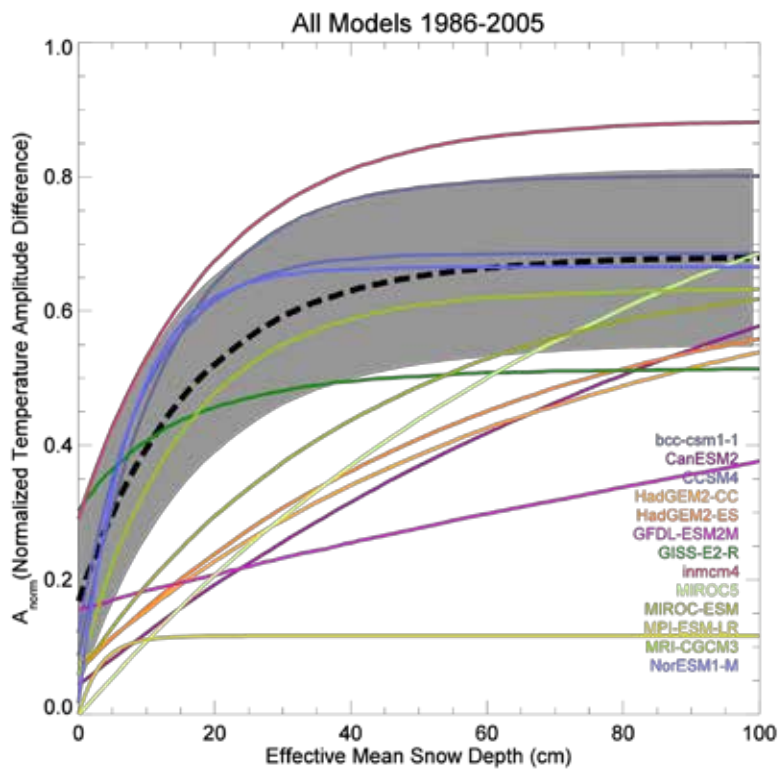


Figure 2.2. A metric for heat transfer through snow. The dashed line and gray shading show observed relation between the normalized difference in the amplitude of the annual cycle of air temperature versus near-surface soil temperature at different levels of effective mean snow depth. Colored lines represent the snow heat transfer relationship as obtained from CMIP5 models (Figure 4 of Slater et al., 2016).

Many model evaluation packages are open source community tools, and such a free and open framework facilitates wide use of the benchmarking system because users can add evaluation metrics or sub-select from existing metrics as desired. Challenges to adoption, integration with other tools, and cooperative development include standards for file formats and data conventions, programming languages, and the diversity of computational architectures required to support single-point to high resolution global analyses. Given that most Earth system modeling centers do not presently share evaluation packages, building flexibility into the structure of news tools is likely to minimize redundant effort across centers. Opportunities to leverage developments across modeling centers should be pursued by engaging with ongoing data infrastructure efforts for CMIP and more broadly the World Climate Research Programme (WCRP).

2.2 Other Model Evaluation Capabilities in Use at Modeling Centers

Modeling centers presently employ a patchwork of model evaluation methodologies. A survey conducted prior to the 2016 ILAMB Workshop, designed to gauge the philosophies and approaches used for model evaluation, confirmed unanimous community interest in comprehensive evaluation tools, with all modeling centers reporting that evaluation played multiple roles in the model development process. Although the primary reported use for model evaluation was to diagnose errors in the model, modeling centers also use their evaluation packages to tune model parameters and to aid with model analyses.

Responses from the modeling centers also revealed the need for community-based approaches to share best practices. Although most modeling centers had their own model evaluation package, some of these packages are slanted toward general diagnostics rather than land-specific diagnostics. Of these packages, roughly half included quantitative metrics and scoring; however, most of the packages also relied significantly on expert judgment, such as for interpreting graphical comparisons between model output and observational constraints. An impediment to quantitative comparisons was the perception that data quality varies widely from one dataset to another. For quantitative comparisons, several modeling centers had already begun to rank variables by the availability and quality of observations (e.g., Figure 2.3), both for prioritizing the integration of new variables into their package and for

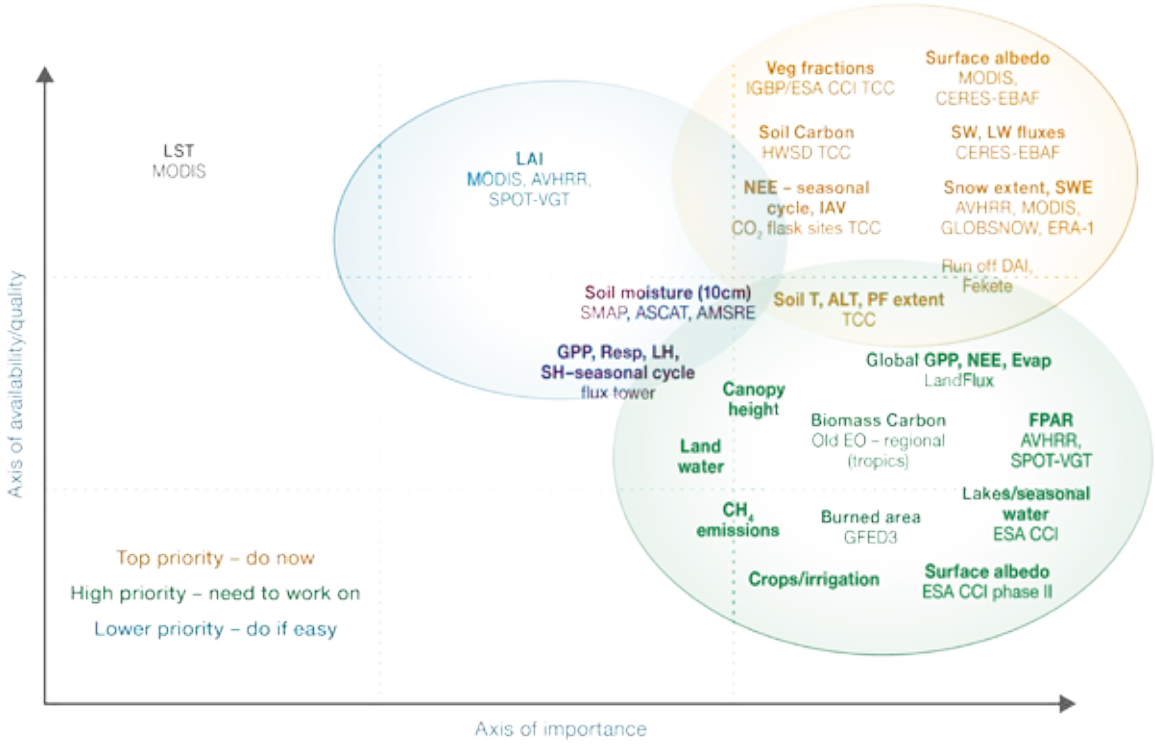


Figure 2.3. Ranking system employed by UKMO in determining land variables to incorporate into their metrics package. Adapted with permission from Martin Best and Chris Jones (UK Met Office).

gauging the relative importance of an observational constraint that had already been incorporated into the package. An important contribution from ILAMB may therefore be parsing the appropriate uses and limitations of various datasets that can be used for model evaluation.

The workflow through which model evaluation packages are employed suggests that another important contribution from a system like ILAMB is in facilitating comparisons for both coupled and uncoupled model runs. Most modeling centers reported that they develop their model sequentially, first focusing on uncoupled simulations, and later tuning for coupled simulations. A challenge for this sequential approach is that there are significant uncertainties in driver datasets that likely propagate to biases in land model output. Thus, ILAMB capabilities to evaluate both coupled and uncoupled runs is likely advantageous. Further development of functional response metrics would facilitate both types of comparisons.

A crucial component of benchmarking workflows is the ability to confront models with observational datasets that may reside in one or more data archives or repositories, and may evolve in time as new observations are added or as data processing methods are improved. Currently, this process is *ad hoc*, with modeling centers or individual scientists typically accessing a given dataset once, possibly converting its format to one that is most consistent with model output, and then storing the data locally for use in analysis. This process could be considerably streamlined through the development of an application programming interface (API) that allows benchmarking toolkits to rapidly and traceably access specific versions of datasets wherever they may be archived, align the data with model formatting requirements, and track whether updated dataset versions are available or new quality control issues with a given version of a dataset have been identified. This functionality is particularly important for model developers, like those in DOE's ACME project, who wish to track the evolution of model performance over time. If automatically downloaded observational data change without the user being informed, the fidelity of the model will appear to change even though no change was made to the model code or input data. We advocate for Federated data centers that support interoperable services as a means for solving the myriad of challenges associated with integrating observational data for model benchmarking (Williams et al., 2016). Obs4MIPS, a Federated archive built for data sets created or reprocessed specifically for use in comparison with model results, represents early work toward meeting these data management, versioning, and provenance challenges (Teixeira et al., 2014; Ferraro et al., 2015).

ILAMB promises to address barriers to sharing model evaluation packages across centers. A few modeling centers have already adopted ILAMB as a primary or secondary model evaluation package. Several centers desired better integration with other centers; however, a difficulty is that a diversity of software is used, including the NCAR Command Language (NCL), Ferret, Fortran, R, and Python. Thus, an open source evaluation system will likely facilitate cross-center interactions and drive community standards for model evaluation.

2.3 Synergies Across Different Benchmarking Activities

Several modeling groups have well-developed efforts focusing on land model assessment and benchmarking. These projects are all moving forward in parallel with ILAMB development. While some overlap exists across these projects, each package has a particular set of capabilities and strengths. PALS focuses on benchmarking in the true sense of the word by defining, through statistical models, an *a priori* expectation of minimum land model performance and assessing the prognostic models against that *a priori* expectation. LVT focuses primarily on water and energy cycling metrics and includes uncertainty and ensemble diagnostics, as well as more advanced statistical measures based on information theory, spatial similarity and scale decomposition techniques. ILAMB, on the other hand, emphasizes breadth through compilation and use of a comprehensive array of land datasets that cover a wide spectrum of terrestrial system processes and space and time scales. The ESMValTool and PCMDI Metrics Package provide mechanisms for routine analysis of coupled model output and include a set of diagnostics packages that collectively provide a comprehensive assessment of a wide range of essential climate variables—including some land variables—that are simulated by Earth system models. Each of these benchmarking efforts is serving unique as well as complementary purposes.

At this stage, coordination of these distinct and international land model benchmarking/assessment activities is challenging due to the diversity of approaches and the complexities of the international funding environment. Nonetheless, the 2016 ILAMB Workshop provided a good opportunity for everyone to share progress and ideas.

Over the longer term, it may be possible and beneficial to integrate existing land diagnostics packages under a loosely coordinated framework, potentially in a manner similar to that employed by ESMValTool for analysis of the coupled climate system. Under this scenario, the independently developed diagnostics packages (ILAMB, PALS, LVT) could be brought together under a single umbrella. Transitioning to this mode of operation would have the benefit of reducing effort related to the overhead of benchmarking (e.g., workflow processes such as reading in, processing, and reformatting model and observational data), which would allow more time, effort, and funding to be devoted to metrics development. One idea, as a first step toward a more coordinated international land model benchmarking activity, would be a joint benchmarking analysis project, wherein each of the existing packages is applied to a set of multi-model output that would enable direct comparison and evaluation of precisely how each package uniquely contributes to our understanding of model strengths and weaknesses.

3.0 Current Status of the ILAMB Software Packages

The complexity of today's process-rich Earth system models poses a verification challenge to developers implementing new parameterizations or tuning process representations, and a validation challenge to modelers for comprehensive and multifaceted evaluation of model predictions. Model developers and software engineers require a systematic means for evaluating changes in model results to ensure that their developments improve the fidelity of the target process representations while not adversely affecting results in other parts of the model. To objectively assess the performance of such models and identify model weaknesses—supporting the goals of the ILAMB project—a first-generation prototype benchmarking package and a second-generation package re-architected for better modularity and increased extensibility were developed. Called ILAMBv1 (Mu et al., 2016) and ILAMBv2 (Collier et al., 2016), respectively, both open source packages evaluate scientific model performance on 24 variables in four categories from about 45 data sets; produce graphical global-, regional-, and site-level diagnostics; and provide a hierarchical scoring system (Mu et al., in prep.).

At the previous ILAMB Workshop in the United States—held in Irvine, California, in January 2011—a methodology was developed for targeting aspects of model performance to be evaluated, identifying a set of benchmarks to test model performance, and guiding model improvements (Luo et al., 2012). Since that workshop, which advocated for near-term research efforts directed at developing a set of widely accepted benchmarks, the team of ILAMB developers and contributors have worked to design critical metrics for terrestrial model evaluation and to build software tools to evaluate those metrics and generate graphical diagnostics. Leveraging prior work on the Carbon-Land Model Intercomparison Project (C-LAMP; Randerson et al., 2009), the ILAMBv1 and ILAMBv2 packages were developed with support from the Biogeochemistry–Climate Feedbacks Scientific Focus Area (SFA) project (**https://www.bgc-feedbacks.org/**; Appendix F.6). ILAMBv1 is written in the National Center for Atmospheric Research (NCAR) Command Language (NCL) and was released as a prototype at the American Geophysical Union (AGU) Fall Meeting in 2015. ILAMBv2 is written in Python and was released at this workshop.

Both ILAMBv1 and ILAMBv2 assess model performance for variables in categories of biogeochemistry (aboveground live biomass, burned area, carbon dioxide, gross primary production, leaf area index, global net ecosystem carbon balance, net ecosystem exchange, ecosystem respiration, and soil carbon), hydrology (evapotranspiration, latent heat, and terrestrial water storage anomaly), radiation and energy (albedo, surface upward shortwave radiation, surface net shortwave radiation, surface upward longwave radiation, surface net longwave radiation, surface net radiation, and sensible heat), and climate forcing (surface air temperature, precipitation, surface relative humidity, surface downward shortwave radiation, and surface downward longwave radiation). For each of these variables, the packages generate graphical diagnostics and score model performance for the period mean over whole years and its bias (Figure 3.1),RMSE, spatial distribution, interannual coefficient of variation, and seasonal cycle and long-term trend (Figure 3.2). Variable-to-variable comparisons, or functional relationships, are also diagnosed to show how well models capture global or regional prognostic behavior in relation to one or more forcing variables (e.g., gross primary production vs. precipitation).

Model performance scores are calculated for each metric and variable and are further scaled based on the degree of certainty of the observational data set, the scale appropriateness and spatial and temporal coverage, and the overall importance of the constraint or process to model predictions. Scores are aggregated across metrics and data sets, producing a single scalar score for each variable for every model or model version. In ILAMBv2, these scores are also presented graphically to indicate absolute performance in stop-light colors and intra-model relative performance (Figure 3.3). Both graphical representations are useful because the absolute performance shows which variables are captured well by the models, while the relative performance or Z-score indicates which models or model versions are doing a relatively better or poorer job of reproducing the variable in question.

ILAMBv1 has been applied to analyze results from a suite of models that participated in the 5th phase of the Coupled Model Intercomparison Project (CMIP5) and new model development underway for ALM and the Community Land Model (CLM). ILAMBv2 is routinely used to study the evolving performance of both ALM and CLM. While ILAMBv1 is continuing to be used for individual studies, all new metrics development is expected to take place in the ILAMBv2 package because it runs in parallel across multiple compute nodes and is more modular, flexible, and extensible.

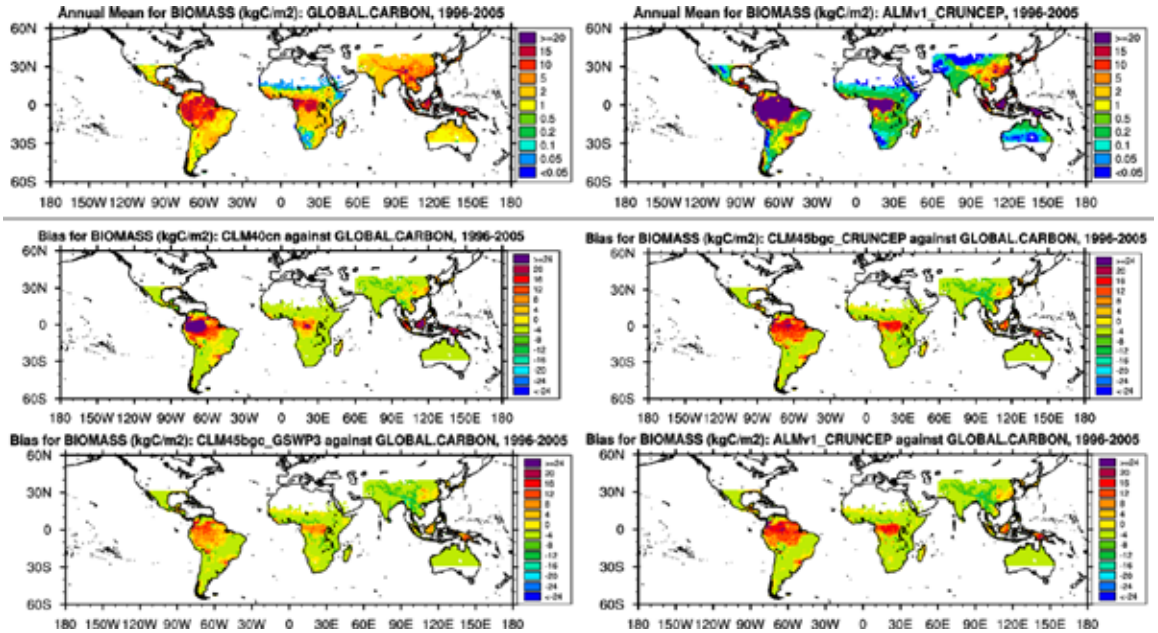


Figure 3.1. Shown here is the year 2000 pantropical forest biomass benchmark data (Saatchi et al., 2011) (top row left) and the Accelerate Climate Modeling for Energy (ACME) Land Model version 1 (ALMv1) annual mean biomass for years 1996 to 2005 (top row right). Below the horizontal line are maps of the bias from four models (CLM4.0-CN, CLM4.5-BGC, CLM4.5-BGC forced with GSWP3, and ALMv1). These biases are computed by subtracting the benchmark from the model annual mean biomass for years 1996 to 2005.

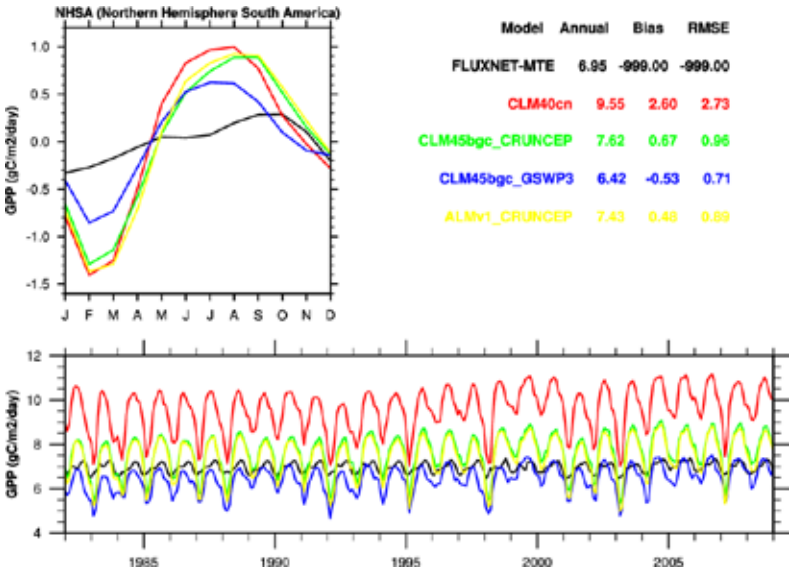


Figure 3.2. The ILAMBv1 prototype compares the model and FLUXNET (Lasslop et al., 2010) mean annual amplitude and phase of gross primary production (GPP) (top left); computes the annual mean, bias, and root-mean-square error (RMSE) of GPP (top right), and compares the full time series of GPP for prescribed regions.

Within US Department of Energy (DOE)-sponsored projects, the ILAMB framework is not only being leveraged by the ACME project but is bridging with large measurement and modeling projects in DOE's Terrestrial Ecosystem Science (TES) Program, including the Next Generation Ecosystem Experiments (NGEE) Arctic, NGEE Tropics, and Spruce and Peatland Responses Under Climatic and Environmental Change (SPRUCE). The ILAMB framework is developing and implementing metrics for new features of ALM, as a standard means for verifying model improvements. It is being adopted by TES projects to assist in development and testing of new process parameterizations, and as a mechanism for rapidly delivering observational data sets collected in the project to the modeling community. In addition, ILAMB is being used routinely to evaluate community-contributed enhancements to CLM within the Community Earth System Model (CESM) framework at NCAR. Both ACME and CESM are incorporating ILAMBv2 into their new workflow packages, so it will be run automatically as a standard post-processing step after executing a model simulation for rapid model development and assessment.

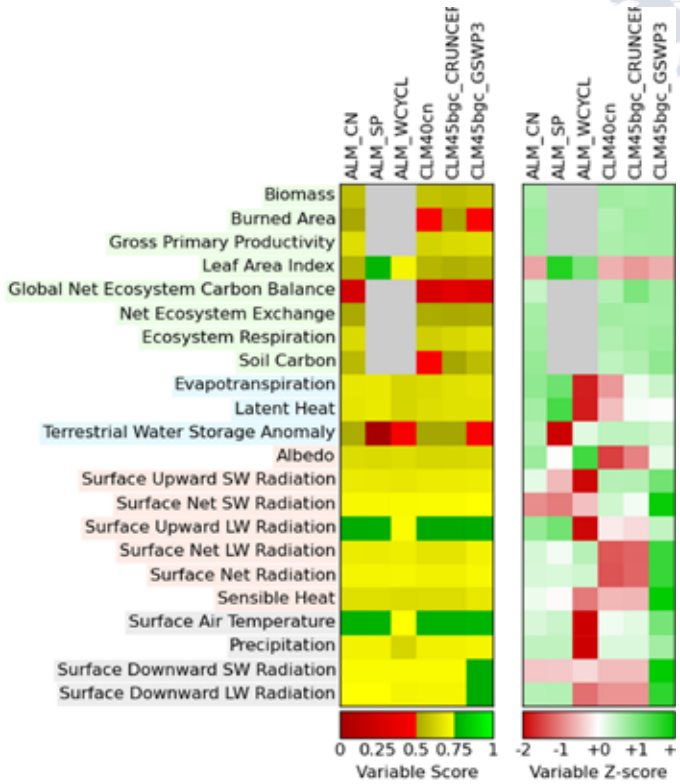


Figure 3.3. The ILAMBv2 package produces a summary graphic depicting model performance across a wide variety of variables, emphasizing absolute performance (left) as well as relative performance (right) with respect to comparisons with observations. This figure compares results from the ACME Land Model (ALM) run offline with carbon–nitrogen (CN) biogeochemistry (ALM_CN), run offline in satellite phenology (SP) mode (ALM_SP), and fully coupled in SP mode (ALM_WCYCL) with the Community Land Model (CLM) run offline for CLM-4.0 (CLM40cn), for CLM-4.5-BGC (CLM45bgc_CRUNCEP) and for CLM-4.5-BGC with Global Soil Wetness Project version 3 (GSWP3) forcing (CLM45bgc_GSWP3).

4.0 Next Generation Benchmarking Challenges

Maintaining and improving the scientific performance of today's complex Earth system models (ESMs) requires comprehensive, multifaceted, and systematic evaluation, analysis, and diagnosis of model results. A widening range of *in situ* measurements and remote sensing observations is available for use in judging the fidelity of land surface and terrestrial ecosystem models. A variety of statistical approaches have been adopted to evaluate model accuracy, including calculations of bias, root-mean-square error (RMSE), phase, amplitude, spatial distribution, Taylor diagrams and scores, functional relationship metrics, and perturbation and sensitivity tests. While many of these statistical measures are not independent, each provides slightly different information about contemporary model performance with respect to observational data and about implications for future projections from ESMs.

Developing metrics that make appropriate use of observational data remains a scientific challenge because of the spatial and temporal mismatch between models and measurements, poorly characterized uncertainties in observationally constrained data products, biases in reanalysis and forcing data, model simplifications, and structural and parametric uncertainties. The modeling community, in direct collaboration with the observation community, should develop clear guidelines on how these measures may best be used and how they complement each other for different benchmarking purposes. For example, functional relationships or variable-to-variable comparisons can partially compensate for errors in forcing data and provide information on ecosystem responses by comparing the relationships between two variables from models and observations, thus offering a zeroth order characterization of overall model behavior with reduced sensitivity to biases in atmospheric driver variables. A second example is the use of results from perturbation experiments, which can be used to probe specific process representations in the models.

This chapter outlines important challenges and benchmarking opportunities identified by the research community for assessing the performance of ESMs. At the workshop, a set of breakout group meetings was held on benchmarking major Earth system processes and another set focused on cross-cutting benchmarking themes. For this report, the summary of a separate plenary presentation and discussion about eddy covariance flux networks was added to the section on Integrating and Cross-cutting Themes. The breakout group meeting reports—contained in the Appendix—provide supporting details for the following benchmarking topics:

KEY RECOMMENDATIONS

- Developing metrics that make appropriate use of observational data remains a scientific challenge that should be addressed through synthesis activities in collaboration with the modeling and observational communities.
- Common benchmarking challenges highlighted the need for collocated measurements and uncertainty information, functional response metrics, emergent constraints, combining observational products, upscaling measurements, and collaborations with modeling and measurement communities.
- Develop "super site" benchmarks—integrated with AmeriFlux and FLUXNET—with detailed process-specific observations and robust model driving data to attribute model biases to underlying mechanisms.
- Create benchmarks for soil carbon turnover and the vertical distribution and transport of soil organic matter.
- Develop benchmark metrics for extreme event statistics, and on the response of ecosystems to extreme events.
- Synthesize data for vegetation recruitment, growth, mortality, and canopy structure, including disturbances, for benchmarking forthcoming demographic models.
- Develop a set of focused benchmarks for critical high latitude ecosystems, focusing on the dynamics of the coupled soil physical and biogeochemical system in permafrost-affected ecosystems.
- Create a set of focused benchmarks for tropical forest ecosystems, including observational targets for size-structured vegetation models, and coupled carbon–nitrogen–phosphorus cycle models.
- Leveraging efforts in observational projects, construct a roadmap and new methods for creating remote sensing data products for benchmarking models.
- Develop meta-analyses of perturbation experiments (e.g., nutrients, hydrology, temperature, CO_2) and related protocol for model comparisons.

Major Processes (Appendix B)

- ecosystem processes and states (Appendix B.1)
- hydrology (Appendix B.2)
- atmospheric CO_2 (Appendix B.2)
- soil carbon and nutrient biogeochemistry (Appendix B.4)
- surface fluxes (energy and carbon) (Appendix B.5)
- vegetation dynamics (Appendix B.6)

Integrating and Cross-cutting Themes (Appendix C)

- process-specific experiments (Appendix C.1)
- metrics from extreme events (Appendix C.2)
- design of new perturbation experiments (Appendix C.3)
- high latitude processes (Appendix C.4)
- tropical processes (Appendix C.5)
- remote sensing (Appendix C.6)
- eddy covariance flux networks (Appendix C.7)

The most important new metrics, benchmarking approaches, and observational data needs—distilled from the workshop breakout group meeting reports—are identified below. A number of common challenges and opportunities emerged from these reports, and they are described in the sidebar on *Common Benchmarking Challenges and Opportunities*. Workshops or sustained research working groups organized to address these topics could be conducted in the same fashion as working group meetings offered by national research synthesis centers in the US. Such workshops would bring together topical experts (e.g., modelers, ecologists, observationalists, remote sensing experts, mathematicians, and computer scientists) to make rapid research progress on the science topics identified in the two subsections below.

4.1 Major Processes

4.1.1 Carbon and Energy Fluxes

Surface fluxes of carbon and energy are key inputs from land to atmosphere models, and observations of these variables have been used to benchmark carbon cycle, land surface, and Earth system models for several decades. Routine observations of these fluxes come primarily from eddy covariance flux measurement tower sites. Networks of

COMMON BENCHMARKING CHALLENGES AND OPPORTUNITIES

A variety of common challenges and opportunities emerged from the individual breakout group meeting reports. Common themes focused on the following:

- need for collocated measurements, particularly around a core set of FLUXNET sites with a sustained record of observations for repeated model testing;
- lack of quantified uncertainty information for observational data;
- utility of functional response metrics and variable-to-variable comparisons;
- value of metrics for future projections based on emergent constraints;
- unrealized opportunities for global observational datasets based on satellite remote sensing synthesized with ancillary databases, using new algorithms;
- importance of applying statistical and machine learning methods to upscaling sparse measurements from sites to regions to the globe;
- need for process-level benchmarks and metrics for extreme events;
- opportunities for collaboration with Earth system model developers (e.g., ACME, CESM, and others); and
- opportunities for collaboration with important field and laboratory experiments and monitoring activities, including AmeriFlux and FLUXNET, the Integrated Carbon Observation System (ICOS), Next Generation Ecosystem Experiments (NGEE) Arctic, the Arctic–Boreal Vulnerability Experiment (ABoVE), the Spruce and Peatland Responses Under Climatic and Environmental Change (SPRUCE) project, Critical Zone Observatories (CZOs), Long-Term Ecological Research (LTER) sites, the National Ecological Observatory Network (NEON), NGEE Tropics, and the Tropical Responses to Altered Climate Experiment (TRACE).

these sites, such as AmeriFlux (**http://ameriflux.lbl.gov/**) and the FLUXNET (**https://fluxnet.ornl.gov/**) network-of-networks, have expanded rapidly over the last 25 years, and the data and meta-data they collect have been used in numerous model intercomparison and model–data comparison studies. Long term observations (>15 years) are available from an increasing number of sites, offering the opportunity to consider new studies of interannual to decadal variability, long term flux trends, ecological succession, multivariate climate response, and regional to global upscaling. While most of these sites are located in mid-latitude regions in North America and Europe, new sites are being deployed in the tropics, at high latitudes, and the undersampled Southern Hemisphere. The increasing density and widening spatial extent of sites, especially through organized and funded activities like ICOS (**http://www.icos-infrastructure.eu/**) and NEON (**http://www.neoninc.org/**), further enable studies of storm systems and convection, monsoons, and large scale extreme events, as well as providing significant improvements in estimates of regional and global gross primary production and ecosystem respiration.

Scaling flux observations to regions or the globe produces very important data products for constraining models. Machine learning techniques that account for nonlinearities, like artificial neural networks and model tree ensembles, have produced the most promising results, but provide limited explanatory information. The FLUXNET-MTE product (Beer et al., 2010), considered to be a best estimate of global GPP distribution, is widely used both for model evaluation—including within the existing ILAMB system (Ghimire et al., 2016)—and model tuning, suggesting the need for complementary approaches (e.g., Kumar et al., 2016). Current upscaling approaches do not incorporate information about disturbance, canopy structure, and other legacy effects (e.g., wildfire, effects of extreme events, insect infestation, disease, blowdowns). However, ancillary databases now contain observations of disturbance and detailed biological metadata that could be combined with flux observations to improve upscaled estimates or model predictions of surface fluxes.

These data and improved process representation in ESMs present opportunities for new synthesis activities directed toward carbon and energy benchmarking. Significant progress in improving process understanding and constraining models could be made through studies focused on the following:

- Multifactor ecosystem responses to climate change, extreme events, and changes in seasonality, which should integrate new phenocam observations (Brown et al., 2016), remote sensing products (Reed et al., 2009), data from the National Phenology Network (NPN; **https://www.usanpn.org/**; Schwartz et al., 2012), similar observations from citizen science programs (Fuccillo et al., 2015), and ancillary databases
- Roles of extreme events and "return times" on ecosystem resilience (Zscheischler et al., 2013)
- Long term trends in light use efficiency, water use efficiency, evapotranspiration, and other quantities, some of which may yield new emergent constraints
- Relationships between forcing and response variables (e.g., stand age and net ecosystem exchange; Noormets et al., 2007)
- Top-down approaches to constraining surface fluxes using vertical measurements of atmospheric CO_2 and other trace gases, and employing atmospheric inversion models (Xu et al., 2016)
- Synthesizing new observations from many data sets across space and time scales (e.g., FLUXNET, remote sensing, disturbance maps, etc.)
- "Super site" benchmarks developed around stable, long-running flux tower sites with a diversity of collocated measurements (e.g., AmeriFlux and FLUXNET, CZOs, LTER sites, or NEON sites)
- Upscaling point measurements, incorporating ancillary databases, to study areas, regions, continents, and the globe (Beer et al., 2010; Langford et al., 2016; Kumar et al., 2016)

A long-standing challenge to synthesis has been the reluctance of some researchers to share their eddy covariance flux data through openly distributed databases, like the FLUXNET2015 Dataset (**http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/**). While flux tower operators are increasingly convinced contributing their data is to their advantage, many researchers prefer direct involvement in synthesis working groups or workshops, which typically demonstrate the value of integrated analyses through high profile publications. Synthesis workshops would optimally involve modelers, flux tower operators, remote sensing observationalists, ecosystem ecologists, and mathematicians, and would be designed from the outset to produce original research papers and new synthesis or meta-analysis datasets for parameter optimization and model benchmarking. Additional details are contained in Appendixes B.5, C.7, B.3, B.1, and C.2.

4.1.2 Soil Carbon and Nutrient Biogeochemistry

Earth's soil holds roughly 2,000 Pg C, and soils have sequestered a significant fraction of CO_2 emissions from fossil fuel burning and human land use change since the start of the industrial era. However, under continued climate change and human intervention, soil carbon (C) is expected to have strong feedbacks to the atmosphere, shifting the balance to make soil a significant source instead of sink of carbon. The soil sequestration strength is determined by turnover rates, which are functions of plant inputs from litter and losses via microbial decomposition. Both of these mechanisms are regulated by nutrient availability. Understanding how the C balance may shift is limited because many key processes that regulate soil C stocks are poorly represented or missing in ESMs. The soil C stocks produced by current ESMs (CMIP5 models) are in only fair agreement with global soil C distributions, and the models are unable to reproduce local to regional scale spatial soil C patterns or to quantify bulk C stocks (Todd-Brown et al., 2013).

Traditionally, model evaluations have focused primarily on whether models can reproduce observed time series or spatial patterns in observational data (e.g., soil C stocks). Such benchmarks provide initial insights in model–data discrepancies, but offer limited insights into the sources of these differences. Benchmarks should be designed to test the representation of important controlling mechanisms (e.g., soil carbon age determined from isotope measurements; He et al., 2016) and environmental factors (Mishra and Riley, 2015), and benchmark datasets should include metadata to determine the appropriateness of comparisons and offer robust estimates of data uncertainties. To address challenges to ESM representation of soil C stocks and fluxes, scientific priorities for synthesis studies include the following:

- Developing improved benchmarks of soil C turnover through evaluation of soil nutrient biogeochemical processes, including (1) cycling of nitrogen (N) and phosphorus (P) and their interactions with ecosystem productivity and decomposition (e.g., Bouskill et al., 2014; Zaehle et al., 2014; Yang et al., 2016) and (2) competition for nutrients among microbes, plants, and mineral surfaces (Tang and Riley, 2013; Zhu et al., 2016)
- Representing the vertical distribution and transport (e.g., bioturbation and cryoturbation of soil organic matter (SOM), particularly at high latitudes, and synthesizing data on radiocarbon ages and C stocks to evaluate these parameterizations (Braakhekke et al., 2014; Koven et al., 2013; 2015; Riley et al., 2014; Tang et al., 2013; He et al., 2016)
- Evaluating models on their ability to simulate ecosystem responses to natural or anthropogenic disturbances and extreme events to highlight or expose processes critical to important phenomena
- Developing and applying emergent constraints based on carbon storage and turnover times to provide limits or bias corrections on future projections (Hoffman et al., 2014; He et al., 2016)
- Improving and harmonizing mapping and upscaling of global soil properties, especially for wetlands, tropical and boreal peatlands, and permafrost regions (Mishra et al., 2013; 2016; Mishra and Riley 2015)

Synthesis activities involving modelers, soil biogeochemists, microbial ecologists, and mathematicians could address the topics above. New collaborative research in these areas should focus on meta-analyses and developing new datasets useful for benchmarking models. Additional details are contained in Appendixes B.4, B.1, and C.2.

4.1.3 Hydrology

The key role of hydrology in land surface models (LSMs) is to partition incoming precipitation water into evapotranspiration, runoff (streamflow), and changes in soil moisture storage. These water cycle calculations are intrinsically tied to energy and carbon balance calculations. Soil moisture lies at the heart of land surface control over moisture fluxes, including both evapotranspiration and runoff. Hydrological processes operate across a range spatial and temporal scales, and LSMs in most ESMs attempt to approximate their effects using one-dimensional physics with varying degrees of complexity in the vertical direction. Groundwater formulations are restricted by the lack of lateral fluxes, surface reservoirs and impoundments are absent, and river and dam management is not considered in these models. Nevertheless, current process representations can be evaluated using a growing collection of *in situ* and remote sensing data. The greatest opportunities for improving water cycle benchmarking lie in synthesizing multiple datasets and developing metrics for related variables that indirectly constrain water fluxes. To improve hydrological evaluation of models and inform future model development, scientific priorities for hydrology benchmarking include the following:

- Benchmarking runoff and streamflow-related processes with Model Parameter Estimation Experiment (MOPEX; Duan et al., 2006) data for headwater watersheds in the US and Global Runoff Data Center (GRDC; Fekete et al., 2002) data globally
- Evaluating model performance in reproducing slow versus fast hydrological responses and capturing the impact of managed streamflow, including mapping of unmanaged watersheds
- Producing benchmark datasets for weather and climate extremes (WCEs), including shifts of the ITCZ and other circulation patterns, hydroclimatic intensity, flood inundation extent and duration, rainfall deficits, and experimentally induced extremes (e.g., throughfall exclusion and warming)
- Synthesizing many *in situ* soil moisture measurements from a wide collection of field activities with satellite remote sensing (e.g., SMOS, SMAP, ASCAT, GRACE) into a long-term dataset
- Developing a global-scale snow water equivalent (SWE) dataset
- Designing indirect benchmarking metrics for global-scale hydrology (e.g., estimate evapotranspiration from streamflow and diurnal temperature cycles from latent heat flux)

Synthesis studies involving modelers, hydrologists, ecohydrologists, and mathematicians could address the topics above. New collaborative research in these areas should focus on collecting and constructing new datasets, particularly for managed systems, and on developing new indirect metrics, particularly from remote sensing, for benchmarking models. Additional details are contained in Appendixes B.2, C.2, C.3, C.6, and B.1.

4.1.4 Vegetation Dynamics and Biomass

Vegetation dynamics refers to changes in ecosystem composition and structure through processes that include recruitment, succession, growth, mortality, and disturbance. In many LSMs, vegetation distribution is prescribed, making metrics of vegetation dynamics valuable only for testing model behavior of dynamic vegetation models (DVMs) that prognostically redistribute plants, or plant functional types (PFTs), across the landscape. In the last decade, vegetation demographic models (VDMs) have emerged that simulate light-competition driven coexistence and competition of PFTs through representation of varying tree size (e.g., cohorts or individuals) in the vertical canopy structure and successional dynamics through representation of disturbance history.

Over time, new data suggest that previous estimates of global vegetation biomass, both above and belowground combined, may be too high. Since most ESMs project higher global live biomass in the contemporary era than recent observations, the carbon storage potential in terrestrial vegetation and the turnover time of vegetation are in question (Negron-Juarez et al., 2015; Koven et al., 2015) . Many regional biomass products exist, but they tend to be limited to forests only or account only for aboveground live biomass. Additional studies that further constrain biomass inventories and how they evolve over time and respond to increasing atmospheric CO_2 are needed. To improve evaluation of vegetation dynamics in ESMs, particularly as VDMs become available, synthesis activities should address the following:

- Developing synthesis datasets for recruitment, mortality, and canopy structure from plot-scale measurements (e.g., Forest Inventory and Analysis (FIA); Johnson, Xu, McDowell et al., in prep.), AmeriFlux and FLUXNET, ForestPlots, ForestGEO, and national inventories for constraining models
- Comparing models with multiple burned area fire products, including GFED3, L3JRC, MCD45A1, Fire_cci, and the Global Fire Assimilation System
- Developing metrics based on multiple satellite remote sensing products for phenology, canopy height, and land cover to allow for characterization of uncertainties across classifications
- Creating metrics for vegetation responses to weather and climate extremes (WCEs), including disturbances from tornadoes and straight line winds, early/last frosts, hail streaks, and flooding
- Searching for emergent constraints based on organic carbon inventories and turnover times to provide limits or bias corrections on future projections (Hoffman et al., 2014)
- Developing benchmark datasets on repeated observations of remotely-sensed biomass to constrain biomass change over time (the most direct cumulative measure of carbon sink activity over time and a high priority to distinguish between different model predictions of the control of the terrestrial carbon sink)

- Participating in FireMIP to support new fire-related metrics and encouraging similar model intercomparisons for ecological networks like Drought-Net and Nutrient Network (NutNet)
- Developing maps of plant traits, land use change, disturbance, and mortality from wildfire, deforestation, drought stress, insects, and disease

Working groups involving modelers, ecosystem ecologists, foresters, and mathematicians could address the topics above. New collaborative research in these areas should focus on developing meta-analyses from widely dispersed field measurements to characterize recruitment, mortality, canopy structure, and biomass inventories, and developing metrics from remote sensing products for phenology, canopy height, and land use/land cover change. Additional details are contained in Appendixes B.6, C.2, C.6, B.5, C.7, and B.1

4.2 Integrating and Cross-cutting Themes

4.2.1 High Latitude Processes

Northern high latitude soils contain about twice as much carbon as in the atmosphere (Hugelius et al., 2014). This enormous carbon pool is vulnerable to accelerated losses through mobilization and decomposition under anticipated warming scenarios, with potentially large global carbon and climate impacts (Koven et al., 2011; Schaefer et al., 2011; Schuur et al., 2015). Many processes control the response of this carbon pool to changing environmental conditions. For example, active-layer dynamics, thermokarst formation, thermal erosion, shrub expansion, fire disturbance, soil moisture heterogeneity, and the overall rate of wetting and drying that will accompany warming. These processes impact the vulnerability of permafrost carbon pool through different mechanisms. Active layer thickness determines the volume of SOC available for microbial decomposition, and has been projected to go deeper under future warming. Thermokarst formation on the permafrost landscape enhances methane emissions to the atmosphere. Thermal erosion due to permafrost collapse can increase microbial decomposition and translocate large amounts of soil carbon to river networks. Increased wildfire occurrence has been projected under future warming scenarios; wildfires can directly combust the carbon in the surface organic layers and may alter the soil moisture dynamics. Similarly, many studies projected shrub expansion northwards under future warming, which can further destabilize the existing permafrost.

Because high latitude ecosystems are governed by extremely strong gradients in temperature and moisture, both vertically and horizontally, benchmarks must assess the coupled nature of biophysical and biogeochemical processes through variable-to-variable relationships in these regions (Harden et al., 2012; Koven et al., 2013; Bouskill et al., 2014). A wide variety of datasets are needed for next-generation benchmarking of ESMs at high latitudes, including maps of soil carbon that provide vertical profiles of carbon and isotopic age data, geographic distributions and dynamics of vegetation across boreal–tundra ecotone, relationships between snow properties and soil thermal dynamics, traits for vascular and nonvascular plants, and large-scale distributions of permafrost extent and active layer thickness. Research in DOE's NGEE Arctic project is directed at understanding the heterogeneity of polygonal tundra ecosystems, representing that heterogeneity in ESMs, and developing benchmarks to testing land models at high latitudes. To advance benchmarking of critically important processes with potentially large climate–carbon cycle feedbacks, collaborative research and synthesis activities should be focused on the following:

- Leading or strongly contributing to an independent research working group addressing synthesis of existing research and assessment of high latitude terrestrial processes affecting permafrost stability and feedbacks and developing potential emergent constraints in similar fashion to the Permafrost Carbon Network (PCN; **http://www.permafrostcarbon.org/**)
- Developing meta-analyses and synthesizing data to create high latitude benchmarks from *in situ* field measurements and experiments and remote sensing data in direct collaboration with researchers from DOE's NGEE Arctic, NASA's ABoVE (Xu et al., 2016), and NSF's Arctic science, observational, and monitoring projects
- Improving and harmonizing mapping of SOM and other soil properties in boreal peatlands and permafrost regions (Mishra et al., 2013; 2016)

- Developing and improving benchmarks of the coupled physical–biogeochemical dynamics of energy, moisture, nutrient, and carbon exchange across the permafrost–organic layer–snow–atmosphere system, and across heterogeneous landscape features that characterize patterned ground, to test models that increasingly represent the complex feedbacks that result from these coupled processes
- Applying statistical and machine learning methods to remote sensed and *in situ* data to understand the representativeness of measurements and intelligently scale sparse, difficult-to-obtain observations across the Arctic (Hoffman et al., 2013; Kumar et al., 2016)
- Implementing a model–data integration framework that addresses key indicators of high latitude ecosystem change as part of NASA's ABoVE program

Synthesis activities involving modelers, Arctic ecosystem ecologists, soil biogeochemists, hydrologists, and mathematicians could address the topics above. New collaborative research in these areas should focus on developing datasets and evaluating ESM fidelity for high latitude processes related to vegetation, soil biogeochemistry, and the physical snow-soil-hydrological system. In particular, functional relationships between biological, chemical, and physical variables and emergent characteristics (e.g., active layer thickness) should be examined to improve understanding of the process interactions and assess the credibility of model responses. Additional details are contained in Appendixes A.4, C.4, B.4, C.1, and C.6.

4.2.2 Tropical Processes

Tropical ecosystems present many processes that overlap with those in other biomes but also have additional complexity that makes modeling and benchmarking a distinct challenge from that of other regions. These include challenges related to high biodiversity, its representation in simulations, and its role in buffering ecosystem responses to perturbations. Advanced modeling and benchmarking have revealed challenges in representing carbon metabolism and the wide variety of above and belowground traits as they relate to water acquisition and use. Benchmarking has exposed these challenges through comparison to drought experiments and atmospheric constraints, with previous and current MIPs providing insights into the advantages and disadvantages of various numerical representations. While advances have been made, most work has pointed to the critical need for more extensive benchmarking of a range of processes at a range of scales, along with associated UQ and new model development.

Representing these processes is particularly crucial since tropical forests are predicted by the CMIP5 generation of ESMs to be particularly important for both the carbon–climate and carbon–concentration feedbacks. This importance led to the focus of the NGEE Tropics project to develop and synthesize key datasets required to test the representations of tropical forest dynamics in ESMs, as well as to develop and integrate into ESMs novel modeling approaches for representing these processes. To advance benchmarking of tropical ecosystem processes important to climate–carbon cycle feedbacks, collaborative research and synthesis activities should be focused on the following:

- Synthesizing spatially distributed inventories of size distributions, recruitment, growth, mortality, litterfall, and other ecosystem processes from the RAINFOR, CTFS-ForestGEO, AmeriFlux and FLUXNET, and GEM networks in direct collaboration with the NGEE Tropics project
- Collecting and developing benchmarking datasets for perturbation experiments and extremes in the tropics, including drought (e.g., Drought-Net), increased atmospheric CO_2 (e.g., Amazon FACE), nutrients (e.g., N, P), and increased temperature
- Modeling climate change to search for carbon cycle tipping points and possible emergent constraints associated with tropical ecosystems
- Taking advantage of naturally occurring events, (e.g., El Niño-induced tropical drought) to synthesize observational data for comparison with ecological forecast and retrospective modeling
- Combining inventory estimates, *in situ* process measurements, flux data, and remote sensing to characterize plant traits and physiological processes at larger scales and for regions with poor spatial coverage (e.g., western Amazon, tropical Africa, and Indo-Pacific) through statistical and machine learning upscaling methods

Research teams involving modelers, tropical ecosystem ecologists, soil biogeochemists, hydrologists, and mathematicians could address the topics above. New collaborative research in these areas should focus on developing improved inventory datasets and creating benchmarks for new demographic models for growth and mortality, tree

height and biomass, turnover of litter and stemwood, sap flow, tissue water potential and root water uptake, and nutrient constraints on carbon cycling. Additional details are contained in Appendixes C.5, B.6, B.5, C.7, B.2, C.2, B.1, and C.6.

4.2.3 Remote Sensing

The large extent and high diversity of vegetation comprising Earth's biomes present a significant challenge for local to global-scale terrestrial ecosystem process modeling efforts, including benchmarking and evaluation of model projections. To provide the knowledge and understanding necessary to improve model parameterizations, representation and evaluation of alternative model structures and observations are needed at the relevant spatial and temporal scales for controlling processes. The general goal of remote sensing from leaf to global scales is to provide critical information on ecosystem dynamics (e.g., seasonality, response to perturbations), and states (e.g., composition, structure, biomass), as well as to scale, map, and monitor important ecosystem properties and processes across space and through time. Compared with other observational and model evaluation datasets (e.g., inventory, eddy covariance, manipulation, and global change experiments), remote sensing data provide the synoptic, continuous, and temporally frequent observations needed for site to global model benchmarking. Moreover, the relative magnitude of remote sensing datasets of various types and temporal extents has helped to usher in the current data-rich era in ecology and global modeling, providing large volumes of information across scales that could be leveraged within data assimilation frameworks for model calibration and development activities.

Remote sensing observations and products useful for model evaluation span a fairly broad range of scales (temporally and spatially) as well as biophysical properties such as leaf area index (LAI) and the fraction of photosynthetically active radiation absorbed by vegetation (e.g., Myneni et al., 2002; Baret et al., 2007), states such as biomass (e.g., Saatchi et al., 2011), soil or canopy moisture (Petropoulos et al. 2015; Schimel et al., 2015), energy balance products such as surface albedo (Schaaf et al., 2002), to process-level observations, including evapotranspiration (Mu et al., 2011), photosynthesis (e.g., Running et al., 2004; Ryu et al., 2011; Guanter et al., 2014; Serbin et al., 2015), and plant functional traits (e.g., Asner et al., 2015; Singh et al., 2015). Calibration of algorithms for the retrieval of measurements using remote sensing observations vary in approach and complexity but generally require some degree of the physical relationship as well as independent information from ground or other observations for evaluation prior to any scientific or modeling use. In addition to other smaller campaigns, past and ongoing global change manipulations (e.g., DOE's FACE and SPRUCE), field experiments, and large-scale projects such as the DOE's NGEE Arctic and Tropics projects, as well as NASA's ABoVE, provide ample opportunities to refine remote sensing methods and products for use within ILAMB and elsewhere (Schmid et al., 2015). To accelerate and standardize the use of remote sensing for model benchmarking, collaborative research and synthesis activities should be focused on the following:

- Constructing a roadmap for remote sensing data product generation that takes into account enhanced cyberinfrastructure for large-scale remote sensing data (Williams et al., 2016) and new data product development for evaluation of process models from site to global scales (Schimel et al., 2015)
- Developing satellite simulators within ESMs that calculate an observable variable expected from remote sensing instruments under the given conditions
- Leveraging remote sensing efforts in DOE's NGEE Arctic, NGEE Tropics, and SPRUCE projects (and in collaboration with NASA's ABoVE and NSF's NEON projects) to develop and test algorithms for image processing, calibration, and uncertainty characterization, and to evaluate approaches for data retrieval and scaling
- Developing community guidelines for appropriate use of remote sensing data as benchmarks and observations for data assimilation
- Fusing data from multiple instruments (e.g., visible, TIR, LiDAR), data streams, or products for new synthetic observational datasets for hydrologic states and fluxes, carbon cycle fluxes, and vegetation trait and other properties

Remote sensing working groups involving modelers, ecosystem ecologists, geographers, remote sensing experts, and mathematicians could address the topics above. New collaborative research in these areas should focus on developing remote sensing products for plant traits, canopy structure, ecosystem responses to extreme events, solar-induced fluorescence, and carbon cycle fluxes (e.g., GPP, NPP, NEE). Additional details are contained in Appendixes C.6, C.1, C.3, C.4, C.5, B.2, B.6, and B.3.

4.2.4 Process-specific and Perturbation Experiments

To become more robust, Earth system models should undergo structural improvements to represent more real world processes (Knutti and Sedlacek, 2013; Luo et al., 2016). Given the enormous complexity of Earth system processes, it is still challenging to (1) specify which processes are more critical than others in regulating Earth system dynamics, such as climate change; and (2) evaluate representation of processes that have been widely incorporated but diversely parameterized in different models. One promising approach to solving this challenge is using process-specific experiments, which can evaluate and improve the model representation of a specific key process through comparison with observations. Key processes to target, for which models are highly parameterized or have major structural uncertainties, include decomposition, nitrogen cycling, autotrophic respiration, chlorophyll fluorescence, phenological sensitivity to climate, and plant trait correlations and trade-offs.

Direct perturbation of environmental properties is one of the most direct ways of assessing ecosystem responses to environmental change. Such experiments—which include perturbation of nutrients, species composition, precipitation, temperature, atmospheric chemistry, CO_2 concentration, or multiples of these factors—have been conducted across a wide range of experimental systems. In some cases, the resulting datasets have been synthesized and are ready for model benchmarking, while others require effort to synthesize and standardize reporting of results. Care is required to avoid scale mismatches and most effectively apply an analog to the experimental perturbation within models. In addition, the mechanistic response of the ecosystem to the perturbation must be understood, so that models exhibiting the correct response for the wrong reason can be recognized. Performance of model runs early in the process of defining an experimental perturbation may be useful in identifying specific processes and assumptions on which models disagree, and they may inform data collection strategies to be most relevant to model benchmarking efforts (Medlyn et al., 2016).

To advance process-level benchmarking of ecosystem models, collaborative research and synthesis activities should be focused on the following:

- Selecting a core set of AmeriFlux and FLUXNET sites that span major biomes to serve as long-term testbeds for ILAMB, collecting all associated data and metadata (e.g., meteorological forcing, soil texture, land use history, and plant traits) necessary for conducting model simulations, and constructing or synthesizing a series of independent benchmark datasets (e.g., net fluxes, biometrics, and experimental data) for diagnosis of model process representations
- Collaborating with DOE's SPRUCE project to collect data and synthesize benchmark datasets for diagnosis of model responses to prescribed perturbations for a northern peatland
- Collaborating with DOE's NGEE Arctic project (i.e., small-scale warming and isotopic tracers) to collect data and synthesize benchmark datasets
- Collaborating with DOE's LBNL TES soil perturbation project to collect data and synthesize benchmark datasets for soil organic matter responses to temperature and moisture
- Synthesize existing nutrient (e.g., Bouskill et al., 2014; Zhu et al., 2016), temperature, and moisture perturbation experiments with meta-analyses appropriate for model benchmarking, and concurrently developing guidance for performing relevant model analyses
- Opportunistically using measurements during weather and climate extremes in lieu of perturbation experiments to develop benchmarks for vegetation and soil biogeochemical responses
- Incorporating the FACE Synthesis (Zaehle et al., 2014) protocol and data into ILAMB in collaboration with original synthesis participants
- Collaboration with TRACE, ITEX, and other soil warming experiment teams to develop modeling protocols, collect forcing data, and synthesis results for benchmarking

Synthesis activities involving modelers, ecosystem ecologists, field and laboratory experimentalists, remote sensing experts, and mathematicians could address the topics above. New collaborative research in these areas should focus on developing simulation protocols, forcing datasets that correspond to the observed meteorology and any perturbation applied, and data for benchmarking ecosystem responses. Additional details are contained in Appendixes C.1, C.3, B.5, C.7, B.4, and C.2.

5.0 Model Intercomparison Projects (MIPs)

5.1 The Roles of Benchmarking in MIPs

Model Intercomparison Projects (MIPs) are important activities for assessing the coherence and reliability of Earth system models. By adopting a common set of protocols with clearly defined inputs and outputs, model predictions can be compared systematically to each other and benchmarked with observations. A number of ongoing and future MIPs are directly relevant to the modeling of the terrestrial water, energy, and carbon cycles, and many of these were discussed at the ILAMB Workshop. Some are conducted under the auspices of the 6th phase of the Coupled Model Intercomparison Project (CMIP6) project, while others are separate activities. The goal of this section is to summarize these MIPs, their different scientific objectives, protocol designs, and the opportunities for land model benchmarking that each presents.

5.2 Descriptions of MIPs and Their Benchmarking Needs

5.2.1 CMIP6 Historical and DECK

As part of the CMIP6 process, each participating model will conduct a set of runs called the Diagnostic, Evaluation, and Characterization of Klima (DECK) experiments (Eyring et al., 2016b). These simulations comprise four experiments: a land–atmosphere only model forced by reconstructed historical sea surface temperatures (i.e., Atmospheric Model Intercomparison Project (AMIP)), a coupled land–atmosphere–ocean preindustrial control, an abrupt quadrupling of CO_2, and an idealized 1% per year CO_2 increase. Furthermore, each model will perform a set of historical simulations with the coupled atmosphere–ocean–land models. For the preindustrial control and historical simulations, models with active carbon cycles will run these with both prescribed atmospheric CO_2 concentrations and prescribed emissions, and this offers a key opportunity to test the ability of the models to predict the evolution of atmospheric CO_2 over the historical period (Hoffman et al., 2014). Furthermore, large-scale dynamics of model-predicted historical climate variables may be compared with corresponding observations from *in situ* and remote sensing methods.

5.2.2 C^4MIP

To isolate carbon feedbacks in the Earth system, the Coupled Climate–Carbon Cycle MIP (C^4MIP) (Jones et al., 2016) will separately force the coupled land–atmosphere–ocean system with CO_2 that acts only on plant-physiological and ocean-solubility processes, and separately only on radiative processes. This allows separating the carbon–concentration feedback, which acts to stabilize the climate system, from the carbon–climate feedback, which acts to destabilize the climate system. Furthermore, fully-coupled future ESM experiments are included, in which CO_2 emissions rather than concentrations are used to force the model and CO_2 is allowed to evolve in time. Previous versions of the C^4MIP experiments (Friedlingstein et al., 2006; 2014a) demonstrated a poor ability of ESMs to agree

KEY RECOMMENDATIONS

- Develop methods to attribute emergent model behaviors such as carbon feedback parameters to specific processes through emergent constraint and traceability approaches.
- Benchmark across coupling and complexity hierarchies—from offline land-only simulations to fully coupled ESMs—to attribute model biases and uncertainties to specific domains and identify feedbacks between domains.
- Develop paired site datasets for benchmarking mode representations of subgrid scale heterogeneity.

on the basic trajectory of terrestrial carbon budgets in response to global change, and this lack of agreement has provided a strong impetus for better benchmarking and validating terrestrial carbon cycle models over the historical period to constrain future trajectories. Furthermore, the CMIP6 iteration of C^4MIP has identified key uncertainties that were poorly represented in prior generation ESMs, including nutrient cycles, permafrost-related processes, and the use of carbon isotopes as a possible diagnostic tool for reducing uncertainty in carbon cycle processes.

5.2.3 LS3MIP

The Land Surface, Snow and Soil Moisture Model Intercomparison Program (LS3MIP) (van den Hurk et al., 2016) contains a series of coupled and offline experiments to isolate the roles of terrestrial energy, water, and carbon cycles in leading to inter-model differences and biases. Included in the LS3MIP protocol are a series of offline experiments, in which the land models will be forced with a common set of meteorological drivers: Tier 1 experiments will be driven by GSWP3 (Kim et al., in preparation); Tier 2 experiments will also include WATCH (Weedon et al., 2011), CRU-NCEP (Viovy and Ciais, 2011), and Princeton (Sheffield et al., 2006) drivers. This will allow both the separation of terrestrial model performance from atmospheric model performance and the role of the uncertainty of historical meteorology on land model performance. In addition, LS3MIP experiments include a set of future land-only time-slice simulations driven by common model-produced meteorology; and prescribed land-surface experiments, following the GLACE protocols (Koster et al., 2004; Seneviratne et al., 2013) for evaluating land-surface feedbacks to climate. Crucial to all of these experiments is accurate knowledge of the soil moisture and snow fields, and moisture and energy fluxes, for diagnosing biases in the land-only model experiments and accurate prescriptions of the land-surface fields in the prescribed land-surface experiments.

5.2.4 LUMIP

The Land Use Model Intercomparison Project (LUMIP) is focused on understanding the complex roles of land use and land cover change (LULCC) as forcing agents in the Earth system. LUMIP includes a series of experiments to better identify and attribute physical and biogeochemical effects of land use, including offline and coupled experiments that are performed with and without land-use change, and a detailed reporting specification of subgrid land model states and fluxes in other CMIP6 experimental runs. Key to benchmarking land use effects are paired observations subject to the similar meteorology but including different land uses and histories, and comparison of these paired sites with sub-gridscale information on land surface heterogeneity due to land use within land models.

5.2.5 MsTMIP

The Multi-scale Synthesis & Terrestrial Model Intercomparison Project (MsTMIP) is designed to evaluate land model skill as driven by common meteorology, spinup, land surface, and other drivers. Experiments include a sequentially-added forcing design, including drivers of climate, CO_2 concentrations, land cover, and nitrogen deposition. MsTMIP is not a CMIP6 project and thus includes participation of models that are run only offline. Phase 1 MsTMIP experiments were focused on the historical period, and Phase 2 consists of future experiments. Phase 2 of MsTMIP will employ a novel computational infrastructure, the JPL "model farm," in which all of the participating models are run on a single machine to ensure that they are treated identically with respect to inputs, outputs, and protocols.

5.2.6 PLUME-MIP

Processes Linked to Uncertainties Modelling Ecosystems (PLUME-MIP) also uses a set of offline climate-driven land models to attribute changes in modeled carbon cycle responses to global change to its underlying drivers. The novel aspect of this MIP is the use of a recently developed traceability framework (Xia et al., 2013) to disaggregate the differences between models into underlying drivers, such as changes in productivity and changes in turnover of various pools (Ahlström et al., 2015; Koven et al., 2015). To accomplish this disaggregation, a new set of model diagnostics is required, in particular to diagnose changes to turnover times under simultaneously changing inputs and model pool transfer rates (Rasmussen et al., 2016).

5.3 New Metrics, Approaches, and Model Output Requirements

A variety of benchmarking metrics approaches have been integrated into the first version of ILAMB to allow testing of models at multiple spatial, temporal, and complexity scales. These include: (1) site-level comparison of water and energy fluxes between model gridcells and flux towers; (2) global- and regional-scale comparison of gridded data products from remote sensing, point-based upscaling, or data assimilation approaches with corresponding fields from offline and coupled land models; (3) comparison of Earth system-integrative measures such as atmospheric CO_2 fields between observations and models.

These multiscale approaches are useful for covering the broad range of scales that encompass observational networks and over which the relevant processes represented in ESMs operate. However, model configurations used in the first generation of ILAMB span only three configurations: (1) offline global model runs forced by bias-corrected historical reanalysis data and historical land use data; (2) coupled global atmosphere–ocean–land model runs forced by time-varying land use and trace gas concentrations; and (3) coupled global atmosphere–ocean–land model runs forced by time-varying land use and fossil fuel emissions, with CO_2 transport either predicted by the atmospheric model or calculated from an offline transport model. Only gridcell-mean properties were tested, and site-level data was based on extracting individual gridcells from global runs.

The larger diversity of model couplings and experimental protocols in the current and upcoming generation of MIPs suggests that a more comprehensive strategy is needed for both model–data benchmarkings and model–model comparisons to best utilize the information in these MIPS. Benchmarking approaches require a high degree of correspondence between the periods of observation and model scenarios, and the ability to benchmark models is always contingent on the fidelity with which the inputs required to simulate the observations correspond to reality; however, this correspondence may span a wide diversity of coupling complexity. Possible couplings include (1) site-level comparisons where the model is driven by site-level observations; (2) offline global models forced by a variety of meteorology datasets; (3) prescribed land-surface experiments as in LS3MIP, where certain land states are initialized to observations in a coupled land–atmosphere framework; (4) AMIP runs where atmospheric model uncertainty is added but sea surface temperatures (SSTs) are constrained to historical dynamics; (5) fully physically coupled runs with atmospheric and ocean model dynamics present; and (6) physically and biogeochemically coupled simulations. Each of these experimental configurations allows potentially different comparisons between models and datasets to be made to benchmark the ESMs across both a wide range of variables and a scale of complexity in Earth system components. To effectively leverage these different MIPs and allow benchmarking approaches to span these complexity hierarchies, it would be ideal to develop within ILAMB the capability to span across different coupling strategies to track which aspects of a given ESM are benchmarked by different comparisons and assign metrics that take a system-centered view of ESMs.

New models outputs will be required for effectively using many of these MIP activities as benchmarking tools. Among the new outputs are model subgrid information, as specified in the LUMIP protocol. This will enable benchmarking with consideration that site-level observations correspond only to a subset of a model gridcell, and of LULCC-related heterogeneity in ESMs. Further, whereas benchmarking with observational datasets can only occur for model variables that correspond directly to observable quantities, non-observable model outputs, such as turnover times and disequilibrium fluxes as identified through a traceability framework, may still be of great use in understanding and diagnosing model behaviors. Furthermore, better instrumenting models to output quantities such as isotopic pools and fluxes, or ecosystem structural information such as tree size distributions (which allow deeper model introspection and process-resolved benchmarking), will be crucial to test increasingly complicated ESMs.

5.4 Available Observations and Data Gaps

ILAMB as it is currently built is able to use a wide variety of global-scale and regional-scale gridded observations, site-specific observations, and integrative observations. Increased use of each of these types of observations would allow a more robust model benchmarking framework. For offline models and MIPS, key required observations are better meteorological driving datasets for the models. These include both global-scale bias-corrected reanalysis products and

site-scale driving data to allow better comparisons of models with site-scale data. Furthermore, driving data of other anthropogenic forcings, such as LULCC, nutrient deposition, aerosol effects, and other processes, have considerable uncertainty that propagates through models and complicates the interpretation of model-benchmark differences. Observations of subgrid scale heterogeneity, for example through the use of remote sensing approaches and paired site-scale observations, will enable better testing of subgrid scale approaches in models, which is crucial as models evolve to have numerous sources of heterogeneity, including topographic position and land-use histories. Moving beyond the mean gridcell value of a given variable will require observations that maintain the full distribution of a given property across a gridcell-sized domain rather than just reporting mean values at the scale of gridcells.

5.5 Expected results from MIPs and ILAMB

The key goal of benchmarking activities is to reduce the uncertainty associated with directly testable model predictions. Although there will always remain an irreducible uncertainty arising from issues such as equifinality, uncertainty in future drivers, and uncertainty in current observations, there is currently a wide divergence in model predictions for things that can be directly and robustly observed that is contributing to the poor predictive capability of terrestrial models (e.g., Hoffman et al., 2014). New MIP activities, and the associated benchmarking opportunities that they represent, offer promise that we as a community can build models that are far more constrained by observations such that the remaining uncertainty will be due to genuinely unknown rather than simply untested processes.

6.0 Model Development and Evaluation Testbeds

Land surface model components of ESMs are experiencing dramatic changes as new process representations are added and software infrastructures are altered to support more detailed demographic and plant trait formulations. Moreover, alternative parameterizations for major submodel components (e.g., soil biogeochemistry) are being introduced into land models to test competing model structures and parameterizations at different spatial and temporal scales. To support this degree of rapid model development, a land model testbed (LMT) capability is needed for calibration and evaluation of process-level submodels at site, regional, and global scales. A well-designed LMT would provide infrastructure similar to that of today's model farms for executing models (e.g., the JPL Model Farm described in Section 5.2.6 and the PEcAn framework described in Appendix E.4), but provide many more features for rigorous benchmarking at varying degrees of model complexity. There is a risk that model development adds parameters and complexity that do nothing to reduce model error and bias. This risk can be overcome by consistently testing simple models against data, and determining the information content provided by more complex parameterizations (Li et al., 2014) facilitated by a LMT deployed on supercomputing computational resources.

One of the key findings of this report is the need to select a core set of AmeriFlux and FLUXNET sites spanning major biomes to serve as the "gold standard" targets of long-term testbeds for ILAMB. A LMT should contain the collection of all associated data and metadata (e.g., meteorological forcing, soil texture, land use history, and plant traits) necessary for conducting model simulations, and have encoded the series of independent benchmark datasets (e.g., net fluxes, biometrics, and experimental data) for diagnosis of model process representations. Efforts to improve the code modularity in ALM and CLM are positioning those models to be able to take advantage of a well-crafted LMT, which must have access to individual submodels and simple input/output mechanisms for exchange of data not typically saved in model history files. For example, residence times for all pools, allocation and turnover of foliage, microbial pool dynamics, respiration of all living pools, trait correlations, N (including biological fixation) and P dynamics are needed for a detailed analysis. This biogeochemical data can then be used to evaluate model dynamics across pools and timescales (Thomas et al., 2013). A LMT could be incorporated into existing automated nightly or weekly model testing to add scientific functionality testing to routine compile, runtime, and restart testing.

In an effort to consider how a LMT may be useful for supporting rapid development of the ACME Land Model (ALM), a table of evaluation variables and benchmark datasets was organized. Table 6.1 contains this sample list of variables and corresponding datasets designed to prioritize incorporation and synthesis of observational data for evaluating the ALM. For each dataset, the citation and data source are listed (when available), and a decision was made about whether the data would be useful as model input or for evaluation or both. Datasets presently available for use are listed as "Ready" and those requiring collection, processing, and synthesis are listed as "Synthesis". As ILAMB is expanded, a database recording the provenance of data should be created and used to track the capabilities of the ILAMB package, and such a database could be part of the supporting infrastructure provided by a LMT.

KEY RECOMMENDATIONS

- Design and build a land model testbed (LMT) for execution, calibration, and evaluation of alternative model formulations and process representations to support rapid model development and testing.
- An initial LMT should be designed around a small number of AmeriFlux and FLUXNET "super sites" for which single point simulations can be executed and evaluated quickly in parallel.
- Eventually a LMT capability should be incorporated into routine model development testing.

Table 6.1. Listed here are example datasets identified for benchmarking the ACME Land Model.

Data Set	Reference	Source	Input or Evaluation	Ready or Synthesis
Soil Nutrients and Age				
Hedley P database	Yang and Post (2011)	http://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1223	Input	Ready
Global soil respiration database	Bond-Lamberty and Thomson (2010)	https://github.com/bpbond/srdb	Evaluation	Ready
Microbial P database	Xu et al. (2013); Hartman et al (2013)		Evaluation	Ready
Vertical soil P profile			Input	Synthesis
Radiocarbon database	He et al. (2016)		Evaluation	Ready
Soil nitrification, denitrification	Ojima et al. (2000)	https://www.nrel.colostate.edu/projects/tragnet	Evaluation	Ready
Soil N deposition and leaching	Suddick and Davidson (2012)		Evaluation	Ready
Sorption-desorption for P by soil order	Agriculture literature		Evaluation	Synthesis
Vegetation Measurements				
Leaf N and P	Kattge et al. (2011)	TRY database	Evaluation	Ready
Fine root N and P	Yuan et al. (2011); Gordon and Jackson (2000)		Evaluation	Ready
Carbon stocks (MgC/ha) tree, root, CWD/dead wood		Forest Carbon Database (CiFOR)	Evaluation	Ready
Fire (burned area)		GFED3 (annual mean, seasonal cycle, interannual variability)	Evaluation	Ready
Wood harvest		Hurtt (annual mean)	Input	Ready
Land cover		MODIS PFT fraction	Input	Ready
Live biomass	Global: Saatchi et al. (2011); Amazonia: Malhi et al. (2006)		Evaluation	Ready
Vegetation Demography				
Demographic data (DBH census, basal area, abundance, species name)	http://ctfs.arnarb.harvard.edu/Public/pdfs/Condit_1998_Census PlotsmethodsBook.pdf	ForestGEO	Input and Evaluation	Synthesis
Basal area by diameter class	http://ctfs.arnarb.harvard.edu/Public/pdfs/Condit_1998_Census PlotsmethodsBook.pdf	ForestGEO, LTER, BOREAS, INPA	Evaluation	Synthesis

Data Set	Reference	Source	Input or Evaluation	Ready or Synthesis
Basal area by wood density class	http://ctfs.arnarb.harvard.edu/Public/pdfs/Condit_1998_Census PlotsmethodsBook.pdf	ForestGEO, LTER, BOREAS, INPA	Evaluation	Synthesis
Basal area by leaf N content	http://ctfs.arnarb.harvard.edu/Public/pdfs/Condit_1998_Census PlotsmethodsBook.pdf	ForestGEO, LTER, BOREAS, INPA	Evaluation	Synthesis
Seasonal LAI	http://ctfs.arnarb.harvard.edu/Public/pdfs/Condit_1998_Census PlotsmethodsBook.pdf	ForestGEO, LTER, BOREAS, INPA	Evaluation	Synthesis
Mean mortality rate (with modes of death captured in RAINFOR database)	http://ctfs.arnarb.harvard.edu/Public/pdfs/Condit_1998_Census PlotsmethodsBook.pdf	ForestGEO, LTER, RAINFOR	Evaluation	Synthesis
Disturbance and mortality	Midrexler et al. (2009)	MODIS Global Disturbance Index (MGDI)	Input and Evaluation	Synthesis
Hydrology				
Soil moisture		De Juer, SMAP	Evaluation	Synthesis
Water storage anomaly		GRACE	Evaluation	Ready
River flow/runoff		Syed/Famiglietti, GRDC, Dai, GFDL, GSCD	Evaluation	Ready
River temperature			Evaluation	Synthesis
Snow cover		AVHRR, GlobSnow	Evaluation	Ready
Snow depth		CMC (North America)	Evaluation	Ready
Snow water equivalent	North America: Ghan et al (2006)	National Operational Hydrologic Remote Sensing Center	Evaluation	Ready
Surface Energy Budget				
Surface skin temperature		MODIS LST, GOES LST	Evaluation	Ready
NLDAS-2 surface air temperature, downward SW and LW	CONUS: Cosgrove et al. (2003)	http://ldas.gsfc.nasa.gov/index.php	Evaluation	Ready
CRU surface air temperature	Harris et al. (20013)	http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_1256223773328276	Evaluation	Ready
Net radiation, LE, H	Lasslop et al. (2010)	Fluxnet	Evaluation	Ready
Albedo		MODIS, CERES	Evaluation	Ready
Radiative fluxes		CERES, SURFRAD, ARM	Evaluation	Ready

Data Set	Reference	Source	Input or Evaluation	Ready or Synthesis
CERES surface SW, LW, and net radiation	Kato et al. (2013)	http://ceres.larc.nasa.gov/order_data.php	Evaluation	Ready
WRMC BSRC surface SW, LW, and net radiation	Konig-Langl et al. (203)		Evaluation	Ready
Carbon Fluxes				
Gross primary production	Lasslop et al. (2010); Jung et al. (2010)	FLUXNET; MPI-BGC MTE product	Evaluation	Ready
Net ecosystem exchange	Lasslop et al. (2010)	FLUXNET	Evaluation	Ready
Litterfall, Litter Content, Litter Decomposition				
Litterfall and litter carbon and nutrients	Holland et al. (2014)		Evaluation	Ready
Litterfall	http://www.ctfs.si.edu/data///documents/Litter_Protocol_20100317.pdf	ForestGEO	Evaluation	Synthesis
LIDET for N	Parton et al. (2007)	http://andrewsforest.oregonstate.edu/research/intersite/lidet.htm	Evaluation	Ready
CIDET for N and P			Evaluation	Ready
Tropical litter decomposition	Manzoni et al. (2010)		Evaluation	Synthesis
Functional Responses				
NPP vs. N availability	Thomas et al. (2010)		Evaluation	Ready
NPP vs. P availability	Quesada et al. (2012); Aragão et al. (2009)		Evaluation	Ready
Aboveground biomass C vs. aboveground NPP	Keeling and Phillips (2007)		Evaluation	Ready
Manipulation Experiments				
FACE synthesis for NPP	Zaehle et al. (2014)		Evaluation	Ready
Ecosystem fertilization	LeBauer and Treseder (2008); Elser et al. (2007); Wright et al. (2014); Vitousek et al. (2004)		Evaluation	Ready
Decomposition	McGroddy et al. (2004)		Evaluation	Ready
Soil incubation	Cleveland and Townsend (2006)		Evaluation	Ready
Soil warming	Rustad et al. (2000); Melillo et al. (2011, 2002)		Evaluation	Ready
CO_2 effect on phosphatase			Evaluation	Synthesis
EucFAC, Amazon FACE			Evaluation	Synthesis
Tropical warming			Evaluation	Synthesis

7.0 Traceability and Uncertainty Quantification Frameworks

In order to understand and explore the uncertainty around predictions made by terrestrial models, it is crucial to improve methods and datasets to quantify the structural and parametric sources of this uncertainty. Two key developments are required to do this: (1) the development of reduced order models to simplify and systematize the relationships within full models, and (2) development of UQ approaches to understand parametric and structural uncertainty. One such reduced-order approach is the traceability framework of Luo and collaborators, which seeks to define a common matrix structure underpinning carbon cycle models, such that both structural and parametric uncertainty can be explored via the values of turnover times, carbon flows, and the correlation structure between these. Other UQ approaches include the identification of parametric uncertainty that most strongly affects model outcomes, so as to focus research efforts on defining these, as well as methods to calibrate model parameters and discriminate between alternate model structures.

7.1 Traceability Framework

To evaluate model fidelity and understand the sources of uncertainty that lie behind carbon cycle projections, the modeling community needs to develop better observational benchmarks of model performance, which has been the focus of ILAMB and related efforts. A key requirement for increased understanding is the ability to tie specific biases in model predictions to underlying process representations. One way to do so is through the development of diagnostic approaches that simplify and generalize model structures into component parts. A promising approach is to consider the carbon cycle at a given location as a cascade in which carbon enters the ecosystem only through leaf-level photosynthesis, and then is transferred from the leaves into the other tissues that comprise a plant, which ultimately grows, dies, and decays in the soil. This common framework allows one to generalize carbon cycle models

KEY RECOMMENDATIONS

- Integrate and report model diagnostics that allow the emulation of carbon cycle models as a matrix of carbon flows and turnover times, in order to attribute uncertainty in carbon responses to specific ecosystem components.
- Apply Bayesian UQ approaches that efficiently utilize leadership-class computing facilities to quantitatively identify uncertainties in LSM output.
- Use UQ results to guide data collection activities and target critical model improvement activities, including new or revised process representations.
- Improve the fidelity of emulators and their use in UQ methods.
- Emphasize the role of inverse modeling and data assimilation to update both model parameters and states as part of Bayesian UQ strategies, and as such, the importance of observational data with associated uncertainties.
- Standardize collection and distribution of observational data. Standardization implies a common data format as well as metadata such as measurement errors and procedures used to compute them. If the data have gaps which were filled in/imputed with a model, provide the model or, at a minimum, the uncertainty bars in the imputed data.
- Incorporate more trait, remote sensing, and other data to provide constraints on model parameter distributions and to enable evaluation of model constraints given existing data sources.
- Suggest a simple and clear web-based graphical user interface (GUI) that provides access to models, UQ, and ILAMB benchmarking tools to facilitate a broader adoption of the approaches and to allow non-modelers but process/domain experts to conduct UQ and data assimilation experiments.
- Leverage several UQ frameworks that have important and complementary tools. Use an improved cyberinfrastructure to link these tools within a broader community-wide model UQ and data integration framework focused on improved land surface model (LSM)/terrestrial biosphere model (TBM) projections.

into a common structure, which can be well represented by the matrix equation (Luo et al., 2003; Luo and Weng, 2011; Luo et al., 2015, 2016; Sierra et al., 2015) as:

$$X'(t) = B\,u(t) - A\,\xi(t)\,K\,X(t), \qquad (1)$$

where $X'(t)$ is a vector of net carbon (C) pool changes at time t, $X(t)$ is a vector of pool sizes, B is a vector of partitioning coefficients from C input to each of the eight pools, $u(t)$ is C input rate, A is a matrix of transfer coefficients to quantify C transfer along the pathways, K is a diagonal matrix of exit rates (mortality for plant pools and decomposition coefficients of litter and soil pools) from donor pools, and $\xi(t)$ is a diagonal matrix of environmental scalars to represent responses of C cycle to changes in temperature, moisture, litter quality, nutrients, and soil texture. The equation describes net C pool change, as a result of C input, distributed to different plant pools via partitioning coefficients, minus C loss through transfer of C, in individual pools.

Overall, this equation can conceptually express all of the soil C transformation processes and summarize structures of classic SOC models, such as the CENTURY (Parton et al., 1987, 1988, 1993) and RothC models (Jenkinson et al., 1987), and—despite the fact the various ESMs may differ in many parameters and processes that determine the terms in this equation—embedded in ESMs (Ciais et al., 2013). Thousands of datasets published in the literature from litter decomposition and soil incubation studies have been used to obtain first-order decay parameters that can be used in ESMs (Zhang et al., 2008; Schädel et al., 2013, 2014). The scalar function, $\xi(t)$, in Equation 1 represents the environmental modifier for decomposition and transfer rates with respect to changes in temperature, moisture, litter quality, and soil texture. Empirical studies have also indicated that temperature, moisture, litter quality, and soil texture are primary factors that control soil C decomposition and stabilization (Burke et al., 1989; Adair et al., 2008; Zhang et al., 2008; Xu et al., 2012).

Equation 1 not only summarizes most of the land carbon cycle models embedded in most of the Earth system models (ESMs) but also contains several mutually independent components. Traceability analysis decomposes the complex terrestrial C cycle into a few traceable components (Xia et al., 2013, 2015a). Traceability analysis helps identify sources of uncertainty in modeled steady-state ecosystem carbon storage due to (1) C input as affected by phenology, physiology, and C use efficiency (Xia et al., 2015a); (2) edaphic and vegetation characteristics as related to baseline C residence time; (3) climate scalars; and (4) environmental variables among models (Figure 7.1). The traceability framework has been applied to assess influences of external variables being represented as parameters, boundary conditions, and diagnostic variables in models so as to disentangle complex representations of external variables in influencing simulated C dynamics in ESMs (Xia et al., 2013).

As an example of how the traceability approach may lead to greater understanding of model behavior, Rafique et al. (2016) applied the traceability framework to two global land models (CABLE and CLM-CASA′) to diagnose causes of their differences in simulating ecosystem carbon storage capacity. Driven with similar forcing data, the

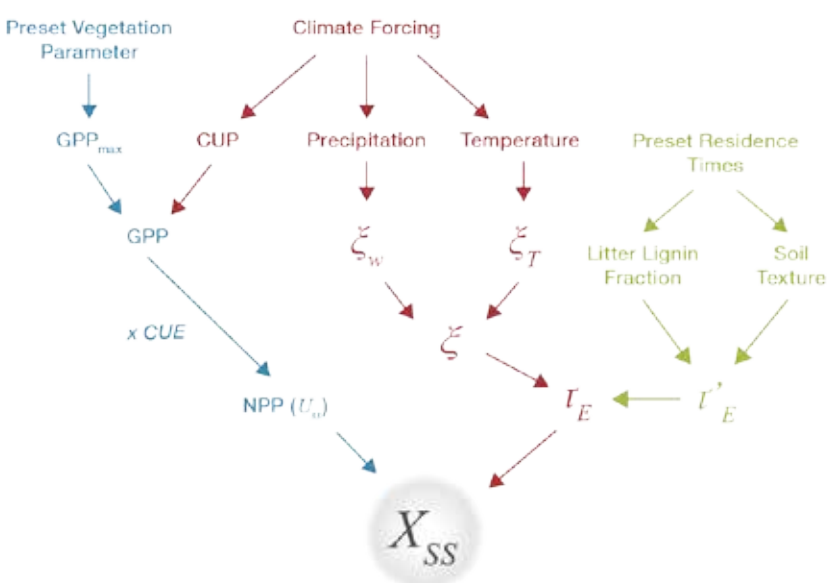


Figure 7.1. Schematic diagram of the traceability framework. The framework traces modeled ecosystem C storage capacity (*Xss*) to a product of net primary productivity (NPP) and ecosystem residence time (τ_E). The latter τ_E can be further traced to (i) baseline C residence times (τ'_E), which are usually present in a model according to vegetation characteristics and soil types, (ii) environmental scalars (ξ), including temperature and water scalars, and (iii) environmental forcing. NPP can be traced to C use efficiency (CUE), C uptake period and the seasonal maximum of gross primary productivity (GPP). Adopted from Xia et al. (2013, 2015).

CLM-CASA′ model predicts ~31% larger carbon storage capacity than the CABLE model. Since ecosystem carbon storage capacity is a product of net primary productivity (NPP) and ecosystem residence time (τ_E), the predicted difference in the storage capacity between the two models results from differences in either NPP or τ_E or both. This analysis showed that CLM-CASA′ simulates 37% higher NPP than CABLE does because of the parameter setting that gives CLM-CASA′ higher rates of carboxylation (V_{cmax}) than CABLE. On the other hand, ecosystem residence time (τ_E) was longer in CABLE than CLM-CASA′. Because τ_E is determined by baseline carbon residence time (τ'_E) and environmental scalars, the difference in τ_E is caused by both longer τ'_E and stronger temperature limitation of soil carbon decomposition (i.e., smaller temperature scalar) in CABLE than CLM-CASA′. The longer τ'_E in CABLE was mainly determined by its longer τ'_E of woody biomass and higher proportion of NPP allocated to woody biomass than CLM-CASA′. Comparatively, environmental scalars have relatively smaller influences than NPP and τ'_E in causing differences in predicted carbon storage capacity between the two models. Overall, the traceability framework offers an effective approach to identify sources of uncertainty among models.

One key issue going forward is that a variety of current and emerging model structures have fundamentally nonlinear dynamics, which may be less conducive to the approach of constructing linear emulators. In particular, both the vegetation components, through the development of cohort-based models (e.g., Moorcroft et al., 2001; Weng et al., 2015), and the soil components, through the development of microbial models (e.g., Wieder et al., 2015c; Sulman et al., 2014) have fundamentally different dynamics because the turnover times become an emergent, nonlinear property that must be diagnosed rather than one that can be calculated from the model. The applicability of the traceability method on this class of models remains a key unresolved question to be explored.

Model intercomparison projects (MIPs) all illustrate great spreads in projected land C sink dynamics across models (Todd-Brown et al., 2013; Tian et al., 2015). It has been extremely challenging to attribute the uncertainty to sources. For example, the CMIP5 protocol did not allow the calculation of all terms required to perform this traceability analysis. Nonetheless, using the available output does allow a first-order separation of the dominant terms of productivity and turnover, which shows an interesting pattern: inter-model spread in the initial carbon stocks was primarily driven by differences in turnover times, whereas inter-model spread in transient changes was mostly due to changes in productivity (Koven et al., 2015). Placing simulation results of a variety of C cycle models within one common parameter space can measure how much the model differences are in common metrics. The measured differences can be further attributed to sources in model structure, parameter, and forcing fields with traceability analysis (Xia et al., 2013; Rafique et al., 2016; Ahlström et al., 2015; Chen et al., 2016). The traceability analysis also can be used to evaluate the effectiveness of newly incorporated modules into existing models, such as adding the N module on simulated C dynamics (Xia et al., 2013).

It will be fruitful to explore how new techniques stemming from the global analysis, such as physical emulators (i.e., matrix expression of global carbon cycle models) and traceability, can enhance benchmark analysis. Furthermore, such emulators may be of use in uncertainty quantification efforts, as the reduced order form of the traceability framework may allow for both computationally-efficient process-based emulators that can be run over large numbers of ensembles, as well as efficient ways of finding steady-state initial conditions to full models when varying parameters for UQ methods.

7.2 Scientific Driver for UQ of LSM

Quantifying the uncertainty in model outputs due to parameters, initial conditions, or model drivers is crucial to robust benchmarking efforts. In particular, inverse modeling and uncertainty propagation are two areas of UQ that should be integrated into the ILAMB framework. LSMs typically contain many parameters and drivers that must first be constrained to obtain meaningful benchmark results. Parameter tuning, part of a larger framework of model–data integration, uses observational data and expert knowledge to identify appropriate model parameters. The use of Bayesian approaches to inverse modeling or calibration (otherwise known as parameter data assimilation, PDA) will further allow determine statistical descriptions of model parameters and potentially reduce parametric uncertainty bounds. Using the improved quantitative description of the parametric uncertainties within an uncertainty propagation analysis then produces meaningful statistical descriptions of the benchmarked metrics for a particular LSM. In particular, it is possible to quantify the robustness of a particular LSM in the presence of model uncertainties.

Moreover, model–data integration activities also include state-variable data assimilation (SDA). In contrast to model calibration or PDA, SDA focuses on updating model states by comparing a model forecast to an observation of that state which serves to move the model closer to the observation weighted by the uncertainties in both the model and data. Following the SDA step, the best estimate of the state of the system is used as the prior for the next model forecast, and the uncertainty in the model projection is reduced based on the confidence in the data and model projection. SDA is particularly useful for capturing processes and perturbations that may not be explicitly captured by a model (e.g., windthrow) and serves to move the model toward the observation, which in turn updates associated model states to better reflect the observed state. SDA together with PDA can be used to test model predictive capacity and evaluate alternative model process representations.

Advanced UQ tools are also important in other aspects of benchmarking, including sensitivity analysis and model diagnostics, especially when the number of model parameters is increasing in tandem with model complexity. Model UQ and variance decomposition can be used to guide data collection and model improvement activities based on the decomposed variance of a particular model forecast. By ordering the dominant drivers of model uncertainty in projections of carbon, water, and energy fluxes and storage model UQ and variance decomposition approaches help to focus on the high-priority model needs first. In addition, UQ activities within ILAMB should be conducted regularly and iteratively to identify model improvements based on previous UQ results and re-prioritize critical new foci based on the latest updated results. For example, UQ can help identify a critical observational need which then reduces the uncertainty of the model. The next UQ cycle would then identify a new area of focus, given that the previous priority is now sufficiently constrained. Finally, it is critical that UQ tools provide both univariate and multivariate approaches to evaluate the covariance among model parameters and drivers.

Applications of UQ techniques are typically constrained by the computational cost of an LSM. At present, advanced UQ techniques, such as Monte Carlo (MC) based methods, can only be used with site-specific models that are computationally inexpensive. At regional and global scales, only scenario-based UQ analyses are computationally tractable. Scenarios are, however, too sparse to draw rigorous conclusions and support decisions with quantified risk/uncertainty bounds. Efficient linear approximation techniques (e.g., maximum likelihood estimation with Gaussian assumption) are often not very useful because LSM responses of perturbed-parameter studies are strongly nonlinear. With the availability of spatially-distributed observational data (e.g., global remote sensing data, Appendix C.6) there is an increasing need to apply advanced UQ techniques at the regional and global scales that can also leverage diverse datasets. This requires new approaches that can approximate the full results using dimensionality reduction, which could be based on climate, vegetation, topographic, or other clustering approaches.

In addition, there have emerged many recent advances in Markov Chain Monte Carlo (MCMC) methods and particle-based MC methods that we can explore and customize to LSMs. Some new efficient methods include implicit particle filter (Chorin and Tu, 2009); stochastic Newton MCMC method (Martin et al., 2012); and MCMC methods that utilize Gibbs samplers (Kuczera et al., 2010), differential evolution samplers (Laloy and Vrugt, 2012), affine invariant ensemble samplers (Goodman and Weare, 2010), and surrogate-based samplers (Goodwin, 2015; Ray et al., 2015). These methods have varying degrees of parallelism that affect their efficient deployments on leadership-class supercomputing facilities. It is unlikely that one method will be suitable under all benchmarking scenarios and for all LSMs. There is thus a need to identify how the various methods can be applied efficiently under the different benchmarking scenarios that will be encountered in regional ILAMB UQ activities.

The number of parameters in an LSM can be large, and this poses another UQ challenge. However, chosen benchmarking metrics are usually impacted only by a small subset of the parameters and drivers that are used within LSMs. Dimensionality reduction can be achieved by identifying a reduced set of salient or relevant contributors through a sensitivity analysis (SA). Global SA requires large perturbed-parameter ensembles (especially for high-dimensional global SA), and the challenge lies in computational resources, data bookkeeping, provenance, and a cyberinfrastructure capable of managing the distributed resources. This can be partly addressed by using sparse grid methods; e.g., Smolyak grids and Curtis-Clenshaw quadratures. However, their use is not widespread in the LSM community.

Emulators or surrogate models are fast-running proxies of LSMs that can be used in studies where LSMs need to be invoked repeatedly (e.g., parameter or data assimilation). Emulators are typically constructed through statistical methods (e.g., Gaussian process regression, and polynomial chaos expansion), machine learning approaches (e.g. random forests, support vector machine regression, deep neural networks, and gradient boosting machines), and model reduction techniques (proper orthogonal decomposition method, reduced basis method, and discrete empirical

interpolation method). The use of emulators (typically generated through large ensemble simulations of the full LSM) can help to reduce overall computational costs of large-scale UQ and benchmarking, in particular for regional-scale LSMs. However, emulators must first be trained using outputs from large ensemble simulations of the full LSM. The number of simulations required is typically reduced by utilizing efficient space-filling sampling techniques (e.g., Latin hypercubes, and sparse collocation method) or constraining the parameter space through a global SA or dimensional reduction algorithms. While the use of emulators in UQ analysis is promising, there are several challenges that must first be addressed. First, LSM responses may be too complex to allow accurate emulators to be built. While there are many successful attempts at constructing emulators for scalar responses, methods for field responses (as one might expect in regional LSM runs) are not well studied. Second, approximation errors inherent in emulators need to be accounted for in the UQ methods. Alternatively, we can attempt to identify an optimal pairing of emulator and MC method that will achieve the desired improvement in accuracy and efficiency. For example, the implicit particle filter efficiently constrains the effective parameter space, allowing accurate emulators to be efficiently built with fewer samples. Finally, streamlined construction of emulators is a necessity for practical UQ and large-scale data assimilation, which are hampered by the complexity of LSM structures and responses.

7.3 Observational Data Needs

As a community we have entered into a data-rich era with numerous observational datasets collected at site to regional and global scales (Luo et al., 2011). These include leaf-level datasets, inventory data, tower observations, and remote sensing. However, in many cases these data are not easily available, well documented, web-accessible, standardized, provided with error estimates, or stored in an archival format. Many key datasets are "long tail" data found in student theses, hard copy, researcher hard drives, or other sources that are challenging to bring forward to the benchmarking and modeling community. New technologies and open-science initiatives are quickly eliminating these access problems with contemporary data, but a general investment in improved data retrieval and standardization is needed regardless of the observation of interest. In particular, proper documentation of datasets is critical for the appropriate use of any observation and to avoid erroneous benchmarking or comparisons. Moreover, data standardization is critical, and knowledge of data collection, instrumentation, post-processing, etc., is necessary to provide comparable data from multiple sources with uncertainties. Web-accessible tools should be prioritized to foster transparency such that the larger community can utilize and evaluate these synthetic observations, which serve to iteratively improve the datasets and curate standard products used across research groups, thereby serving to enhance reproducibility and direct comparisons across scales.

The following is a list of specific data requirements for maximizing the use of observations in model uncertainty quantification efforts:

- Collaborating with DOE's SPRUCE project to collect data and synthesize benchmark datasets for diagnosis of model responses to prescribed perturbations for a northern peatland
- Include estimates of measurement errors in any data that is distributed. This should also mention distribution of the errors.
- Access to scripts/codes for gap-filling, or generate gap-filled data and documentation of the gap-filling algorithm.
- Metadata: how it was collected (instrument), how the measurement error estimates were computed (assumptions, etc.), what missing data has been filled in, and how that was done, etc.
- Include data and associated metadata in the same file/package.
- Standardized, documented, and web-accessible meteorology driver data available at multiple temporal resolutions able to drive the models within ILAMB
- Web-accessible orbital and suborbital remote sensing datasets useful for model evaluation, calibration, and benchmarking, including LiDAR, microwave, hyperspectral, and thermal (Appendix C.6; Schmid et al., 2015). This includes new fusion products designed to test model outputs and functional responses within a UQ framework
- "Sensor simulator" to provide direct comparison between internal model structure and canopy radiative transfer and low-level observations from suborbital and orbital remote sensing platforms. By comparing direct observations (i.e., surface reflectance) as well as derived products (e.g., LAI), the uncertainty in model structure can be evaluated and can as well as facilitate direct data assimilation to improve model fidelity.

7.4 Algorithm Needs

The main algorithms needed can be classified into five categories: sensitivity analysis algorithms, inverse modeling algorithms, data assimilation algorithms, Monte Carlo-based algorithms, and training algorithms for emulators. There are potential overlaps in these categories. Existing packages for these algorithms exist in R (e.g., abc), Python (e.g., Sciki-Learn [**https://www.scikitlearn.org/stable/**], pyMC), MATLAB (e.g., UQLab [**https://www.uqlab.com/**]), and C++ (e.g., DAKOTA, UQTK).

Most of the scripting languages contain packages that implement different deterministic and Bayesian calibration methods. Bayesian calibration develops estimates of LSM parameters as probability density functions (PDF); they are usually much narrower than the bounds that constitute prior beliefs regarding their values. Many new Bayesian methods are implemented in R and Python. When Gaussian assumptions regarding the PDF are acceptable, scalable ensemble Kalman filters (e.g., OpenDA [**https://www.openda.org/**]) are routinely used. However, if distributional assumptions are not to be imposed, Markov chain Monte Carlo (MCMC) and particle filters (PF) are required to solve the calibration problem. New MCMC-based algorithms are available in Python, for example the Differential Evolutionary Monte Carlo method (in spotpy), and the affine invariant Monte Carlo method (in emcee). Approximate methods for Bayesian calibration (e.g., Approximate Bayesian Computation [Csilléry et al., 2010]) that could employ LSMs (not emulators) are available in R (Csilléry et al., 2012). Bayesian calibration of LSMs is still in its infancy; the thrust seems to be in assessing whether parameter-estimates-as-PDFs confer much benefit in terms of predictive skill.

Approaches for constructing emulators through statistical, regression and machine learning techniques exist mostly in R and Python (e.g., Scikit-Learn). DAKOTA (**https://dakota.sandia.gov/**) and UQTK (**https://www.sandia.gov/UQToolkit**); however, Karhunen-Loeve (KL) approximations of multivariate Gaussian random fields are potentially suitable for field-scale emulation, although it is unclear how the large eigensolves required for KL decompositions of regional LSM runs can be efficiently performed by serial UQ software. Random field models for non-stationary random fields (e.g., wavelet based) are not supported by any UQ package. Despite the availability of multiple, well-implemented packages, there is currently no framework that allows streamlined construction of emulators that take into account the complexity of LSM structures and responses. To make advances in the application of UQ techniques for LSM, the following priorities should be pursued:

- Access to scalable packages for EnKF, MCMC, approximate Bayesian computation
- Automatic packages for constructing surrogate models based on Gaussian process, neural nets, deep learning, random forests, support vector machine regression, and non-parametric methods
- New parsimonious parameterizations for spatially variable fields; e.g., flux, permeability, and sparsity-enforcing inference methods such as Bayesian compressive sensing
- A connected cyberinfrastructure to link multiple existing tools, frameworks, and approaches within ILAMB to provide synthetic workflows that provide advanced UQ and assimilation algorithms and approaches

7.5 Computational, Visualization, and Data Analysis Needs

Perturbed parameter ensembles result in large datasets, and UQ is assisted substantially by a combination of physical intuition (i.e., expert knowledge) and data patterns observed in the ensembles. Exploratory data analyses of the ensembles is a necessity for efficient UQ analyses, but existing visualizations tools for large data (e.g., Ensight, Paraview) are geared toward interrogation of individual datasets, not ensembles. Large data analysis tools such as Spark [Spark] can script/automate much of the preliminary data processing required in exploratory data analysis but lack any visualization capabilities. A scripting and visualization capability such as R, Python, or Matlab, but customized to ensemble analysis, would be helpful. The following objectives should be considered to overcome computational and visualization challenges:

- Parallelization of LSMs: task-based parallelization of LSM, distributing each site or ensemble member on each core of a graphics processing unit (GPU) to speed calculations

- Data analysis and visualization: possibility to combine big-data analysis software (e.g., Spark) with visualization capabilities (e.g., like the statistical scripting language R) to enable detailed diagnostic figures. In addition, packages such as R-Shiny provide interactive data wrangling and plotting for big data. Possibility of having Big Data analytics clusters to be co-located with HPC platforms?
- Web-accessible GUI to run models and model UQ tools within ILAMB. This will facilitate more direct connection between modeler, measurers, and domain experts.
- Leverage existing tools for interactive data analysis (e.g., R-Shiny) to improve interaction and real-time analysis of model benchmarking results. Provide web-accessible tools for analysis and visualization capable of generating publication-ready graphics

8.0 Computational Needs and Requirements

Comprehensive analysis of ESM output at increasing resolutions is already challenging the computational infrastructure commonly used by modelers and analysts. As observational data sets continue to grow in temporal length and spatial resolution, data storage and processing capacities will limit their use in model benchmarking without appropriate investments in data management and computational infrastructure. Scalable algorithms and machine learning techniques should be developed for evaluating and benchmarking high resolution and long time series ESM results.

Combining integrating, and synthesizing data across Earth science disciplines offers new opportunities for scientific discovery that are only starting to be realized (Hoffman et al., 2011). The rise of data-intensive scientific pursuits, in Earth sciences and other disciplines, has led some visionaries to proclaim it the fourth paradigm of discovery alongside the traditional experimental, theoretical, and computational archetypes (Hey et al., 2009). The promise of scientific advances in predictive understanding of environmental change has stimulated an enormous increase in the volume of both model and observational data. ESM simulations, especially for community modeling activities like CMIP, can generate tens of terabytes to several petabytes of output in raw form (Overpeck et al., 2011). Satellite remote sensing data tend to be very large and their size has grown as spatial and temporal resolutions have increased; however, small ecological data sets, often the most useful for synthesis, may be the most difficult to preserve, distribute, and use (Reichman et al., 2011). Research organizations must address these data collection, curation, archiving, discovery, and distribution challenges, and plans for creating a Virtual Laboratory infrastructure promise solutions that could enable new knowledge discovery (Williams et al., 2016).

Today's large and complex Earth science data often cannot be synthesized and analyzed using traditional methods or on individual workstations. As a result, data mining, machine learning, and high performance visualization approaches are increasingly filling this void and can often be deployed only on parallel clusters or supercomputers (Hoffman et al., 2011). However, supercomputer architectures designed for compute-intensive simulations, usually containing large numbers of cores with high speed interconnects between nodes, are not typically optimal for large scale analytics. Instead, such applications demand large and fast on-node memory, high bandwidth input/output (I/O), and fast access to large local disk volumes. To realize the promise of new scientific discovery from very large, long time series Earth science data, a distinct balance of increasing computational, storage, and bandwidth capacity from high performance computing resources is required. Scientific computing enterprises should be advised to strike the right balance of these resources for their application communities as they plan their expansion to exascale computing (Lucas et al., 2014).

As described above, UQ presents significant computational challenges that lead to development of reduced complexity and surrogate models that may fail to reproduce model behavior in unpredictable ways. Methods that can exploit leadership-class computing should be developed to address these challenges. Facilities supporting large scale data management and server-side manipulation and computation (e.g., Google Earth Engine) will become increasingly important as growing data volumes eliminate the possibility of transporting data to a researcher's site for analysis. Data assimilation, *in situ* visualization, and benchmarking should function independent of the locations of

KEY RECOMMENDATIONS

- Scalable algorithms and machine learning techniques should be developed for evaluating and benchmarking high resolution and long time series ESM results.
- Research organizations should develop cyber infrastructure to support large scale data collection, curation, archival, discovery, and distribution, and it should support automated model–data comparisons and online data assimilation for parameter estimation through supercomputing facilities.
- Scientific computing facilities should strike a balance between resources for compute-intensive vs. data-intensive applications as they plan their expansion to exascale computing.
- New development for ILAMB should include improved support for remote retrieval and version tracking for observational data.

the data streams or observational data products needed to drive the simulation or evaluate its results. Realizing this vision requires investment in both cyber infrastructure for simulations and data storage and retrieval (e.g., obs4MIPs) and the software components of models and benchmarking packages. New development for ILAMB should include improved support for remote retrieval and version tracking for observational data.

9.0 Conclusions and Next Steps

ADVANCING BENCHMARKING SCIENCE

- A combination of small, targeted working groups, and larger, but less frequent meetings with the full community can increase visibility, participation, and science impact of ILAMB over the next several years.
- Supporting the 6th Phase of the Coupled Model Intercomparison Project (CMIP6) is one of the most critical ILAMB goals for the next 3–4 years.
- In the next 10 years, the community needs a synthesis center that will lower the barrier to information flow between measurement and modeling communities, with ILAMB serving as a core capability.

9.1 Workshop Conclusions

The May 2016 ILAMB Workshop was very successful in bringing the international community together to identify scientific challenges and priorities for future research. The workshop demonstrated that there is a vibrant community of scientists, spanning many disciplines, who are committed to reducing barriers for information flow between the measurement and modeling communities. The integration of ILAMB packages into the workflow of several major modeling centers highlights the growing importance of this effort for the science of Earth system prediction.

A variety of **Benchmarking Approaches** have been adopted to evaluate model accuracy through comparison with observations, including the following:

- Statistical comparisons (bias, root-mean-square error (RMSE), phase, amplitude, spatial distribution, Taylor diagrams and scores)
- Functional relationship metrics or variable-to-variable comparisons
- Emergent constraints
- Reduced complexity models and traceability analyses
- Formal uncertainty quantification (UQ) methods
- Meta-analyses of perturbation and sensitivity experiments.

While many of these statistical measures are not independent, each provides slightly different information about contemporary model performance with respect to observational data and about implications for future projections from ESMs. *Reduced complexity models, traceability analysis, and UQ methods* could be combined into useful frameworks to achieve the following goals:

- Integrate and report carbon cycle model diagnostics as a matrix of flows and turnover times to attribute responses to specific ecosystem components
- Apply Bayesian UQ approaches that utilize leadership-class computing facilities to quantify model uncertainties
- Employ UQ results to guide data collection activities and target process representations needing improvement
- Investigate integration of emerging UQ frameworks with future ILAMB package releases.
- *Developing metrics* that make appropriate use of observational data *remains a scientific challenge* because of the following:
- Spatial and temporal mismatch between models and measurements
- Poorly characterized uncertainties in observational data products
- Biases in reanalysis and forcing data

- Model simplifications
- Structural and parametric uncertainties.

In the near-term, an important step will be to target specific areas within the fields of ecosystem ecology and hydrology for synthesis and further detailed ILAMB metrics development. Recommendations identified for next-generation **Benchmarking Challenges and Priorities** included the following:

- Develop supersite benchmarks integrated with AmeriFlux and FLUXNET
- Create benchmarks for soil carbon turnover and vertical distribution and transport
- Develop benchmark metrics for extreme event statistics and responses of ecosystems
- Synthesize data for vegetation recruitment, growth, mortality, and canopy structure
- Create benchmarks focused on critical high latitude and tropical forest ecosystems
- Leverage observational projects and create a roadmap for remote sensing methods.

Small, targeted working groups should be formed to research and publish analyses supporting these priorities. Other priority areas that the community identified as important included photosynthesis, aboveground biomass and litter, permafrost processes, atmospheric radiation measurements, the three-dimensional structure of atmospheric CO_2, and the use of radiocarbon as a constraint on soil processes.

Specific Enabling Capabilities identified as required to address the next generation **Benchmarking Challenges and Priorities** included the following:

- Model development of new process representations and new output variables
- Deployment of land model testbeds (LMTs)
- Directed field measurements and monitoring activities
- Perturbation experiments and laboratory studies
- Standardize collection, processing, archiving, and distribution of observational data in Federated data centers
- Advanced computational resources and infrastructure.

New model development and verification activities could be more rapidly advanced through frequent and systematic simulation and testing. In particular, priority capabilities identified included the following:

- LMTs for automated execution, calibration, and evaluation of alternative or competing model formulations
- *In situ* diagnostics to summarize simulation results and avoid output of large data sets, which can greatly reduce computational efficiency
- Initial LMT development that implements AmeriFlux and FLUXNET supersite evaluation of single-point offline simulations
- LMT capabilities incorporated into existing routine model testing (e.g., nightly or weekly automated integration testing).

Computational needs and requirements identified for model development, testing, and advanced benchmarking included the following:

- Scalable algorithms and machine learning techniques for evaluating and benchmarking high resolution and long time series ESM results
- Cyber infrastructure to support large scale data collection, curation, archiving, and distribution, supporting automated model–data comparisons and online data assimilation for parameter estimation through supercomputing facilities

- A balance between resources for compute-intensive vs. data-intensive application as scientific computing facilities plan their expansion to exascale computing
- New development for ILAMB that includes improved support for remote retrieval and version tracking for observation data through repositories like obs4MIPs.

Additional field measurements and monitoring activities, as well as perturbation experiments and lab studies, could provide valuable observational data for constraining models. High priority measurement needs identified for developing benchmarks and improving ESMs included the following:

- Long-term energy, carbon, and water flux measurements at AmeriFlux and FLUXNET sites with standardized instrumentation and methods, and additional frequent or continuous ancillary *in situ* measurements of soil moisture, sap flow, tree height and diameter, litterfall, and soil nutrients
- High latitude and tundra soil core measurements of carbon and nutrient distributions, including isotopes and soil ice/water content, and observations of vegetation growth and expansion of woody vegetation
- Characterization of tropical ecosystem traits and canopy structure and chemistry; observations of tropical ecosystem responses to drought, increased temperatures, and elevated atmospheric CO_2; and measurements of nutrient cycling and hydrology in tropical forests, focusing on strong land–atmosphere interactions
- Remote sensing algorithms and processing infrastructure for generating data products useful for large-scale ecosystem characterization and monitoring, scaling up *in situ* measurements, and informing future measurement site selection.

9.2 Long-term Vision for Model Benchmarking

A productive approach for achieving breakthroughs in the areas described above would be to organize small working groups that bring together key individuals at the cutting edge of the target discipline along with ILAMB developers. Priorities for these synthesis activities are identified in Section 4. Over the course of several meetings, the teams would have a goal of creating new metrics. The teams also would use the ILAMB system to create figures and tables highlighting these metrics for a synthesis paper, for which all the contributing participants would share in co-authorship. The community also expressed enthusiasm in bringing the full community together for larger meetings, and there was consensus that this would be complementary to the smaller targeted working groups, especially if the larger meetings were organized on a 3–5 year cycle.

On a 3-year horizon (FY 2017–2020), the 6th phase of the Coupled Model Intercomparison Project (CMIP6) will be nearly complete, generating a vast archive of model simulation output from its suite of core DECK simulations and numerous associated MIPs (Section 5). The combined CMIP6 collection will provide information essential for governments around the world to limit the magnitude and impact of climate change. In this context, supporting CMIP6 must be a central ILAMB goal over the next three years, and it is expected to generate many unique challenges. For example, C^4MIP, LS3MIP, and LUMIP, as described in Section 5, all have unique objectives, simulation characteristics, and variable requests. Participants in these MIPs are interested in the ability of models to predict land surface changes on vastly different time scales and across a widely varying set of processes.

To successfully support these MIPs, further development and a unique tailoring of the ILAMB system for individual MIPs may be necessary. Within each MIP, ILAMB may help to identify robust responses that occur across multiple models as well as persistent biases. Using the DECK simulations and other closely related simulations, ILAMB also may be helpful in documenting improvements in the representation of the land surface and atmospheric processes over time, from CMIP5 to CMIP6. This information will be of broad interest to Earth system scientists, policy makers, funding agencies, and the general public. Another important goal will be to use the emergent constraints that are currently being integrated within the ILAMB system to constrain future predictions of carbon dioxide and other biogeochemical variables. In doing so, ILAMB participants may be able to enhance the value of CMIP6 for the Intergovernmental Panel on Climate Change 6th Assessment Report, and other international and national synthesis efforts.

Another necessary step is to create a closer coupling between obs4MIPs (Teixeira *et al.*, 2014; Ferraro *et al.*, 2015) and ILAMB. This can be achieved by integrating ILAMB datasets into the obs4MIPs online repository and converting existing ILAMB datasets to follow well-established netCDF Climate and Forecast (CF) conventions (Eaton *et al.*, 2011). Whereas obs4MIPs currently includes many datasets valuable for constraining the physical climate system, many ecosystem variables have not yet been integrated into this system. This step will make it easier for ILAMB developers to build new modules, and it will increase the transparency and traceability of the system as it evolves.

Over a 5–10 year time horizon, the ILAMB system could serve as a core capability within a US or international center dedicated to increasing information flow between international measurement and Earth system modeling communities. Other important capabilities, complementing ILAMB, would include the ability of the center to solicit small synthesis proposals from the community for new working groups, host MIP-related activities, and support expanded Earth system model use and access by a broader cross section of scientists within disciplines of ecosystem ecology, biogeochemistry, and hydrology.

Appendix A. Benchmarking Tools

A.1 PALS/PLUMBER

Gab Abramowitz and Martin Best

The Protocol for the Analysis for Land Surface models (PALS; Abramowitz, 2012) is an online web application for the automated evaluation and benchmarking of land surface model (LSM) simulations. PALS hosts a collection of "experiments," each of which contains a collection of data sets required to force (if running offline) and evaluate a LSM at the particular spatial resolution or location prescribed by the experiment. Users create model profiles within the PALS system, and then upload their LSM simulation and associate it with one of their model profiles and the appropriate experiment. Once uploaded, the analysis script associated with the experiment automatically analyzes the uploaded model output, comparing it to evaluation data sets and/or model outputs from other users that are already associated with the experiment. Results of the analysis are available to all users with access to the experiment.

There are several motivations for creating this type of system. Running model intercomparison projects (MIPs) in this environment means the following:

- Analyses are transparent to all involved because analysis scripts are downloadable and editable. Standardization of evaluation can therefore be a community-based effort.
- Contributions to MIPs can be ongoing, without additional analysis effort.
- Additional analyses can be performed by anyone with access to the experiment.
- The entire history of MIPs on the PALS system remain "live" and available.
- A version history of data sets, analysis scripts, and experiment metadata are accessible to all experiment users.
- Ancillary data associated with models and model outputs can potentially be data-mined as part of the analysis.
- Ancillary data associated with models and model outputs improves provenance information and reproducibility.

This makes achieving the broader goals of a MIP, such as understanding why some models perform better than others, or whether or not models share particular weaknesses, more attainable.

Another obvious use of such a system is for model development. PALS' implementation of "workspaces" to limit access to experiments to a subset of users means that development teams can use this type of system for fast, repeated analysis of model developments to share online with co-developers, as follows:

- The automated nature of analysis allows continuous integration testing for scientific model development through application programming interface (API) access (e.g., using Jenkins).
- Equity: access to the evaluation system is not contingent upon the ability to successfully install an analysis package or local computing resources. This increases the potential for international standardization of model evaluation and avoids duplication of analysis infrastructure.
- As noted above, ancillary data associated with model versions and model outputs improves provenance information and reproducibility and opens up the potential to data-mine ancillary data.
- The ability to nominate benchmarks for each analysis—other model outputs already associated with a particular experiment—makes comparing against different model versions easier.

Success of this type of system is clearly dependent upon the adoption of model input/output standards. PALS currently supports the Assistance for Land-surface Modeling Activities (ALMA) NetCDF standard to which many land surface modeling groups adhere. Work is underway to ensure full Climate and Forecast convention for NetCDF files (CF-NetCDF) compliance and Coupled Model Intercomparison Project (CMIP) interoperability in the next version of the ALMA standard.

In its first phase, PALS focused solely on single site (flux tower) analysis. It attracted about 230 users from more than 60 institutions in 20 countries, of which about 20% were active users. This version of PALS has not been available since late 2014 after a Struts vulnerability forced us to take it offline. However, while limited in scope, this resulted in two successful MIPs: PLUMBER (Best et al., 2015; Haughton et al., 2016) and SavMIP (Whitley et al., 2016).

For PLUMBER, land surface models were benchmarked for 20 observational FLUXNET sites ranging in geographical locations, climates, and land cover. Both simple physical models and empirical relationships were used to provide benchmarks for the sensible and latent heat fluxes in this study. The land surface models were not evaluated against each other but were individually ranked in comparison to the benchmarks.

The results showed that for standard statistical metrics, all of the land surface models had a similar performance relative to the benchmarks. The models had a better overall ranking compared to the simple physical models but were out-performed for both surface fluxes by a three variable piecewise linear regression. In addition, for the sensible heat flux, the models were outperformed by a single variable regression between the flux and the downward shortwave radiation. This demonstrates that further improvements can be made to the models without introducing additional complexity, but rather by making better use of the information contained in the forcing data.

Furthermore, assessing the performance of the model relative to the benchmarks for alternative statistical metrics based upon distributions showed that the models had differing overall rankings compared to the benchmarks. This suggests that previous development efforts among the international community have focused on optimizing for standard statistical metrics, but this does not necessarily result in overall better performance.

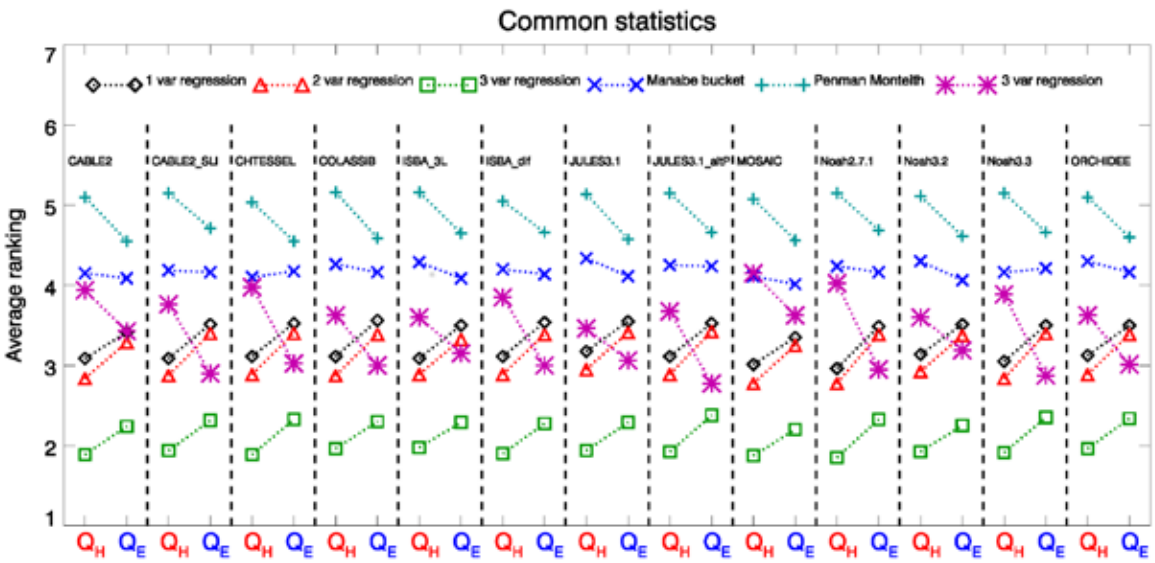


Figure A.1.1. Common statistics for each model are shown by average ranking from the PLUMBER benchmarking activity.

The second phase of PALS aims to broaden its focus and introduce new features. First, the system will not be specifically tailored to LSMs, so it will likely launch as **http://modelevaluation.org/**. All the existing PALS site-based LSM experiments, and additional global and regional LSM experiments, will still be available.

Next, experiment owners will be able to control the operation of the master analysis script. This means that as long as the JavaScript Object Notation (JSON)-based input/output requirements of the master analysis script are met, any analysis package can be used to perform the analyses for a given experiment. This means that incorporating evaluation packages such as ILAMB or Land surface Verification Toolkit (LVT) into the **http://modelevaluation.org/** environment is possible.

We are also building this system to avoid the bottleneck that uploading large model outputs inevitably creates. By using a distributed architecture, where the "worker" nodes that actually perform the analysis (e.g., using Python or R) can be co-located with or at centers producing large model outputs, "uploading" a model output to this system need not involve the transfer of large files. Instead, the central web server optimally manages a collection of worker nodes to minimize analysis time. Once results are complete, analysis images and summary data are then sent to the central web server for display to users.

An initial working version of the second phase system is running and undergoing testing. All code is available in a collection of open source GitHub repositories. Any suggestions, contributions or collaborations are actively encouraged.

A.2 PCMDI Metrics Package (PMP)

Peter Gleckler

A more routine benchmarking and evaluation of models is envisaged to be a central part of the sixth phase of the Coupled Model Intercomparison Project (CMIP6). One purpose of the Diagnostic, Evaluation and Characterization of Klima (DECK) and CMIP historical simulations is to provide a basis for documenting model simulation characteristics (Meehl et al., 2014). In addition to scientifically targeted tools under development like the ILAMB package, two capabilities (Eyring et al., 2016a; Gleckler et al., 2016) will more broadly characterize CMIP DECK and historical simulations as new model experiments are published on the Earth System Grid Federation (ESGF). The foundation that will enable this to be efficient and systematic is the community-based experimental protocols and conventions of CMIP, including their extension to obs4MIPs, which serves observations in parallel to the CMIP output on ESGF. Here we summarize some aspects of one of these capabilities—the Program for Climate Model Diagnosis and Intercomparison (PCMDI) Metrics Package (PMP; Gleckler et al., 2016).

The PMP is built on US Department of Energy (DOE)-supported tools (Williams et al., 2014) and emphasizes the implementation of a diverse suite of summary statistics to objectively gauge the level of agreement between model simulations and observations. The PMP software is open source, has a wide range of functionality, and is being developed as a community tool with the involvement of several institutions. Collectively, the PMP, Earth System Model Evaluation Tool (ESMValTool), and ILAMB packages offer valuable capabilities that will be crucial for the systematic benchmarking of the wide variety of models and model versions contributed to CMIP6. This evaluation activity can, compared with early phases of CMIP, more quickly and openly relay to analysts and modeling centers the strengths and weaknesses of the simulations including the extent to which long-standing model errors remain evident in newer models. In addition to being strongly integrated with the data conventions of CMIP, obs4MIPs and the ESGF, a priority for the PMP is to make all aspects of the analysis as traceable and reproducible as possible. All results from the PMP include a trail of the codes and dataset versions used to generate them.

We illustrate the type of summary statistics available via the PMP with three examples. The first (Figure A.2.1) is based on a recent paper (Ivanova et al., 2016) that examines how well simulated sea-ice agrees with measurements on sector scales and demonstrates that the classical measure of total sea-ice area is often misleading because of compensating errors. The second (Figure A.2.2) is also based on a recent paper (Covey et al., 2016) that highlights the amplitude and phase of the diurnal cycle of precipitation. A third example is given by a simple "portrait plot" comparing different versions of the same model (Gleckler et al., 2016) in Atmospheric Model Intercomparison Project (AMIP) mode.

The PMP is under rapid development with a priority of providing a diverse suite of summary statistics for all historical and DECK simulations to researchers and modeler developers soon after each simulation is published on the ESGF. The package is designed to enable community contributions. All the PMP code is hosted at **https://github.com/PCMDI/pcmdi_metrics**.

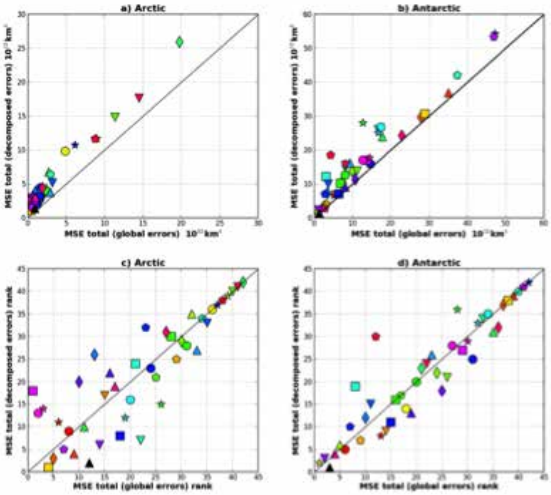


Figure A.2.1. Model ranking using mean-square error (MSE) of the total sea-ice area annual cycle: (a) Arctic scatter plot of decomposed and global errors; (b) Antarctic scatter plot of decomposed and global errors; (c) Arctic scatter plot of decomposed and global errors model ranking; and (d) Antarctic scatter plot of decomposed and global errors model ranking.

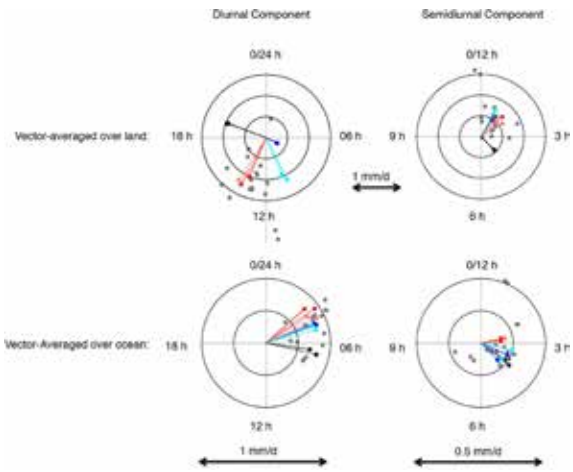


Figure A.2.2. Harmonic dial plots of the amplitude and phase of Fourier components, after vector averaging over land and ocean areas separately, for Tropical Rainfall Measurement Mission (TRMM) 3B42 observations (black lines and dots), for the four highest-resolution CMIP5 models (colored lines and dots), and for the other 17 Atmospheric Model Intercomparison Project (AMIP) models from CMIP5 with only July results shown for clarity (gray dots). For TRMM and the highest-resolution models, solid lines mark January results, whereas dashed lines mark July results.

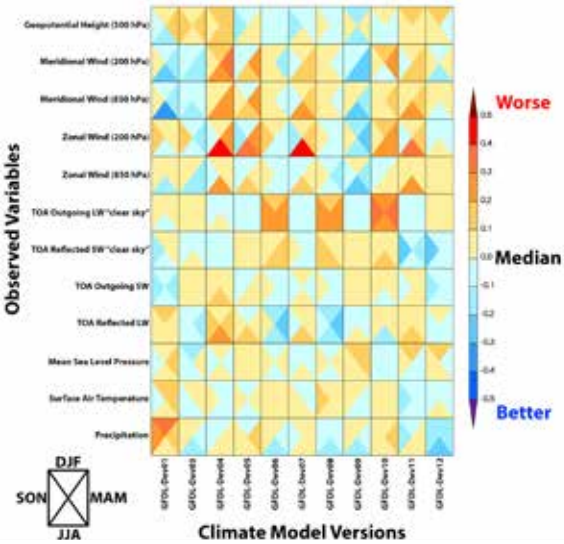


Figure A.2.3. Figure A2.3: Relative error measures of different developmental tests of the Geophysical Fluid Dynamics Laboratory (GFDL) model in AMIP mode. The error measure is a spatial root-mean-square error (RMSE) that treats each variable separately. The color scale portrays this as a relative error by normalizing the result by the median error of all model results (Gleckler et al., 2008). For example, a value of 0.20 indicates that a model's RMSE is 20% larger than the median error for that variable across all simulations, whereas a value of –0.20 means the error is 20% smaller than the median error. Credit: Erik Mason/GFDL.

A.3 ESMValTool

Veronika Eyring

A community diagnostics and performance metrics tool for the evaluation of Earth system models (ESMs) has been developed that allows for routine comparison of single or multiple models, either against predecessor versions or against observations. The priority of the effort so far has been to target specific scientific themes focusing on selected essential climate variables (ECVs), a range of known systematic biases common to ESMs, such as coupled tropical climate variability, monsoons, Southern Ocean processes, continental dry biases, and soil hydrology–climate interactions, as well as atmospheric CO_2 budgets, tropospheric and stratospheric ozone, and tropospheric aerosols. The tool is being developed in such a way that additional analyses can easily be added. A set of standard namelists for each scientific topic reproduces specific sets of diagnostics or performance metrics that have demonstrated their importance in ESM evaluation in the peer-reviewed literature. The Earth System Model Evaluation Tool (ESMValTool; doi:**10.17874/ac8548f0315**; Eyring et al., 2016a) is a community effort open to both users and developers encouraging open exchange of diagnostic source code and evaluation results from the CMIP ensemble. This will facilitate and improve ESM evaluation beyond the state of the art and aims at supporting such activities within CMIP and at individual modeling centers. Ultimately, we envisage running the ESMValTool alongside the Earth System Grid Federation (ESGF) as part of a more routine evaluation of CMIP model simulations while using observations available in standard formats (e.g., obs4MIPs) or provided by the user.

The ESMValTool consists of a workflow manager and a number of diagnostic and graphical output scripts (Figure A.3.1). The workflow manager is written in Python, whereas a multilanguage support is provided in the diagnostic and graphic routines. ESMValTool takes the name of a namelist file as a single input argument, and the namelist files are text files written using the eXtensible Markup Language (XML) syntax to define the model and observational data to be read, the variables to be analyzed, and the diagnostics to be applied. A large collection of standard namelists are included in ESMValTool (v1.0) for analyzing a wide collection of ECVs across atmosphere, ocean, sea ice, and land components. For example, one namelist can be used to reproduce the figures from the climate model evaluation chapter of IPCC AR5 (Chapter 9, Flato et al. [2013]) (Figure A.3.2). Another XML namelist will produce a plot comparing the RMSE over

different sub-domains for net biosphere productivity, leaf area index, gross primary productivity, precipitation, and near-surface air temperature like that of Anav et al. (2013) (Figure A.3.3).

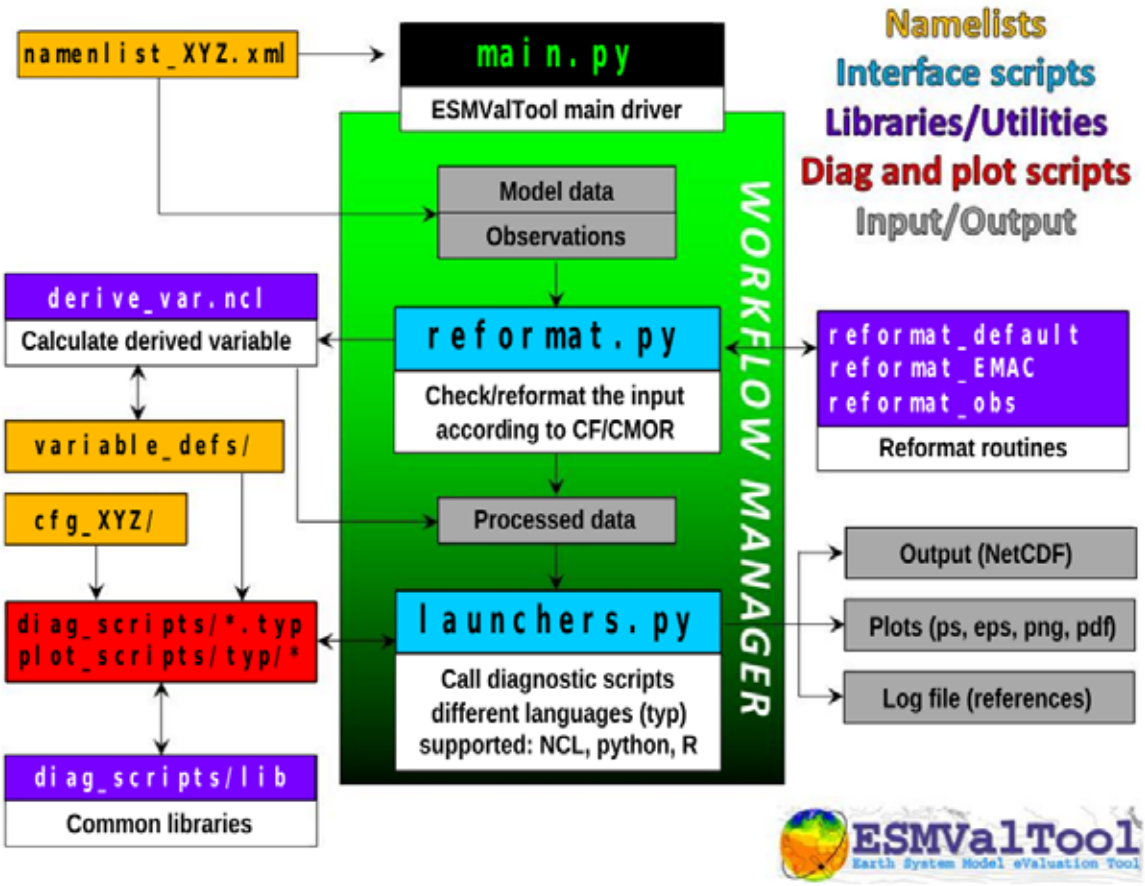


Figure A.3.1. Schematic overview of the ESMValTool (v1.0) structure. The primary input to the workflow manager is a user-configurable text namelist file (orange). Standardized libraries/utilities (purple) available to all diagnostics scripts are handled through common interface scripts (blue). The workflow manager runs diagnostic scripts (red) that can be written in several freely available scripting languages. The output of the ESMValTool (gray) includes figures, binary files (NetCDF), and a log file with a list of relevant references and processed input files for each diagnostic.

We aim to move ESM evaluation beyond the state of the art by investing in operational evaluation of physical and biogeochemical aspects of ESMs, by using process-oriented evaluation, and by identifying processes most important to the magnitude and uncertainty of future projections. Our goal is to support model evaluation in CMIP6 by contributing the ESMValTool as one of the standard documentation functions and by running it alongside the ESGF. In collaboration with similar efforts, we aim for a routine evaluation that provides a comprehensive documentation of broad aspects of model performance and its evolution over time and to make evaluation results available at a time scale that was not possible in CMIP5. The ability to routinely perform such evaluation will drive the quality and realism of ESMs forward and will leave more time to develop innovative process-oriented diagnostics – especially those related to feedbacks in the climate system that link to the credibility of model projections.

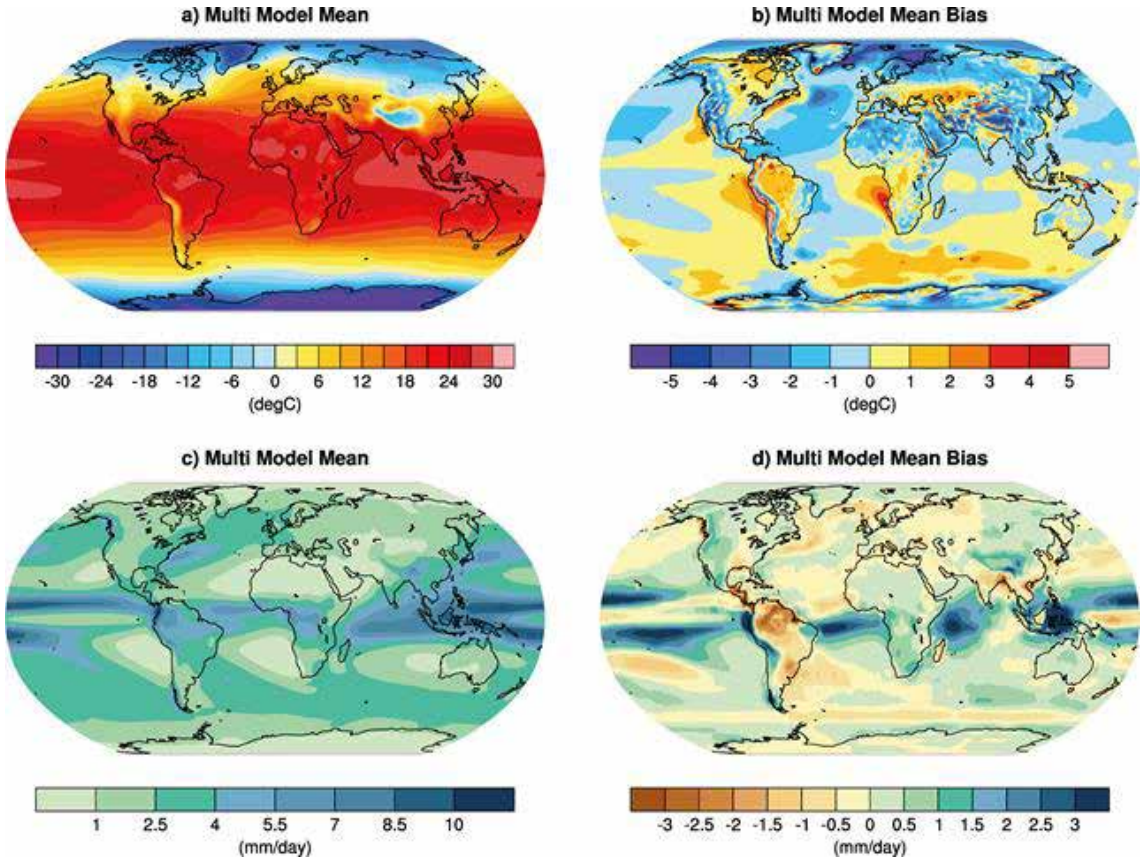


Figure A.3.2. Annual-mean surface air temperature (upper row) and precipitation rate (mm day^{-1}) for the period 1980–2005. The left panels show the multi-model mean and the right panels the bias as the difference between the CMIP5 multi-model mean and the climatology from ERA-Interim and the Global Precipitation Climatology Project for surface air temperature and precipitation rate, respectively. The multi-model mean near-surface temperature agrees with ERA-Interim mostly within ±2°C. Larger biases can be seen in regions with sharp gradients in temperature, for example in areas with high topography such as the Himalaya, the sea ice edge in the North Atlantic, and over the coastal upwelling regions in the subtropical oceans. Biases in the simulated multi-model mean precipitation include too low precipitation along the equator in the western Pacific and too high precipitation amounts in the tropics south of the equator. Similar to Figures 9.2 and 9.4 of Flato et al. (2013) and produced with ESMValTool *namelist_flato13ipcc.xml*.

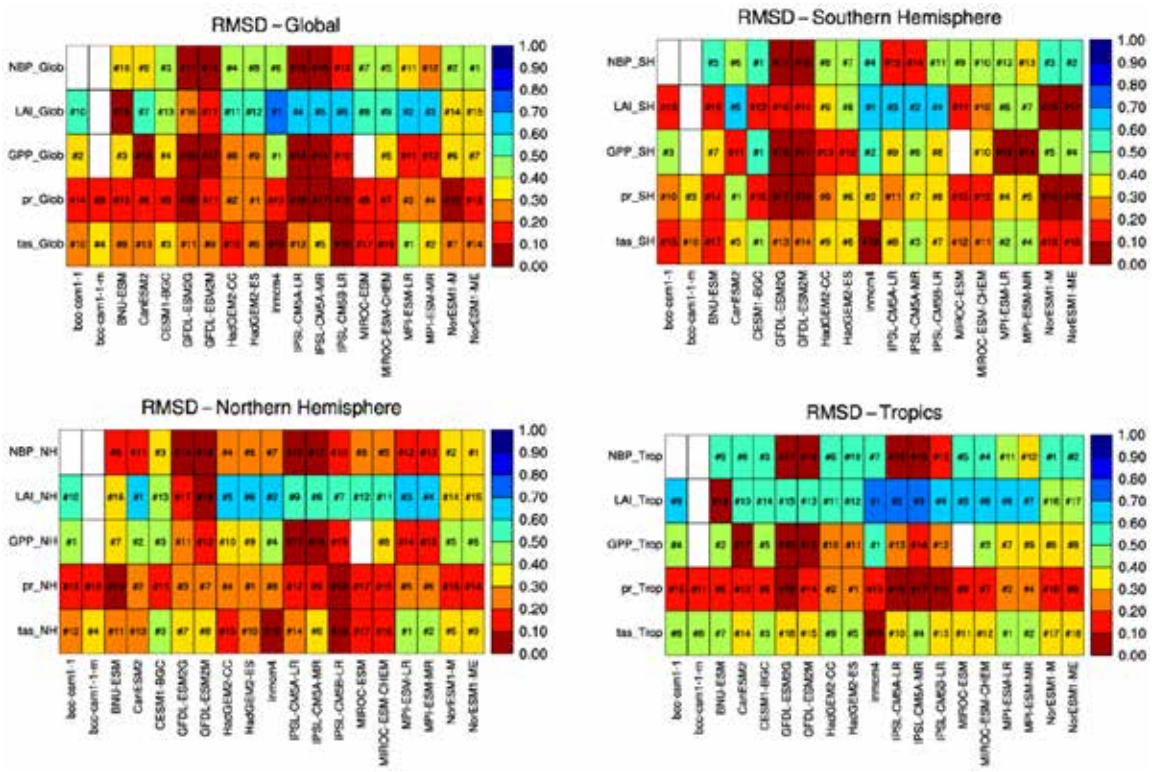


Figure A.3.3. Relative space–time RMSE calculated from the 1986–2005 climatological seasonal cycle of the CMIP5 historical simulations over different sub-domains for net biosphere productivity (NBP), leaf area index (LAI), gross primary productivity (GPP), precipitation (pr) and near-surface air temperature (tas). The RMSE has been normalized with the maximum RMSE to have a skill score ranging between 0 and 1. A score of 0 indicates poor performance of models reproducing the phase and amplitude of the reference mean annual cycle, whereas a perfect score is equal to 1. The comparison suggests that there is no clearly superior model for all variables. All models have significant problems in representing some key biogeochemical variables such as NBP and LAI, with the largest errors in the tropics mainly because of a too weak seasonality. Similar to Figure 18 of Anav et al. (2013) and produced with ESMValTool *namelist_anav13jclim.xml*.

A.4 NASA Land Surface Verification Toolkit (LVT)

Sujay Kumar

The NASA Land surface Verification Toolkit (LVT; **http://lis.gsfc.nasa.gov/software/lvt**; Kumar et al., 2012) is an open-source, formal system for land surface model evaluation and benchmarking. LVT is designed to provide an automated, consolidated environment for model evaluation and includes approaches for conducting both deterministic and probabilistic verification. A key motivation in the development of LVT is the concept of "model–data fusion" (MDF; Raupach et al., 2005; Williams et al., 2009), which is the paradigm for combining the information from models and observational data, to aid the formulation, characterization, and evaluation of models in a structured manner. The evaluation step is a critical element that is necessary to complete the MDF paradigm. LVT was initially developed to augment the land surface modeling and data assimilation framework known as the Land Information System (LIS; Kumar et al., 2006). LIS includes several key components of the MDF paradigm, including a suite of land surface models, computational tools such as data assimilation, optimization and uncertainty estimation. Together, LVT and LIS provide a comprehensive environment to enable the MDF paradigm.

LVT is implemented using object oriented framework design principles as a modular, extensible, and reusable system. The software is designed with explicit interfaces for incorporating support for observational datasets and evaluation metrics. The interoperable nature of the LVT design allows the reuse of existing features with new components that are developed. For example, a newly incorporated support for an observational dataset can take advantage of all available analysis metrics without needing any additional implementation.

A key design consideration in LVT is the support of observational datasets in their native formats, enabling the continued use of the system without requiring additional implementation or data preprocessing. Currently a large suite of *in situ*, remotely sensed, and other model and reanalysis datasets are implemented in LVT. The spatial and temporal scales of these measurements vary significantly. LVT handles the geospatial and temporal transformations of these datasets from their native formats to enable flexible analysis configurations.

In recognition of the need for having a variety of performance evaluation metrics for model evaluation, LVT supports a suite of analysis metrics. Aside from the traditional accuracy-based measures, LVT also includes metrics to aid model identification, such as entropy, complexity, and information content. These measures can be used to characterize the tradeoffs in model performance relative to the information content of the model outputs. The accuracy-focused metrics that quantify model performance using residual-based measures often do not provide insights on the robustness of the model under future or unobserved scenarios. The availability of metrics such as those based on information theory helps in mitigating these limitations. In addition to model verification, LVT also provides an environment for model benchmarking, where benchmark values for each metric are established a priori (Best et al., 2015). The development of such benchmarks is facilitated in LVT, using regression and machine learning techniques. More recently, application-oriented, end-user focused diagnostic measures have been developed. For example, LVT can be used to produce estimates of drought/flood risks by analyzing the distribution of soil moisture, streamflow, or evaporative fluxes from the land surface model. Finally, LVT also includes uncertainty and ensemble diagnostics based on Bayesian approaches that enable the quantification of predictive uncertainty in land surface model outputs.

LVT is an evolving framework and continues to be enhanced with the addition of new analysis capabilities and the incorporation of terrestrial hydrological datasets. The capabilities in LVT provide novel ways to characterize LSM performance, enable rapid model evaluation efforts, and are expected to help in the definition and refinement of a formal benchmarking and evaluation process for the land surface modeling community.

A.5 ABoVE Benchmarking System

Joshua B. Fisher

The Arctic-Boreal Region (ABR) is a major source of uncertainties for terrestrial biosphere model (TBM) simulations. These uncertainties are precipitated by a lack of observational data from the region, affecting the parameterizations of cold environment processes in the models. Addressing these uncertainties requires a coordinated effort of data collection and integration of the following key indicators of the ABR ecosystem: disturbance, flora / fauna and related ecosystem function, carbon pools and biogeochemistry, permafrost, and hydrology. We are developing a model-data integration framework for NASA's Arctic Boreal Vulnerability Experiment (ABoVE), wherein data collection for the key ABoVE indicators is driven by matching observations and model outputs to the ABoVE indicators. The data are used as reference datasets for a benchmarking system which evaluates TBM performance with respect to ABR processes. The benchmarking system utilizes performance metrics to identify intra-model and inter-model strengths and weaknesses, which in turn provides guidance to model development teams for reducing uncertainties in TBM simulations of the ABR. The system is directly connected to the International Land Model Benchmarking (ILAMB) system, as an ABR-focused application.

Appendix B. Metrics for Major Processes

B.1 Ecosystem Processes and States

Nancy Y. Kiang and Ben Bond-Lamberty

Ecosystem processes are the full suite of interactive components of an ecosystem that determine a column mass budget and fluxes into and out of the system vertically and horizontally. Ecosystem components are typically distinguished in land models into modules for soil biogeochemistry coupled with vegetation dynamics (biophysics, phenology, growth, ecology), and with these biological components coupled to surface hydrology and the atmosphere. Thus, system processes are (1) the vertical interactions between these components from the ground hydrology to the atmosphere (e.g., the exchange of water, litter, nutrients, and sum of energy and gas fluxes) and (2) the horizontal exchanges and external forcings that lead to heterogeneous boundary conditions for these column physics (e.g., edge effects, transport, disturbance, and dispersal, the latter being covered under the section on Vegetation Dynamics).

Ecosystem states are the magnitudes of these fluxes and mass storage pools at a point in time, as well as their trajectories with respect to time or another driver. The pools may be categorized according to system components and various classifications of their respective compositions, such as biodiversity, chemical mix, and geometric structure.

Table B.1.1 provides a summary of ecosystem processes addressed in this section, focusing on processes that couple ecosystem components with each other. Table B.1.2 provides a summary of ecosystem state variables that are targets for benchmarking, together with data sets that could serve as these benchmarks. There is some natural overlap with other sections of this report that focus on the ecosystem components. Further details on identifying appropriate model ecosystem diagnostics and suitable data for model benchmarking serves a primary goal of improving ecosystem process representation.

Table B.1.1. Ecosystem coupling processes.

	Physics	Biophysics and Biogeochemistry	Ecology
Land–Atmosphere	Observed or GCM meteorology Canopy albedo Surface energy balance Water vapor conductance	CO_2 exchange Autotrophic respiration Heterotrophic respiration	Fire emissions Anthropogenic forcings
Vegetation–Soil	Canopy air: temperature, humidity, CO_2 concentration	Litterfall mass and quality (C:N ratio, lignin content) Nitrogen dynamics	Microbial-vegetation nutrient competition
Hydrology–Soil–Vegetation	Layers vs. catchments Interception/throughfall Root water uptake, stomatal conductance	Multi-pool, multi-layer soil (Dissolved organic carbon) Leaching of NO_3^-	
Horizontal Exchange	Edge effects in meteorology	General circulation of CO_2 and fire emissions	Managed land dynamics, land use; Natural and anthropogenic disturbance, fire

Table B.1.2. Ecosystem State Model Diagnostics vs. Measurements.

	Equilibrium spin-up state Preindustrial control Partitioning/ classification of mass balances and fluxes.	**Responses/Sensitivities** Elevated CO_2	**Uncertainties**
Land–Atmosphere	Model CO_2, surface fluxes CO_2 record: flasks, ice cores Products from FLUXNET	Model mean, seasonal timing latitudinal gradients Airborne fraction	
Vegetation Canopy	Vegetation structure Net zero flux FLUXNET, inventory, satellite	Seasonal timing, net fluxes Land use and land cover change (LULCC)?	
Soil	Model litter layer, SOC, soil N Soil carbon databases Land Use Model Intercomparison Project (LUMIP) management data sets	dC/dX, dC/dt Soil flux databases	High observational uncertainties

Specific Points and Recommendations

Key recommendations to improve evaluation, benchmarking, and process representation of ecosystem processes and states in ESMs are as follows:

- To interpret and compare the performance of models relative to benchmarks, it is necessary to analyze the component parts of each model and not merely their emergent behavior. There should be more focus on comparing process representation and not just diagnostic variables.
- To create standards for benchmarks, the land modeling community must develop clear guidelines on how different statistics and visualizations (e.g., bias, RMSE, Taylor diagrams) are used and how they complement each other for different benchmarking purposes.
- Observational data often lack quantified uncertainties. These should be required as an essential component of data products in benchmarking tools like ILAMB to be useful to inform, constrain, and benchmark models. Uncertainty in forcings, boundary condition data sets, and parameter sets is needed to quantify weights properly in propagation of uncertainty in model simulations.
- In model development, it is critical that tests are designed to eliminate confounding factors that would affect interpretation of the effects of new model physics. Examples of confounding factors that influence model performance other than a new model update include forcings data sets and boundary conditions, for which controls should be selected to identify model improvements versus other factors.
- To improve model process representation, the observation and modeling communities should communicate regularly their perspectives with each other so that (1) the measurement community develops functional relationships from data sets that are suitable for use in models and (2) modelers can keep informed of insights from new data. Modelers need to provide the observation community with a clear definition of needs, such as through a scaled-based matrix of measurement needs for models.

B.1.1 Scientific Challenges and Opportunities for Model Evaluation

Accuracy: A number of statistical and visualization approaches have been used to evaluate model performance (e.g., bias, RMSE, phase, amplitude, spatial distribution, scores, Taylor diagrams, and functional relationships/perturbation sensitivity) (Gleckler et al., 2008; Doney et al., 2009; Luo et al., 2012). To create standards for benchmarks, the land modeling community must develop clear guidelines on how different statistical measures are used and how they complement each other for different benchmarking purposes. With regard to known issues with specific data, the ability of both measurements and models to close energy and carbon budgets is advocated as a continued important accuracy criterion.

Uncertainty: Uncertainty in observational data is often lacking and should be demanded as an essential component of data products in benchmarking tools like ILAMB to be useful to inform, constrain, and benchmark models. Uncertainty in forcings, boundary condition data sets, and parameter sets is needed to quantify weights properly in propagation of uncertainty in model simulations.

Sensitivity: Insight into model behavior can be gained through checking relationships: variable vs. variable, vs. time, vs. drivers, turnover/response rates. Because process representations generally directly encode sensitivities found in observations, directly examining the different models' physics should be the first analytical step for evaluating and anticipating their different behaviors. However, sensitivity between coupled ecosystem components is an area worth developing for benchmarking for emergent properties of ecosystems.

Scaling—temporal: Understanding at which time scale a process has significant influences is vital to representing it appropriately in models. To discern these time scales from both observational data as well as model outputs, a suggested approach is Fourier transforms of time series and periodicities. This has been used, for example to analyze patterns of diurnal, seasonal, and interannual cycles.

Scaling—spatial: In scaling up (e.g., from sampling points at a site, from sites to regions, and regions to the globe), land modelers must remain cognizant that each change in scale entails different relevant ecosystem processes (cf., Moorcroft et al., 2001). From sampling within a field site, the distribution and variability of point measurements with microclimate and individual plant heterogeneities need to be quantified well to scale up model processes to the ecosystem scale (cf., Shao et al., 2013; Keenan et al., 2012; Todd-Brown et al., 2013). Scaling up from site-based studies to the regional and global scale must account for disturbance effects, anthropogenic forcings, and teleconnections that are not observed at the site scale but that operate at the larger scale. At the same time, to account for numerical issues, approaches must be developed to downscale or tune column physics at the ESM grid scale for processes that operate at the subgrid scale, such as soil moisture and precipitation.

B.1.2 New Metrics and Benchmarking Approaches

Benchmarking metrics provide a standardization for model evaluation and a bridge between what land modelers can simulate and what the observational community can measure. The advent of size-structured and patch-age based second generation dynamic global vegetation models (DGVMs) and trait-based vegetation models, the introduction of more ecosystem types and land use change, and the availability of more measurements from long-term sites and recent satellites, all motivate re-evaluation of old benchmarking metrics and addition of new ecosystem metrics.

Table B.1.3 provides a summary of key ecosystem process and state metrics for standardization in the land modeling community. These draw upon also the efforts of the various model intercomparison projects (MIPs) of the Coupled Model Intercomparison Project 6 (CMIP6), particularly the Coupled Carbon Cycle Climate Model Intercomparison Project (C^4MIP), where the goal is to constrain future climate projections (e.g., identify emergent constraints). As with C^4MIP, we recommend the community develop standard model diagnostic variables, units, and time scales of averaging.

The metrics include pre-industrial spin-up benchmarks where there are no observations to compare to, but an equilibrium model state must be defined, such as potential biomass and equilibrium soil carbon. The metrics also include variables suitable for evaluation against the observational record. We recommendation that models also develop instrument simulators to output the fundamental measurements observed by remote sensing instruments. These could include updating the fundamental canopy radiative transfer model (RTM) code such that it outputs canopy reflectance or thermal brightness temperatures based on the internal canopy structure, optical properties, or thermal properties. In addition, an important component would be the capability to simulate basic LiDAR waveform

information based on canopy properties. That capability should be based on the model's RTM representation, to best compare with LiDAR observations, instead of converting to estimates of height or biomass. For example, LiDAR waveforms of the simulated vegetation structure could be produced for direct comparison with LiDAR measurements by using an internal radiative transfer model. Other examples are simulation of solar-induced fluorescence (SIF) or shortwave albedo in the same band as measurements.

Table B.1.3. New Metrics/Model Diagnostics/Benchmarks.

	Activity	Physical Properties	Ecosystem Structure	Temporal Diagnostics	Spatial Diagnostics
Land–Atmosphere				Seasonal timing	Horizontal column Vertical regional
Vegetation Canopy	Fluorescence	Albedo	Age since disturbance. Plant age, geometry, demography, biomass. LiDAR waveforms	Seasonal timing Decadal-centennial prediction	RMSE, uncertainty
Soil			Parameter values - data repository	Seasonality of fluxes	RMSE, interpolation
Vegetation–soil			Litterfall mass, litter layer	Seasonality	Requires data

B.1.3 Observational Data Needs and Priorities

Current best-available datasets must be selected based the relevant time scale (annual mean, seasonal cycle, interannual variability, trend) and the spatial extent and resolution for comparison (site, regional, global). New *in situ* or remote sensing measurements are needed for global soil depths, isotope tracers, leaf area index, and many other state variables. A wide variety of measurements are needed to characterize specific phenomena of interest, including drought. Appropriate metadata (e.g., site history) must accompany all field data. Synthesis of data from a variety of sources (e.g. FLUXNET, TRY, Allometree, NECTAR), and coordination among data centers providing open standard APIs is crucial.

Table B.1.4. Observational Data Needs.

	Ecosystem Structure	Physics	Biogeochemistry	Ecology	Scaling Up
Atmosphere					Flux inversion products
Vegetation Canopy	Age distribution of disturbance, plant demography. Height. Root exudates. Reproduction. Allometric leaf area index and seasonality of traits.	Seasonality of leaf traits Hyperspectral data	Vegetation structure Site: Airborne: Remote:	Cover change	Beyond PFTs "Decomposition functional types" (Bond-Lamberty et al., 2016b)
Soil	More soil state and response data needed: C, N, bulk density. Partitioning of soil hetero- vs. autotrophic respiration.		Soil respiration. Updated gridded soil respiration observational data	Peatlands	

B.1.4 Model Development and Output Requirements

To improve ecosystem process representation, the land modeling community should investigate advances to these aspects of coupling ecosystem components:

- **Energy exchange**: Second generation vegetation models that represent canopy heterogeneity and seasonally prognostic leaf albedo should be evaluated to determine if they improve the prediction of surface albedo, canopy and ground temperature, and surface energy balance.
- **Water exchange**: First, litterfall is a poorly constrained ecosystem exchange process between vegetation and soil. The mulching effect of a litter layer to insulate the soil and conserve soil moisture is well known but lacks a mechanistic modeling approach for ESM grid scales. Matthews (1997) produced a benchmark estimate of litter production and pools with regard to annual dry matter production according to vegetation type and climate. However, seasonal variation in the physical properties of a litter layer (mass, heat capacity, moisture conductivity) by ecosystem type and seasonally is poorly known. Eddy flux sites should be monitored to develop relationships between temporally varying litterfall quantity, decomposition processes, and litter layer physical properties. Second, water stress remains a tuned control on plant stomatal conductance relative to a particular land model's soil hydrology. ESM land surface models do not typically have very deep soils, meaning that water stress and conductance of deep-rooted plants are inadequately represented. Pelletier et al. (2016) have produced the first global gridded map of soil thicknesses to bedrock, and implementation of this soil depth map in more ESM land models will enable deeper-rooted soil–vegetation–atmosphere coupling in the conductance of water vapor.
- **Carbon exchange**: Litterfall from vegetation as an input to soil biogeochemistry is subject to high uncertainty in model simulations due to uncertainty in leaf mass per leaf area and weak performance of leaf phenology models for the timing of senescence. Introduction of deeper roots with deeper soils will alter vegetation–soil and water–carbon coupling in modeled ecosystems, as it will motivate revision of each DGVM in its distribution of soil carbon from senescing roots, and in plant allometry and carbon allocation to roots. Phenological timing remains poorly simulated but the approaches of Stöckli et al. (2008, 2011) and Caldararu et al. (2014) are worthy of experimentation in more land models.
- **Nutrient exchange**: For those ESMs that include soil–plant nitrogen dynamics, plant biomass pools typically have fixed C:N ratios, and their growth drives demand for soil N. N inputs are generally from deposition. Improved representation with varying C:N should be explored.

Table B.1.1. Ecosystem coupling processes.

	Physics	Biogeochemistry	Ecology
Atmosphere		CH_4	
Vegetation Canopy	Beyond PFTs Leaf physiology Phenology Respiration partitioning SIF	C:N:P	Community structure: height-stratified canopies Managed land dynamics Wetlands Herbivory, insects Climate change/elevated CO_2 responses "Decomposition functional types" (Bond-Lamberty et al., 2016b)
Soil	Layers vs. catchments Permafrost Deep soil Erosion	C:N:P, CH_4, N_2O	Other functional pools?
Ocean coupling	Runoff	Nutrient fluxes	

B.2 Hydrology

Randal D. Koster and Hongyi Li

B.2.1 Scientific Challenges and Opportunities for Model Evaluation

The key role of hydrology in land surface models (LSMs) is to partition incoming precipitation water into evapotranspiration (ET), runoff (streamflow), and changes in soil moisture storage. These water cycle calculations are intrinsically tied to energy balance calculations (e.g., through the connection between ET and latent cooling) and carbon balance calculations (e.g., through the control of stomatal conductance on transpiration). Soil moisture (its vertical profile and spatial variations) lies at the heart of land surface control over moisture fluxes, including both ET and runoff.

A wide variety of terrestrial processes are relevant to surface hydrology: ET and its component parts, streamflow generation, snow, permafrost, subsurface moisture transport, and human water management and disturbance, to name just a few. Also of relevance are groundwater dynamics, with different timescales connecting deep and shallow groundwater processes with surface hydrology. River routing is a key process to consider; evaporation from stream surfaces provides moisture to the atmosphere, and the streams and rivers themselves inject fresh water into oceans and lakes, a needed input flux for ocean models. Rivers also transport and transform nutrients through the Earth system, and lakes and wetlands slow these transport times. Additional relevant processes are discussed below.

Current State of Process Representations in Models

Today's LSMs compute a broad suite of hydrological fluxes (e.g., infiltration, interception loss, surface runoff, baseflow, soil moisture storage change). However, the accuracy of these fluxes is arguably limited by key disparities in model complexity. For example, in many models the "vertical" treatment of the land surface is highly detailed, with multiple stacked soil layers overlain by a complex canopy structure. One-dimensional physics can thus be said to be well-represented. However, many aspects of hydrological behavior are affected equally by horizontal complexity—spatial variability (not explicitly resolvable in climate model-based land surface schemes) in topography, vegetation (including root distributions), soil properties, and soil moisture itself. Emphasizing complexity in the vertical at the expense of the horizontal leads to poor model performance. Balancing process complexity for strongly coupled processes (e.g., ET versus runoff formulations) is also important for good model performance.

Poor representation of runoff is also reflected in (1) the lack of appropriate complexity in groundwater modeling and (2) underrepresented aquatic processes, especially in rivers. Groundwater formulations are restricted by the lack of lateral fluxes between land grid cells and the lack of realistic, spatially variable depths to bedrock. Both lead to poor simulation of groundwater table dynamics, which can interact with runoff generation processes. Riverine processes are also oversimplified, leading to a lack of lateral water fluxes between terrestrial water bodies (e.g., rivers, lakes, wetlands) and land, which will modulate the soil moisture at certain spatial and temporal scales. It also leads to underrepresented linkages to the atmospheric model via water, energy, and carbon fluxes from the river water surface (particularly when inundation is not properly modeled), and to the ocean model via terrestrial discharges at river mouths.

Existing Approaches for Assessing Model Performance

Many approaches are currently used to assess land model performance in producing hydrological fluxes. Flux tower data are used to assess ET, for example, and streamflow measurements (once corrected for human influence) are used to assess runoff production. *In situ* soil moisture measurements have been used to evaluate model soil moisture, and the advent of satellite-based soil moisture measurements is allowing such validation to proceed at the global scale. Satellite-based datasets of ET and vegetation phenology (e.g., NDVI) have more recently been used to evaluate land model output.

B.2.2 New Metrics and Benchmarking Approaches

New Metrics, Scores, and Functional Relationships

New work is needed to better evaluate hydrological processes in LSMs. For example, these models produce runoff (streamflow), which is reasonably well measured. While annual and seasonal streamflow in unmanaged systems is already a staple of model evaluation, work is needed to extend current time series analyses to determine if models reproduce slow versus fast responses and capture the impact of managed flows. Similarly, models produce soil moisture information that could be evaluated in the context of drought identification and potentially lead to a more useful drought index.

Since direct measurements of many hydrological fluxes are unavailable, methods for novel indirect estimation of these fluxes should be developed. For example, satellite-based fluorescence measurements may prove useful for evaluating transpiration, and other vegetation-focused measurements (e.g., NDVI) may be useful for constraining land models with dynamic vegetation. Functional relationships between directly measurable variables and those that are not could be very useful in hydrologically ungauged areas. For example, functional relationships have been reported between the Horton index (the ratio of catchment ET and available soil moisture for ET) and NDVI, between the aridity index (the ratio of evaporative energy and annual precipitation) and floods, etc. The capability of LSMs to reproduce such functional relationships could enable diagnosis not only of the effectiveness of the representation of individual processes but also the balance of complexity in the treatments of model components.

The joint control of soil moisture over ET and runoff in nature and in LSMs suggests one potentially valuable benchmarking approach. Because ET and runoff both vary with soil moisture, they effectively vary with each other. A land model should be able to reproduce observations-based relationships between ET and streamflow production efficiencies, with soil moisture (a largely model-dependent variable) taken out of the picture. Techniques for such benchmarking currently exist.

Since most applications of LSMs and ESMs are large-scale in nature, the influences of human systems on the water cycle are not negligible. Caution is thus necessary regarding the role of human impacts while designing and applying new metrics over large scales. A related issue is potential nonstationarity: a model may validate well for present-day climate, but will it also perform well under a modified climate? Evaluations should proceed with this concern in mind.

Current Best-available Data Sets for Specific New Metrics

Existing datasets can be used as the basis for new metrics. For runoff and streamflow-related metrics, Model Parameter Estimation Experiment (MOPEX) data are largely ideal for pristine headwater watersheds over the United States and Global Runoff Data Center (GRDC) data are the best available for global streamflow metrics, though because the GRDC basins are largely regulated, caution is needed in their use. For soil moisture-related metrics, both in situ measurements and satellite-based datasets (SMOS, SMAP, ASCAT) are of great value.

B.2.3 Observational Data Needs

Gaps in Current Data Availability

The lack of snow water equivalent (SWE) data on the global scale is a significant deficiency. Moreover, direct measurements of ET at large spatial scales are not available; at best we have access to indirect evaluations through, for example, the analysis of streamflow (see above), the upscaling of FLUXNET site data using satellite information (e.g., NDVI), or the interpretation of diurnal temperature cycles in terms of latent heat flux. Furthermore, while streamflow data are available, separate datasets are needed for managed and unmanaged systems. Human impacts also take the form of irrigation, and irrigation data are sparse, if not absent. Collocation of different measurements would greatly increase their value.

New *in situ* or Remote Sensing Measurement Needs

A number of currently underutilized *in situ* datasets would contribute significantly to the evaluation of simulated land surface hydrology. For example, sap flow measurements may provide valuable information on transpiration, and direct or indirect measurements of macropore structures are still lacking. Remote sensing has the potential to provide a number of datasets relevant to evaluating land model hydrological fluxes. The ECOSTRESS mission, for example, focuses on ET, MODIS provides information relevant to both ET and snow, ASO also provides snow information, GPM provides precipitation data, SMAP provides data on surface soil moisture, GRACE data are relevant to terrestrial water storage, and SWOT (and AirSWOT) will provide useful information on surface runoff. The global coverage of these datasets gives them unprecedented value for the evaluation of land model products. Measurements are never perfect, and all measured variables should be provided with associated uncertainty estimates.

Spatial and Temporal Extent and Resolution Requirements

Any metric for evaluating a land model's simulation of hydrology needs to be valid for a large spatial area; local site measurements (e.g., flux towers) are, in isolation, inadequate. This is because: (i) the key hydrological flux, runoff (streamflow), is not measured at local sites; and (ii) land surface models are meant to produce large-area estimates of surface fluxes. Runoff production and ET vary substantially in space as a result of spatial heterogeneity in soil moisture, soil properties, and vegetation properties. Hence the measurement of runoff production at a local site has limited usefulness, even if the measurements are of high accuracy. Stream gauge measurements, in contrast, integrate spatially the runoff generated across a basin and are thus ideal targets for land model hydrological validation; they constitute a useful basis for new metrics and benchmarking. By validating large-scale runoff through streamflow measurements, the modeler is also arguably benchmarking aspects of large-scale ET.

Synthesis Activities Needs and Approaches

Combining different available soil moisture datasets into a single, long-term dataset for model evaluation would be useful. Such a synthesized dataset can be derived from *in situ* soil moisture measurements and a number of different satellite-based soil moisture products. Parallel work on model development is needed to bring the land model's soil moisture variables more in line with these measurements. Another example of a proposed synthesis activity is the development of a global dataset of pristine (unmanaged) watersheds, similar to the MOPEX dataset but extended to the global domain. The content can potentially even be extended to incorporate additional watershed-scale measurements or estimates such as soil moisture or SWE, which might provide new insights not underpinned by the in situ measurements.

B.3 Atmospheric CO_2

Gretchen Keppel-Aleks and William J. Riley

B.3.1 Scientific Challenges and Opportunities for Model Evaluation

Atmospheric CO_2 integrates both land and ocean fluxes over large spatial scales, providing a unique constraint on integrated fluxes. The concentration footprint of atmospheric CO_2 ranges 10^6 km^2 to hemispheric, depending on the location, altitude, and vertical extent of the observation. The fact that atmospheric CO_2 integrates over large areas and is quite sensitive to atmospheric transport complicates the use of CO_2 for benchmarking because model–data mismatch may be attributed to either carbon fluxes or atmospheric transport. Therefore, it would be possible to alias a transport-induced error into a comparison intended to evaluate carbon fluxes. At this point, mismatch in CO_2 diagnostics for predictive models may be dominated by carbon exchange, but constraining error in the atmospheric transport operator is crucial and will become a more significant source of error as carbon cycle models evolve. Despite the complication of these characteristics, atmospheric CO_2 has been used successfully to benchmark simulated time series (e.g., Lindsay et al., 2014), seasonal patterns (e.g., Keppel-Aleks et al., 2013), functional relationships at interannual timescales (e.g., Keppel-Aleks et al., 2014), and multi-decadal trends (e.g., Graven et al., 2013). CO_2 has also been used as emergent constraints (e.g., Cox et al., 2013; Hoffman et al., 2014; Figure B.3.1). Functional relationships may provide insight into linkages between biogeochemistry and physical climate, and thus will be useful for emergent constraints on centennial scale prediction.

There are multiple opportunities to develop atmospheric CO_2 as a benchmark. Some fully coupled ESMs have the capability to simulate the three-dimensional structure of CO_2. Several ESMs include capabilities to simulate isotopic fractionation in terrestrial processes, and including a 3-D $\delta^{13}CO_2$ tracer would facilitate evaluation against observations from surface networks. Transport of CO_2 throughout the atmosphere is relatively facile, because it is a passive tracer and, to first order, chemical formation *in situ* can be neglected. Further, the isotopic composition of CO_2 can be used to attribute variations specifically to certain sources. For example, $\delta^{13}CO_2$ is a useful tracer of terrestrial CO_2. Finally, there are opportunities to better integrate the use of atmospheric CO_2 with local scale constraints, to identify model occasions when mismatches between local-scale observations and the models lead to regionally coherent biases.

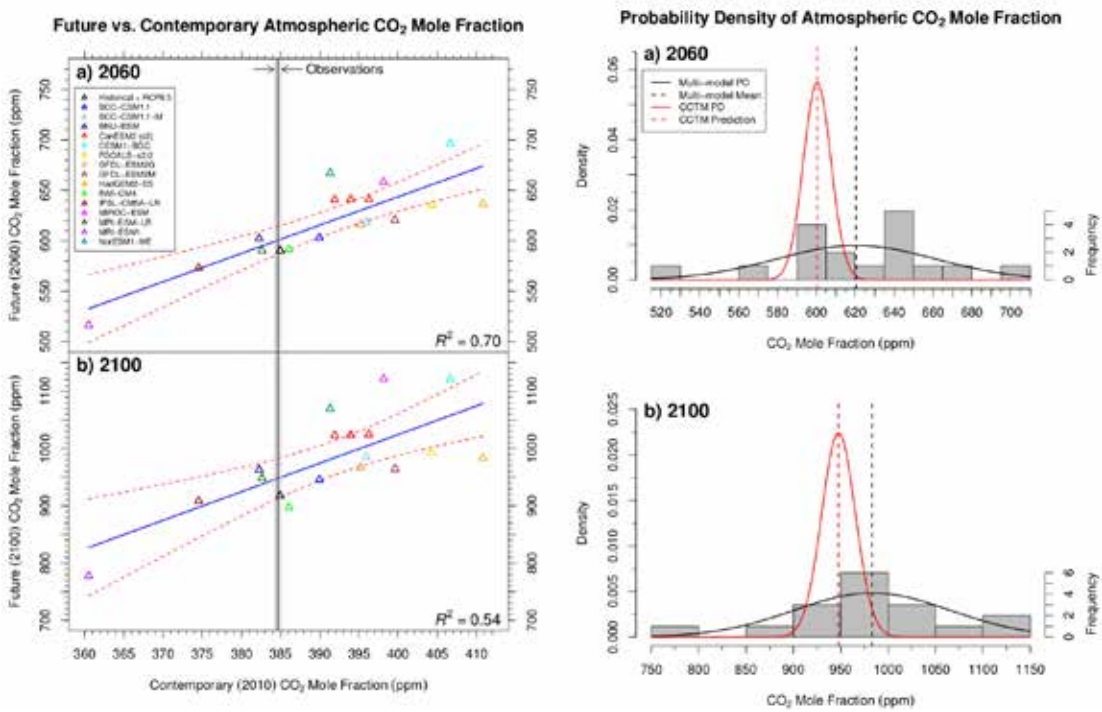


Figure B.3.1. Hoffman et al. (2014) found an emergent constraint based on carbon inventories (left, for (a) 2060 and (b) 2100) and applied it to constrain future atmospheric CO_2 projections from CMIP5 Earth system models, reducing both the mean and uncertainty range of CO_2 mole fractions (right, for (a) 2060 and (b) 2100).

B.3.2 New Metrics and Benchmarking Approaches

Incorporating atmospheric CO_2 observations with vertical resolution above the surface is an important goal for the benchmarking system that will permit disentangling transport-induced biases from the land (or ocean) flux biases the system is designed to constrain. Incorporation of isotopes of CO_2 will also permit accounting of the contribution from land and ocean fluxes. The $\delta^{13}CO_2$ data are available at 95 National Oceanic and Atmospheric Administration (NOAA) flask observing sites, with many time series extending from the early 1990s to the present.

Atmospheric CO2 likely plays a key role in emergent constraints because it integrates over the regional to global scales for which emergent constraints are most likely to provide value for future climate–carbon cycle predictions. Determining robust ways to use atmospheric data for emergent constraints should be an ongoing focus of discussion.

B.3.3 Observational Data Needs

Atmospheric CO_2 data are publicly available and observations from all platforms are tied to the World Meteorological Organization (WMO) calibration standards. Within the past decade, remote sensing observations of atmospheric CO_2 have gained prominence, and characterization of errors in observations have improved, especially from remote sensing. Observations to constrain the atmospheric transport operator are also crucial. Diagnostics for boundary layer depth, convective mixing, and horizontal advection would all provide insights into whether model–data mismatch

is due to carbon fluxes or atmospheric transport. Existing atmospheric CO_2 data are fairly well archived, with data publicly available from NOAA (surface), CDIAC (aircraft campaigns and TCCON), and NASA (OCO-2). Maintaining and growing these archives of surface and atmospheric profile measurements along with estimates of all anthropogenic emissions over time is critical to meet a variety research needs in a warming world. Availability of other observations, including satellite remote sensing, varies by agency.

B.3.4 Model Development and Output Requirements

CO_2 should be output at gridcell resolution with the vertical profile saved for comparison with aircraft campaigns, which occur in regions sparsely sampled by the long-standing surface network, and remote sensing platforms, which reduce potential model–data bias due to misrepresentation of vertical transport. Monthly frequency is the minimum temporal frequency, although process level insights could be gained by benchmarking diurnal and synoptic variations. Components of CO_2 in the atmosphere from land, ocean, and fossil fuel sources (kg/kg), and specific humidity (kg/kg) are necessary for column integration for comparison with satellite observations. Integration of isotopes of CO_2, including $\delta^{13}CO_2$ should be expressed per mil (‰).

B.4 Soil Carbon and Nutrient Biogeochemistry

Gustaf Hugelius, Jinyun Tang, and the International Soil Carbon Network (ISCN)

B.4.1 Introduction

Soils hold the Earth's largest biogeochemically active organic carbon (C) pool, which has the potential for a significant feedback to climate. At roughly 2,000 Pg C, this stock is more than twice as large as the atmospheric C pool (Ciais et al., 2013). Over time and large spatial scales, the soil C stock is determined by its turnover, which is a function of input from plant photosynthesis and losses via microbial decomposition, both of which are mediated by nutrient biogeochemistry. At present, global scale C inputs to soil are roughly balanced by losses to the atmosphere. However, because of its large pool size, even small changes in the soil C balance may cause significant increases in atmospheric greenhouse gas concentrations, contributing to additional climate warming. Since the start of the industrial era, soils have sequestered a significant fraction of CO_2 emissions from fossil fuel burning and human land use change (Ciais et al., 2013). However, under continued climate change and human intervention, soil C is expected to feedback with atmospheric C, and this balance may shift (Davidson and Janssens, 2006). Although urgently needed, quantification of how this balance may shift remains elusive, as many key processes that regulate the soil C stocks are poorly represented or missing in existing ESMs (Lehmann and Kleber, 2015).

B.4.2 Scientific Challenges and Opportunities for Model Evaluation

Broad-scale observations of soil C that span global environmental conditions are useful first order benchmarks for model predictions. For instance, observed global scale patterns provide undeniable evidence of the overarching climatic and biological controls on soil C and nitrogen cycling (Post et al., 1982; 1985). Thus, the degree of agreement between ESM predictions and observed global scale soil organic matter (SOM) patterns provides a baseline assessment of the ESMs' predictive power, even though the range of complex interactions and processes that control SOM cycling in models have not been assessed. The soil C stocks produced by current ESMs (CMIP5 models) are in only fair agreement with global soil C distributions, and the models are unable to reproduce local to regional scale spatial soil C patterns or to quantify bulk C stocks (Todd-Brown et al., 2013). Soil C variability in models can largely be explained by modeled net primary productivity (NPP), but observed soil C stocks cannot be explained solely by NPP and temperature. This model–data discrepancy is partly due to large C stocks in permafrost and peatlands where soil freezing or anoxia limits decomposition, resulting in large accumulations of soil C even under limited NPP. Permafrost and peat formation are examples of strong environmental controls on soil C turnover, and a model that does not address these controls cannot reproduce observed C stocks.

Therefore, ESM development should focus on improving the key controls on soil C turnover such as biogeochemical nutrient dynamics and environmental controls of microbial activity, suggesting that useful benchmarks for ESM soil C dynamics should target soil C turnover. Presently, basic soil nutrient biogeochemical processes are lacking or

insufficiently represented in many existing models, which causes models to behave inconsistently with data (Bouskill et al., 2014; Zaehle et al., 2014). Needed are improvements in modeling the cycling of nitrogen (N) and phosphorus (P) and their interactions with ecosystem productivity and decomposition through limiting plant photosynthesis or microbial processing of SOM. Modeled and observed soil C stocks should be analyzed in the context of both empirical and model data to understand processes affecting both NPP and soil C turnover times.

Soil C turnover in models has traditionally been conceptualized as a spectrum of pools linearly decaying with different turnover rates, which are modified multiplicatively by moisture and temperature effects (Parton et al., 1988). However, recent studies suggest that soil C decomposition across all ecosystems is an emergent response resulting from the interactions between many biotic and abiotic factors, including availability or activity of microbes, minerals, plants, and inorganic chemicals (Schmidt et al., 2011). This new conceptualization may explain why existing ESMs under-predict the climate change effect on carbon turnover (Carvalhais et al., 2014; Koven et al., 2015). Many new modeling approaches are also being explored to explicitly address interactive and emergent factors. Notably, studies show that considering the microbial and environmental dynamics in models e.g., improves global distributions of soil C stocks (Wieder et al., 2013), explains the diverse temperature sensitivity of C decomposition (Tang and Riley, 2015) and improves simulated respiratory response to soil moisture fluctuations (Grant et al., 2012a; Manzoni et al., 2014, 2016). Also, most soil biogeochemical models have only simulated the biogeochemistry in topsoil, but models are developed to resolve the vertical distribution and transport of SOM and they show improved model performance in recreating observed radiocarbon ages or C stocks at high latitudes (Braakhekke et al., 2014; Koven et al., 2013; 2015; Riley et al., 2014; Tang et al., 2013; He et al., 2016).

To date, model evaluations have focused primarily on whether models can reproduce observed time series or spatial patterns in observational data (e.g., soil C stocks). While such benchmarks provide initial insights into whether discrepancies exist, they offer limited insights into why models may or may not mimic observations. The next logical step is to break down the observed spatial and temporal patterns to identify key processes and environmental controls on model predictions. A model should be evaluated for what it was designed to simulate as opposed to what we wish it to simulate. For any given benchmarking activity, the targeted processes should be identified a priori and the empirical benchmarking dataset should be adapted accordingly. For instance, a model that does not include peatland formation should not be directly compared to datasets that include substantial stocks of peatland soil C. Other approaches include evaluating whether models can simulate ecosystem responses to disturbances, which could be either natural or manipulative. The emergent constraint approach is a non-traditional benchmarking method to evaluate and post-correct model performance (Hoffman et al., 2014), but its accuracy and mechanistic underpinning require further examination. Finally, to make the model–data benchmarking informative, benchmarking datasets should also include explanatory support data (metadata) and provide robust estimates of data uncertainties.

B.4.3 Observational Data, New Metrics, and Benchmarking Approaches

Despite its importance, observation-based estimates of the global soil C are highly uncertain. The estimates published between 1951 and 2011 (Scharlemann et al., 2014; median 1,460 Pg C, n = 27) have varied from 500 to 3,000 Pg C. With the recent release of the WISE 3.1 database (Batjes, 2016) the number was updated to 1,408 ± 154 Pg C to 1 m depth and 2,060 ± 217 Pg C to 2 m depth. The WISE database combines earlier products with climate maps and an updated soil profile dataset that integrates the global harmonized soil data with notable improvements at northern high-latitudes. At local to regional scales most modern soil inventories are based on digital soil mapping techniques where soil properties are predicted based on soil profile reference data in combination with environmental data. Hengl et al. (2014) first applied this technique at global scale and produced the SoilGrids 1 km dataset. Although digital soil mapping has many advantages when compared with other approaches, its product is still in early stages of development and needs further evaluation. Hengl et al. (2017) described the technical development and accuracy assessment of the most recent and improved version of the SoilGrids system at 250 m resolution, based on machine learning. Even with these recent advances, global soil C estimates still have large uncertainties, and regional discrepancies are high for wetland soils, tropical and northern peatland soils, and permafrost region soils. Broad-scale characterizations of these soil types are still hampered by pedon data scarcity, access restrictions (licenses), and insufficient data on their spatial distributions. Therefore, there are substantial remaining challenges for the research community working with improving and harmonizing mapping of global scale soil properties (Batjes et al., 2017).

While the community has not decided whether to replace established multi-pool models with models based on emerging conceptualizations of transient environmental and microbial dynamics within ESMs, disparate types of models can be evaluated with some common metrics. Examples include benchmarking model-estimated soil

C residence time with that from radiocarbon datasets and data–model experiments that target soil C responses to various environmental perturbations. Such approaches offer a way forward in comparing the performance of traditional and emerging models for a range of processes and across environmental gradients. Wieder et al. (2015a) present a framework for representing soil microbial processes in ESMs. However, formulating standard protocols for model parameters and output as well as common benchmarking approaches that are applicable across various model designs is a challenge that continues to provide opportunities for innovative ideas and cross-cutting discussions and collaborations.

Several challenges remain for next generation of soil biogeochemistry models. To meet these challenges, both model development and creation of dedicated benchmarking datasets are needed. First, how realistic is model representation of microbial dynamics? Is, for example, the microbial substrate-use efficiency, microbial community population dynamics or microbial and enzyme turnover appropriately represented? Microbial community responses to soil warming and changes in moisture are of particular interest. New experiments, including C-isotope labeling techniques (Dijkstra et al., 2011), for example, will be helpful to constrain these processes. Second, how realistic is model representation of soil mineralogy impacts on C stabilization across wide environmental gradients? Observed correlations between soil mineralogy and C turnover (Torn et al., 1997; Doetterl et al., 2015) could emerge from mineral interactions with, for example, dissolved organic substrates (Mayes et al., 2012), extracellular enzymes (Quiquampoix et al., 2007), root exudates (Keiluweit et al., 2015), or soil aggregates (Nicolas et al., 2014). Datasets are needed to parameterize and evaluate the representation of aggregates and abiotic destabilization effects on soil C dynamics across the full gradient of environmental conditions. High-quality observational datasets of soil mineralogy and soil textures across broad geographical scales are presently lacking. Third, how realistic is model representation of SOM stabilization and microbial activity across gradients from aerobic to anaerobic conditions as well as from frozen to unfrozen states? In response to hydrology, soils continuously fluctuate between aerobic and anaerobic conditions, and the two conditions often coexist at different soil depths (e.g., Grant et al., 2012b). Partial freeze-thaw dynamics of the soil column occur in both seasonally frozen and permafrost soils. Empirical data to support a mechanistic parameterization and evaluation of models with comprehensive redox cycles and dynamic soil freezing are needed. Model approaches that look beyond empirical scaled temperature and moisture responses may provide new ways forward in modeling these complex relationships (Davidson et al., 2012). Useful benchmarks to validate such models could be provided by laboratory incubations of full intact soil cores under varying thermal and hydrological conditions. Fourth, how realistic is model representation of soil transport and turbation processes? This includes bioturbation, cryoturbation, and other physical transport mechanisms. Only limited data are available to benchmark model performance, and observed radiocarbon ages of different SOM fractions across diverse environments are needed. Fifth, how realistic is model representation of nutrient dynamics and competition by microbes, plants, and mineral surfaces? These processes also feedback to plants and alter an ecosystem's capability to sequester atmospheric carbon. Many models that consider stoichiometric demand are limited to C:N dynamics, and increased understanding of C:P dynamics is desirable. Further, data availability limits mechanistic parameterization and the ability to assess models of nutrient competition (Tang and Riley, 2013; Zhu et al., 2016).

B.5 Surface Fluxes (Energy and Carbon)

A. Scott Denning and Daniel M. Ricciuto

B.5.1 Scientific Challenges and Opportunities for Model Evaluation

Surface fluxes of carbon and energy are a key input from land to atmosphere models, and observations of these variables have been used to benchmark carbon cycle, land surface, and Earth system models for several decades. Networks of surface flux observations such as the FLUXNET eddy covariance network have expanded rapidly over the last 25 years and have been used in numerous model intercomparisons and model–data comparison papers. Tools such as ILAMB can indicate when particular models may be agreeing with each other or with observations of surface fluxes, but, absent other benchmarks, cannot explain why they diverge in century-scale predictions. When different types of data are co-located, the benchmarks are even more powerful and should be given more weight. Intensively observed sites or regions, such as Critical Zone Observatories (CZOs), Long-Term Ecological Research (LTER) sites, or National Ecological Observatory Network (NEON) sites that include surface fluxes as part of a diverse set of measurements, may be candidates for a new subset of powerful "super-site" style benchmarks. Additionally, surface flux measurements in combination with experimental manipulations (e.g., warming experiments, rainfall exclusion, or CO_2 additions) may provide powerful constraints on ecosystem responses to climate change.

While eddy covariance measurements are critically important, their footprint (~1 km^2) is still 2–4 orders of magnitude smaller than that of a typical Earth system model grid cell (~104 km^2). Key process and driving variables of surface fluxes at these spatial scales may differ from those at the flux tower scale. It remains difficult to characterize soil, vegetation, and disturbance heterogeneity, and to estimate the effect of this heterogeneity on model predictions. "Bottom-up" approaches to upscaling use observations (e.g., FLUXNET) in combination with gridded driver datasets to estimate fluxes at regional scales. These already comprise a set of important ILAMB benchmarks, but more work remains to characterize associated uncertainties. Atmospheric inversion "top-down" models have progressed rapidly over the past two decades, increasing in resolution from continental scale to scales approaching that of Earth system model grid cells. While the global surface atmospheric CO_2 concentration measurement network remains relatively sparse and atmospheric transport uncertainty contributes to high estimated flux uncertainty, targeted regional networks and new remote sensing capabilities are beginning to enable predictions of surface CO_2 fluxes at higher accuracy and resolution. In the future, a combination of top-down and bottom-up techniques with data assimilation or model–data fusion approaches could produce integrated surface flux benchmarks that are more accurate and spatially relevant than individual approaches.

Specific Points and Recommendations

Measurements of surface exchanges of energy, water, carbon, and momentum at flux towers are uniquely valuable for evaluation of ESMs because these are precisely the quantities that must be provided by land-surface modules for successful coupling to the atmosphere. It is critical that ESMs continue to focus on getting the surface fluxes right, despite the aforementioned problems with heterogeneity and mismatched footprints. Benchmarking models against hundreds of surface flux records can help identify key model shortcomings and guide model development, but the value of these comparisons is greatest when the data are used to understand which processes matter at which spatial and temporal scales. Combining surface fluxes with other key benchmarks to understand their responses to changing climate conditions enhances mechanistic understanding of model deficiencies.

The mismatched footprints of flux towers and ESM grid cells have driven innovations in surface flux benchmarking. One approach involves model evaluation against suites of flux sites across gradients of climate drivers such as moisture or stand age. Upscaling from tower footprints has been done directly using field measurements and remote sensing to characterize spatial patterns and heterogeneity (e.g., Bigfoot Project: Cohen et al., 2003; Turner et al., 2003). Empirical upscaling of tower fluxes to produce global maps of surface fluxes by combining local observations with remote sensing and climate data is an especially promising direction for future model benchmarking (Luyssaert et al., 2007; Beer et al., 2010; Jung et al., 2011). Another important approach involves comparing models to measurements at much larger spatial scales using natural integrators of mass balance such as hydrologic watersheds or atmospheric mixing. Atmospheric measurements of trace gases provide a strong constraint for surface fluxes over large areas, but quantitative benchmarking requires accurate calculation of the effects of atmospheric transport through formal optimization techniques collectively known as inverse modeling. These methods have been used for CO_2 and other trace gases for decades, but have historically been limited by sparse CO_2 measurement networks. Recent developments in greenhouse gas observations from space (e.g., GOSAT, OCO-2) have the potential to dramatically improve ESM benchmarking at larger scales.

Benchmarking based on diurnal, seasonal, and even interannual variations in the recent past does not fully test the ability of models to predict future fluxes in response to climate forcing outside the envelope of recent changes. Unlike hindcasts, ESM *predictions* on decadal and centennial timescales cannot be compared to observations of changes that have not happened yet! Instead, we rely on model intercomparisons such as C^4MIP and CMIP5 to characterize the spread among models of the future. Intercomparisons provide a way to quantify uncertainty in production modeling, and classification of variations in ESM predictions relative to emergent constraints in hindcasts can help stratify models and provide guidance for model development (Hoffman et al., 2014).

B.5.2 New Metrics and Benchmarking Approaches

In addition to simple differences between models and observations, metrics should include separate evaluation of model bias, variance or RMSE, phases of diurnal and seasonal cycles, and spatial covariance. For mechanistic interpretation to propel model improvement, benchmarking should focus on characterizing functional relationships such as changes in surface fluxes with temperature and soil moisture anomalies. For ESMs to make credible

predictions, new benchmarks must quantify long-term responses to climate forcing, rather than just diurnal and seasonal behavior. While interannual variations are notoriously difficult to simulate accurately and very few flux tower records are long enough to characterize decadal variations, benchmarks that explicitly target these slower changes will be important in evaluating and improving decadal to century timescale ESM predictions.

B.5.3 Observational Data Needs

Surface fluxes of heat, water, carbon, and momentum are now routinely measured at more than 700 sites around the world, and flux data are available across an amazing breadth of climate and ecosystem types. Unfortunately, much of the data from these sites is difficult to obtain in a timely way. A number of national and regional networks contribute data to FLUXNET (**http://fluxnet.fluxdata.org/**), which performs high-level processing to fill in missing values and match flux data with other measurements, but flux records are often years behind real time. Moreover, updating of site records is uneven across the networks. These factors make development of benchmarks that relate flux anomalies to climate forcing or other data problematic. Combining flux data with remote sensing and other *in situ* observations (e.g., trace gas sensors or specialized phenocams) is possible, but is not done routinely.

Most flux towers have only operated for a few years, and only a handful have operated long enough to assess decadal changes in surface fluxes. To quantify responses of slower ecosystem processes, it will be critical to maintain the longest-running tower sites into the future, despite the cost and manpower challenges. The few 20-year records now available demonstrate the important roles of ecosystem succession and climate response. Predictive ESMs will be greatly enhanced if these long records can be captured in new benchmarks.

Integrated meta-analyses are required to enable evaluation of changes in surface fluxes from predictive models in response to forcing from climate, land use, and nutrient cycling. Combining flux records with other observations such as climate, remote sensing, land use, and disturbance histories provides the information modelers need to assess mechanisms for slowly changing fluxes. New syntheses can take advantage of ecosystem manipulations (e.g., Amazon throughfall exclusions, SPRUCE, and NGEE), leverage natural experiments (e.g., droughts, heat waves), and collect flux records across gradients in climate, land use, nutrient deposition, and stand age. Benchmarks using these synthetic analyses help indicate the sources of model discrepancies and lead to improved confidence in ESM predictions.

B.5.4 Model Development and Output Requirements

Current models include the calculation of albedo, partitioning of latent and sensible heat, transmittance of radiation to the ground, soil heat flux, and canopy temperature for some approximation of canopy heat capacity. Approaches to calculation of albedo and canopy radiation balance and heat storage vary widely, and evaluating how these different model frameworks calculate surface energy balance should be revisited in light of how second generation vegetation models now represent heterogeneity in plant canopies. In addition, ESM development is now addressing gaps in process representations that pertain to slower responses of ecosystems to changes in forcing, including ecosystem succession, nutrient cycling, and the effects of prolonged physiological stress. Plant mortality and replacement of plant functional types in response to climate change are critical processes that control ESM responses over decadal time scales, yet have typically not been included in ESMs.

B.6 Vegetation Dynamics

Rosie Fisher and Chonggang Xu

B.6.1 Scientific Challenges and Opportunities for Model Evaluation

In the context of this report, we define "vegetation dynamics" as the changes in ecosystem composition and structure—manifested in current ESMs as the distribution of plant functional types (PFTs)—in space, and of the processes leading to that distribution, including recruitment, succession, growth, mortality, and disturbance. In many LSMs, vegetation distribution is prescribed, and thus, vegetation dynamics metrics become a test of model behavior only when dynamic vegetation models (DVMs) make PFT distribution prognostic.

Development of Vegetation Demographic Models

Most land surface models now contain some kind of vegetation dynamics model, typically a first generation model, including Lund-Potsdam-Jena (LPJ)-derived models (in ORCHIDEE, CLM, CTEM), TRIFFID (in JULES), and the JSBACH-DGVM. The majority of CMIP models simulations do not actually include prognostic DVMs (they can typically be turned off and replaced with a static PFT distribution) because of challenges with increasing model degrees of freedom.

In first generation ESMs, the land surface is discretized into tiles, according to PFT, with each PFT represented by a single representative individual. The abstraction of ecosystems into this simplistic structure makes it difficult to simulate light competition, and, thus, exclusion or coexistence of different PFTs. In the last decade, second-generation vegetation demographic models (VDMs) have emerged that capture light-competition driven coexistence and competition of PFTs through the representation of different tree sizes (e.g., cohorts or individuals) in the vertical canopy structure and successional dynamics through the representation of disturbance history. One of these (SEIB-DGVM) is incorporated into an existing CMIP model, and development of several more VDMs is underway. VDMs allow comparison with many more potential data streams than first generation DVMs, and the sections below were written with this in mind.

Existing Large-scale Metrics for First Generation Vegetation Dynamics

In the first generation of ILAMB, the only vegetation dynamics metrics were for burned area. The GFED burned area product is used for comparison with models (Giglio et al., 2013). Hantson et al. (2016) reviewed the availability of benchmarking products related to fire in the context of the planned FireMIP experiment. They highlighted first the existence of four alternative burned area products (GFED3, L3JRC, MCD45A1, Fire_cci) and also the Global Fire Assimilation System biomass-burning fuel consumption product, which includes both fire and radiative power (Kaiser et al., 2012). Expansion of ILAMB to include these metrics would be beneficial to collaboration with FireMIP.

Most existing large-scale metrics of vegetation dynamics are derived from Earth observation measures of canopy greenness and algorithms that imply phenological type from seasonal cycles of canopy greenness (Lawrence et al.., 2012). Further, canopy height metrics allow distinction between short stature and low stature vegetation (trees/shrubs/grass). Both of these metrics can also be used to assess model projections of LAI and canopy height. Numerous alternative land cover maps exist (GLC2000, GlobCover, MODIS). For DVMs, it is traditional to compare model output with land cover maps generated from one or both of these products (Gotangco-Castillo et al., 2012). However, generation of land surface cover products (and their variation through time) is subject to uncertainty in both algorithm structure and PFT classification (Poulter et al., 2011). Integration of all such products into the ILAMB package would allow for characterization of the uncertainties across classifications.

Existing Plot-scale Metrics for Vegetation Dynamics

In the case of vegetation demography models, tree demography/forest inventory data at the site level have been used to compare with model simulations of recruitment, mortality, and canopy structure. Some early syntheses might be suitable for ILAMB integration, notably Forest Inventory and Analysis (FIA) program mortality rates gridded over the USA (Johnson, Xu, McDowell et al., in prep). There are numerous regional forest inventory datasets, but no comprehensive synthesis of these disparate products, meaning global-scale analyses are not possible at present.

B.6.2 Observational Data Needs

Observational data can be divided into two categories: (1) new data that is now available for use by first generation DVMs and (2) data that can be accessed by second generation (demographic) DVMs.

Forest Inventory Data

A critical but challenging source of data for VDM comparisons is the network of national and regional scale forest inventory data. These include FIA (USA), ForestPlots, ForestGEO, and many other national inventories (e.g., Spain, Russia). Data can be used to quantify mortality rates by PFT or size class, equilibrium and transient stand structure (height distributions), and relations among all these and driver variables, plant properties, and changes through time (e.g., van Mantgem et al., 2009). The major challenges here are analysis of the complex raw data, which is routinely conducted for small-scale analyses, and comparison across networks, which is rarely undertaken. This is a long-standing but important challenge (e.g., Purves and Pacala, 2008). Further challenges to the use of inventory plots are the typical absence of model drivers (meteorology) and auxiliary data (soil, plant traits) at individual sites, making direct comparison with models difficult, although this can potentially be overcome by concentrating on cross-network analyses and variable relationships, such as growth/mortality relationships through space and time.

Representation of Functional Diversity and Use of Trait Data

A further development in the LSM community is a proliferation of methods that seek to better capture diversity of plant function via the increasing use of plant functional trait data. These approaches include (1) using trait maps or trait-environment relationships to constrain LSMs (where trait information is an input) (Verheijen et al., 2013; Reich et al., 2014); (2) using optimality models to predict plant traits under given conditions (Xu et al., 2012; Thomas and Williams, 2014), and (3) trait filtering, where plants of different functional types compete within a demographic model (Scheiter et al., 2013; Fisher et al., 2015). For these latter two methods, geographical distribution of plant traits (which is increasingly available from remote sensing data) might be considered a metric or benchmark. Thanks to recent, very large databases of plant traits (Kattge et al., 2011), there has been much progress recently in identifying relationships between plant leaf traits (Wright et al., 2004; Reich et al., 2014). Depending on the choice of model, these data can either be used as input (to trait maps and climate-environment relations, or trait-trade off relationships) or as validation of the geographical distribution of traits predicted by optimal or trait filtering approaches. Despite the abundance of data for the most easily measured traits, such databases are only sparsely populated for many functional variables, in particular for belowground plant properties and for more physiologically complex processes (plant hydraulics information, tissue allocation, carbohydrate storage).

Remote Sensing Products

Remote-sensing based disturbance maps could be useful for benchmarking severe mortality events (e.g., fire and insects; Hansen et al., 2013). Such products are more useful for benchmarking if they attribute the disturbance to different causes of death (fire, deforestation, drought stress, insects/disease). Dynamics of vegetation heights based on LiDAR sensors could be useful to detect the disturbances too. With a demography size-structured model, however, linking height retrievals to model size-class representations introduces elements of uncertainty, particularly if retrievals are only available for the tallest trees. be able to provide the tallest tree info? This is particularly true for new global LiDAR products such as Global Ecosystem Dynamics Investigation (GEDI). Finally, the remote-sensing based functional relationship between traits and vegetation dynamics (e.g., trait distribution vs. mortality rates) could be useful for the third generation of vegetation models.

Paleo and Tree Ring Data

Forest inventory data has time scale limitations. Thus, it would be beneficial to use pollen records to indicate past vegetation distributions (e.g., PalEON for North America; **http://www3.nd.edu/~paleolab/paleonproject/**). It would also be useful to compile the tree ring data across the world for the prediction of tree diameter growth under past climate conditions.

Variable-variable Relationships

Thurner et al. (2016) generated a global product of the plant productivity divided by the estimated carbon stocks. The result is an estimate of carbon residence time, which, although not precisely a metric of mortality, is comparable to the identical model metric and can potentially be used not just for DVMs but also for static vegetation distribution models.

B.6.3 New Metrics and Benchmarking Approaches

In terms of the metrics of benchmarking, it would be beneficial to use the traditional bias and RMSE as score metrics; however, metrics related to the successional trajectories (e.g., basal area and density change through time) with different types of disturbances could be useful to constrain the overall behavior of models. Furthermore, for the demographic type of DVMs, it would ideal to have metrics on the distribution of size and height on the same grid cell, given that it is important to correctly simulate both the mean and distribution to capture the vegetation dynamic under future novel climate conditions.

Appendix C. Metrics for Integrating and Cross-cutting Themes

C.1 Process-specific Experiments

Mathew Williams and Jianyang Xia

In this section we discuss how process-specific experiments—that is detailed lab or field based studies—can provide critical parameters or insights into improved model structure.

The key scientific priority is selecting a group of sites from FLUXNET that span major biomes to serve as testbeds for ILAMB. Each of these sites should have associated data provided (e.g., met forcing, soil texture, land use history, plant traits) to allow model runs over specified time periods. Each site would have a series of independent datasets (e.g., net fluxes, biometrics and experimental data), allowing a careful diagnosis of model process representation. Below we set out the more detailed requirements and activities.

C.1.1 Scientific Challenges and Opportunities for Model Evaluation

It has been widely suggested that Earth system models should be made more robust by improving their structures to represent more real world processes (Knutti and Sedlacek, 2013; Luo et al., 2016). Given the enormous complexity of Earth system processes, it is still challenging to (1) specify which processes are more critical than others in regulating Earth system dynamics, such as climate change; and (2) evaluate representation of processes that have been widely incorporated but diversely parameterized in different models. One promising approach to solve this challenge is using process-specific experiments, which can evaluate and improve the model representation of a specific key process with observations. In this section, we identify a range of key processes where current models are highly parameterized or have major structural uncertainties. This identification then allows targeted links to process-specific experiments for tackling knowledge gaps in the following areas:

- **Decomposition**: Coupling to plant process, particularly priming through microbial dynamics
- **Nitrogen cycling**: Organic uptake, fixation largely unmeasured, not included in models, but likely to be critical
- **Autotrophic respiration**: Fundamental controls are poorly known, climate sensitivity is a major question
- **Fluorescence**: How can these data, soon to be available from space, be used to evaluate canopy processes?
- **Phenological sensitivity to climate**: The model response of plant canopies to changes in precipitation, CO_2, and temperature lacks strong foundations
- **Plant trait correlations and trade-offs**: Trait data are more available, but the trade-offs between traits must be better incorporated into models.

C.1.2 New Metrics and Benchmarking Approaches

Experimental approaches for addressing the key process uncertainties listed above involve using models to simulate processes at selected eddy flux sites, so that direct comparison to local data for process diagnostics are possible. This requirement means that the necessary drivers for all selected sites must be synthesized and distributed.

- **Decomposition**: Priming studies using varied litter quality to monitor microbial responses, time series of soil respiration (trenching experiments would allow more direct monitoring of heterotrophic respiration). Evaluate modeling of decomposition dynamics, climate sensitivity, and litter quality sensitivity.

- **Nitrogen cycling**: C:N ratio for all pools, 15N tracer studies to quantify uptake, allocation, and turnover. Evaluate modeling of N pools and dynamics.
- **Autotrophic respiration**: Plant tissue respiration measurements, links to whole-plant economy, C isotope tracer experiments. Evaluate capacity of models to distinguish between growth and maintenance respiration for various plant pools, and their seasonal patterns.
- **Fluorescence**: This quantity needs to be co-observed with eddy flux data to allow direct relations to gas exchange to be evaluated. There are issues with representativeness when comparing site to satellite data. Evaluate leaf level process representation in models.
- **Phenological sensitivity to climate**: Models could usefully provide output of leaf out date and senescence date that would be comparable to remote sensing indices. Below-ground phenology is a major uncertainty, so rhizotron data would be valuable. Information on non-structural carbohydrate can inform on plant allocation potential. Evaluate phenological timing against local data.
- **Plant-trait correlations and trade-offs**: Use local trait data to calibrate and evaluate models.

C.1.3 Experimental/Observational Data Needs

Field experimentation is a useful approach to explore new mechanisms underlying Earth system changes (Medlyn et al., 2015; 2016). However, there are challenges to connecting experimental data to models due to scale mismatches and gaps in records. Hence the need for carefully constructed driver and evaluation datasets at selected sites for developing diagnostics of model process representation. There are clear areas for novel experimental focus, particularly around isotopic tracers and fluorescence.

Gaps in Current Data Availability

There are difficulties in accessing experimental data in forms of value for model calibration and evaluation. Likewise, climate forcing for experimental data are often unavailable. The measurements are usually non-consecutive, and only a few variables or processes, e.g., soil respiration, are measured with standardized tools among different sites. Isotopic data remain relatively rare, but offer opportunities for tracing flows of C and N, allocation and residence times (Trumbore, 2006).

New *in situ* or Remote Sensing Measurement Needs

In situ experiments should focus on isotope tracer studies that quantify the residence time and pathways for N and C in ecosystems. Leaf and canopy scale studies of fluorescence are needed to inform use of satellite data (Guanter et al., 2014; Yang et al., 2015). Measurement of non-structural carbohydrate can inform on how plants invest and hedge against risk. It is highly valuable to have *in situ* remote sensing data over instrument sites, for comparison with satellite observations. Drone based sensors now make it possible to record similar data to that collected by satellite sensors, and thereby to determine atmospheric, scale, and spatial location errors between platforms.

Spatial and Temporal Extent and Resolution Requirements

There is a need for detailed *in situ* evaluations of model processes to test and parameterize models consistently; this means being able to isolate specific model processes so their decoupled sensitivity to particular forcing (experiments) can be evaluated and calibrated. Temporal requirements are closely related to residence times of carbon pools. Data extending over years are critical for understanding dynamics of the long-lived soil and wood pools. Weekly data are needed to track key phenological events. Hourly data provide insights into leaf level processes and sensitivity. We note significant data gaps in tropical ecosystems (Schimel et al., 2015) leading to major unknowns in the C cycle of these biomes.

Integrating Extant Meta-analyses into Benchmarking Approaches

Meta-analyses of field experiments results have been recently used for benchmarking terrestrial ecosystem models (e.g., Piao et al., 2013). Plant trait databases are growing and providing important data on plant traits (Kattge et al., 2011). Their focus is mostly on leaf traits, particularly structural traits. These databases will become more valuable as they include broader plant traits, and functional traits (e.g., respiration determinants, carboxylation rates). We particularly need to understand trait trade-offs, and use these to guide model parameterisation and structural improvements. We need to be able to simply characterize response patterns of different C and N processes for benchmarking model response functions.

Synthesis Activities Needs and Approaches

Exploration of full economic modelling for C allocation and C-N linkages provides a means to introduce optimality constraints on biological processes consistent with competitive interactions (Thomas and Williams, 2014). Effective modeling of plant-microbe-soil interactions, addressing priming, N fixation, exudates among other processes (Wieder et al., 2013), requires a concerted experimental effort, and particularly the use of isotopic tracers to unravel belowground processes.

C.1.4 Model Development and Output Requirements

For model development we require testbeds for calibration and evaluation of submodels at site scale, allowing simple connections between model inputs/outputs and site data. We need to evaluate plant trait correlations to determine process trade-offs (e.g., wood density versus hydraulic resilience). There is a risk that model development adds parameters and complexity, but thereby does not reduce model error and bias. This risk can be overcome by consistently testing simple models against data, and determination of the information content provided by more complex parameterizations (Li et al., 2014).

For output requirements, we need residence times for all pools, allocation and turnover of foliage, microbial pool dynamics, respiration of all living pools, trait correlations, N dynamics (including biological fixation). The biogeochemical data can then be used to evaluate model dynamics across pools and timescales (Thomas et al., 2013).

C.2 Metrics From Extreme Events

Hyungjun Kim and Maoyi Huang

C.2.1 Scientific challenges and opportunities for model evaluation

In the context of ILAMB, we define extreme events as the terrestrial and societal impacts (e.g., floods, streamflow and soil moisture drought, vegetation dieback, and fire) of weather and climate extremes (WCEs), and their feedbacks to the atmosphere. The WCEs are estimated as the occurrence of a value of a weather or climate variable above (or below) a threshold value near the upper (or lower) ends ("tails") of the range of observed values of the variable (Seneviratne et al., 2012). Also, WCEs are identified as single and compound events. The latter occurs when (1) two or more extreme events occur simultaneously or successively, (2) combinations of extreme events lead to conditions that amplify the impact of the events, or (3) combinations of events that are not themselves extremes but lead to an extreme event or impact when combined. For example, floods most likely occur when heavy precipitation falls over saturated soils, so that it is desirable to analyze precipitation and soil moisture extremes simultaneously. One special case of compound events is associated with feedbacks within the climate system such as the possible mutual enhancement of droughts and heat waves in transitional regions between dry and wet climates that can be attributed to the interactions among soil moisture, surface energy budget partitioning, and near-surface temperature (Seneviratne et al., 2010).

Infrequent extreme events may play a particularly important role in structuring terrestrial ecosystems, for example in controlling severe fires and contributing to drought-related vegetation mortality events (Figure C.2.1). Thus, it is necessary to include these long-term effects and their role in governing vegetation dynamics. Current models,

particularly those that do not have a dynamic vegetation component, only represent short-term responses to WCEs, such as depressed growth during the period of the WCE. However, datasets to benchmark these long-term ecosystem responses to WCEs are sparse, and the framework to test ecosystem model responses to WCEs is not well developed.

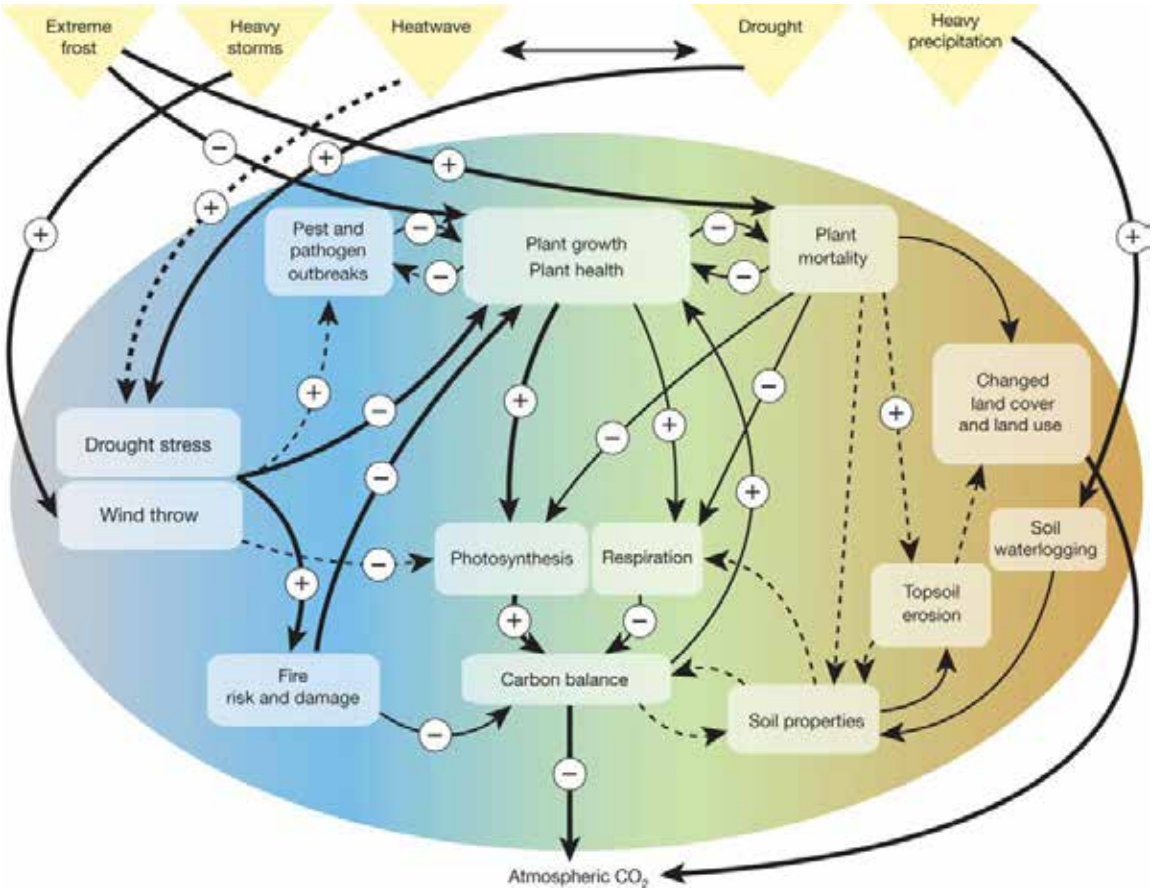


Figure C.2.1. Processes and feedbacks triggered by extreme climate events, including droughts and heatwaves, heavy storms, heavy precipitation, and extreme frost. Solid arrows show direct impacts; dashed arrows show indirect impacts. The relative importance of the impact relationship is shown by arrow width (broader lines indicate stronger feedbacks). Adopted from Reichstein et al. (2013).

To distinguish causal processes of extremes and to evaluate how they are well represented in a model, we suggest a logical framework to categorize them into different spatiotemporal scales and scopes of their footprints and impacts, and list examples which have relatively large uncertainties or are missing representations in current ESMs.

A. Climate scale features: Macro-scale features having long persistence (> seasonal) and large horizontal length scale (> 2,000 km), such as the spatial distribution and intensity of SST anomalies (e.g., El Niño and other climate modes), locations of ITCZ on meridional migrations, intensity of Hadley circulation, and latitudinal temperature gradient

B. Synoptic and mesoscale features: Persistence up to seasonal time scale and continental scale in the spatial domain, such as monsoons, tropical/extratropical cyclones, frontal systems, and sand/dust storm, as well as their impacts, such as excessive precipitation (i.e., meteorological drought) and heat/cold waves

C. Basin-scale land processes: Processes spanning up to seasonal or sub-seasonal scale such as excessive deficits and surpluses of water (e.g., flood), dry (i.e., hydrological and ecological droughts)/wet spells, extreme sea level, cryosphere- and ecosystem-related impacts (snow and snowmelt, fire, vegetation dieback), and landslides

D. Socioeconomic impacts: Processes which are directly related with human-society, such as inundation and crop failure (i.e., agricultural drought)

C.2.2 New Metrics and Benchmarking Approaches and Observational Data Needs

Considering the potential objects listed above to be benchmarked, we propose several metrics on WCEs below:

A. ITCZ displacement: Meridional distance of ESM simulated ITCZ location from atmospheric reanalysis datasets or satellite observations (e.g., QuikSCAT). Location of the ITCZ is defined as places where the temporal mean of the meridional component of surface wind (v) is zero.

B. Zonal shift of Walker circulation: Zonal displacement from ascending/descending kernel locations of atmospheric reanalysis datasets. Convergence and divergence of near surface (e.g, 950 mb) and high atmosphere (e.g., 300 mb) and 500 mb pressure velocity will be used to identify the kernels.

C. Reproducibility of weather systems: Skills of ESM representations of weather systems in terms of geographical location, intensity, and duration. Objectively detected weather systems (Utsumi et al., 2014) generated by ESM will be evaluated through comparison with observations (e.g., best track records for tropical cyclones; Utsumi et al., in revision; Figure C.2.2) and/or objective detections based on atmospheric reanalysis datasets.

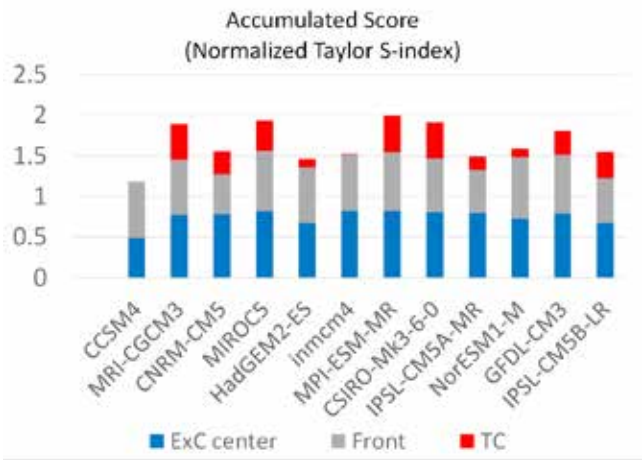


Figure C.2.2. Benchmarking for weather system reproducibility of CMIP5 models.

D. Hydroclimatic intensity: Giorgi et al. (2011) suggested an index to estimate the intensity of hydroclimatic cycles as a ratio of mean precipitation intensity and mean dry spell length. ESM-reproduced precipitation intensity and temporal variability will be validated by using an observational precipitation-based index for each model gridcell.

E. Flood inundation extent and duration: ESM calculated inundated area will be compared with satellite-based surface water extent (Prigent et al., 2007). ESMs without the inundation process can utilize an off-line method using a standalone river model (e.g., CaMa-Flood; Yamazaki et al., 2011; Figure C.2.3) to validate their runoff generation. The anomaly of water storage combined with the other components (e.g., soil moisture) can be compared with the terrestrial water storage anomaly monitored by the GRACE satellite (Kim et al., 2009; Figure C.2.4).

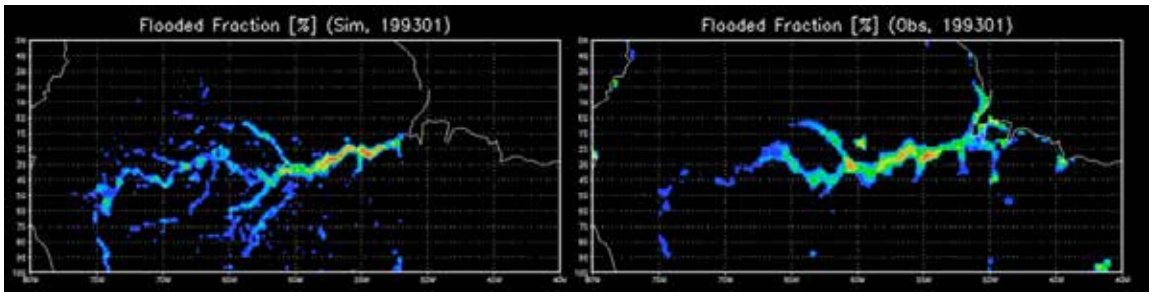


Figure C.2.3. Comparison of flood inundation extent over Amazon by CaMa-Flood (left) and satellite remote sensing (right).

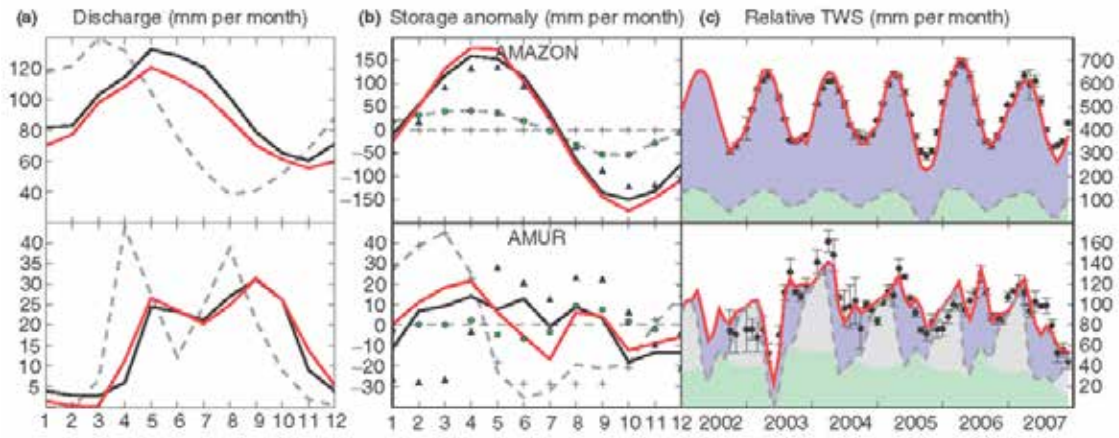


Figure C.2.4. (a) Comparison of seasonal cycles of observed GRDC discharge (black solid line), discharge routed by the Total Runoff Integrating Pathways (TRIP) model (red solid line), and runoff without routing (gray dashed line). (b) Comparison of seasonal cycles of GRACE TWSA (black solid line), simulated TWSA with river storage (red solid line), simulated TWSA without river storage (gray dashed line), and the major water storage components in TWS. Gray crosses, green circles, and blue triangles represent snow water, soil moisture, and river storage, respectively. (c) Inter-annual variations of relative TWS: GRACE observation (black dot), and the TWS simulations with river storage (red solid line) and without river storage (gray dashed line). Each area shaded by blue, gray, and green indicates the portion of river storage, snow water, and soil moisture in the simulated relative TWS, respectively.

F. Cumulative rainfall deficit: Maeda et al. (2015) suggested combining GRACE observations with *in situ* river discharge data to estimate water storage deficit on a basin-scale. The deficit reflects a cumulative amount of precipitation needed to satisfy evapotranspiration requirements through consecutive months (Figure C.2.5). A combination of ESM-simulated precipitation and evapotranspiration will be compared to benchmark how the model properly represents the intensity and the duration of dry spells.

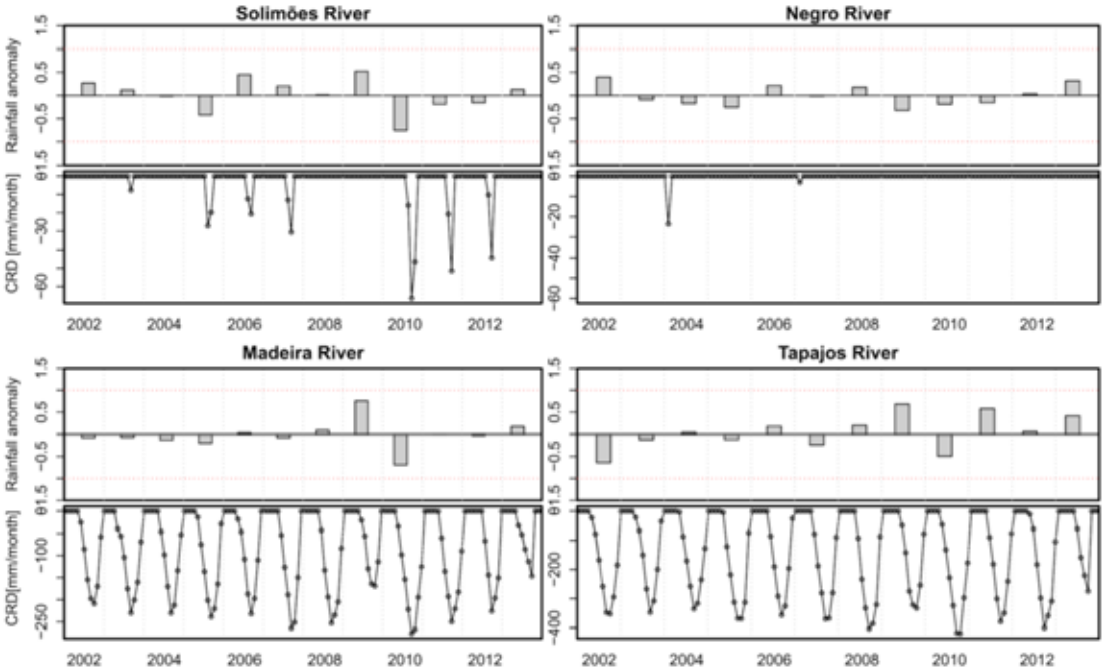


Figure C.2.5. Cumulative Rainfall Deficit and annual rainfall anomalies in four watersheds over the Amazon basin.

G. Event oriented benchmarks: Compile standard dataset libraries for well-studied extreme events for comprehensive benchmarking through multiple state and flux variables between onset and offset of the extremes. A 2003 heatwave in Europe, California drought, Alaska fire events, 2010 Russian drought, and 2011 flood in Australia would be candidates. Crucial to the use of naturally-occurring WCEs as model benchmarks is to compile both the short-term water, energy, and carbon responses of the coupled ocean–land–atmosphere system during the event, as well as the longer-term responses of ecosystems and anthropogenic systems to the extreme events, including vegetation mortality responses to drought and heat events, and soil and vegetation carbon losses during fires. Ideally, such observations can be collected in cases where some medium-term predictability allows installation of dense observing systems prior to or during the WCE, for example in examining ENSO-related drought events which may allow several months of predictability about where droughts are likely to occur, which has been a strategy of the NGEE Tropics project for the 2015–2016 El Niño event.

H. Experimentally-induced WCEs: Numerous rain throughfall exclusion experiments have been conducted in terrestrial ecosystems to simulate drought events, and these are a useful benchmark of terrestrial models (e.g., Fisher et al., 2007; Powell et al., 2013). These experiments, and other experimentally-induced WCEs, allow for targeted measurement campaigns and collection of key variables required for testing models, which may not be possible given the opportunistic nature of observational campaigns around naturally-occurring WCEs. Synthesizing these experiments and developing clear model protocols for comparison is a key requirement for better use of these experiments as model benchmarks.

C.3 Design of New Perturbation Experiments

Martin De Kauwe and Ankur Desai

Breakout Meeting attendees: James Simkins, Shawn Serbin, Rosie Fisher, Elena Shevliakova, Ben Bond-Lamberty, Dan Ricciuto, Nick Smith, Kaoru Tachiiri

C.3.1 Scientific Challenges and Opportunities for Model Evaluation

Perturbation experiments directly manipulate ecosystems and by measuring observed responses against a control, they provide direct tests of ecosystem responses to land use and global change (Bonan, 2014). Manipulation experiments short-circuit long-term monitoring experiments and directly test the global changes that ESMs are expected to predict. Despite this, these experiments have been under-used in evaluating ESMs predictions. There are a number of reasons for this disconnect: (i) there are often scale mismatches between the (coarse) model and the experiment; (ii) datasets from experiments are not in a format which can readily be used by modellers; (iii) the necessary meteorological forcing for the experiment may not exist, or may have gaps; (iii) there are data-sharing issues; and (iv) the modelling and experimentalist communities are not sufficiently engaged. Furthermore, attempts to model experiments have traditionally taken place after the conclusion of the experiment (but see Luo, 2001; Parton et al., 2007; Medlyn et al., 2016), which often results in missed opportunities to take measurements that could have distinguished between competing model hypotheses (Dietze et al., 2014).

Field manipulations encompass a broad range of experiments including: nutrient addition/removal, species transplant (addition/removal), precipitation and temperature manipulation, rainfall exclusion, manipulation of atmospheric chemistry and greenhouse gases. Arguably the most well known example of which were the US Department of Energy Free-Air Carbon Dioxide Enrichment (FACE) studies, carried out between ~1996–2010 (Figure C.3.1). For logistical reasons many of these experiments often manipulate a single factor, although a smaller collection of multi-factorial experiments do exist (Dukes et al., 2005; Pendall et al., 2013; see Dieleman et al., 2012 for a review). Such experiments need to be conducted outside of mid-latitude biomes, and the Amazon FACE experiment (**https://amazonface.org/**; Grossman, 2016) is expected to provide valuable information about photosynthetic potential of tropical forests. Ongoing and new studies that look at multiple factors in critical ecosystems such as peatland warming and drying (SPRUCE; **http://mnspruce.ornl.gov/**; Witze, 2015; Figure C.3.2), drought, nutrient addition, active warming, including the Tropical Responses to Altered Climate Experiment (TRACE; **http://forestwarming.org/**; Cavaleri et al., 2015; Figure C.3.3), or passive warming, including the International

Tundra Experiment (ITEX; **http://ibis.geog.ubc.ca/itex/**) and the Zero Power Warming (ZPW) experiment (**https://www.bnl.gov/envsci/test/zpw-liveupdates.php**) have high potential for constraining ecosystem model responses in ways that are difficult to do with traditional benchmarks from long-term observations. There are also a new generation of FACE experiments focused on mature ecosystems, which cover a wider range of biomes and climatic space than the first generation did (Norby et al., 2016).

Figure C.3.1. Four rings at the Oak Ridge National Laboratory FACE experiment.

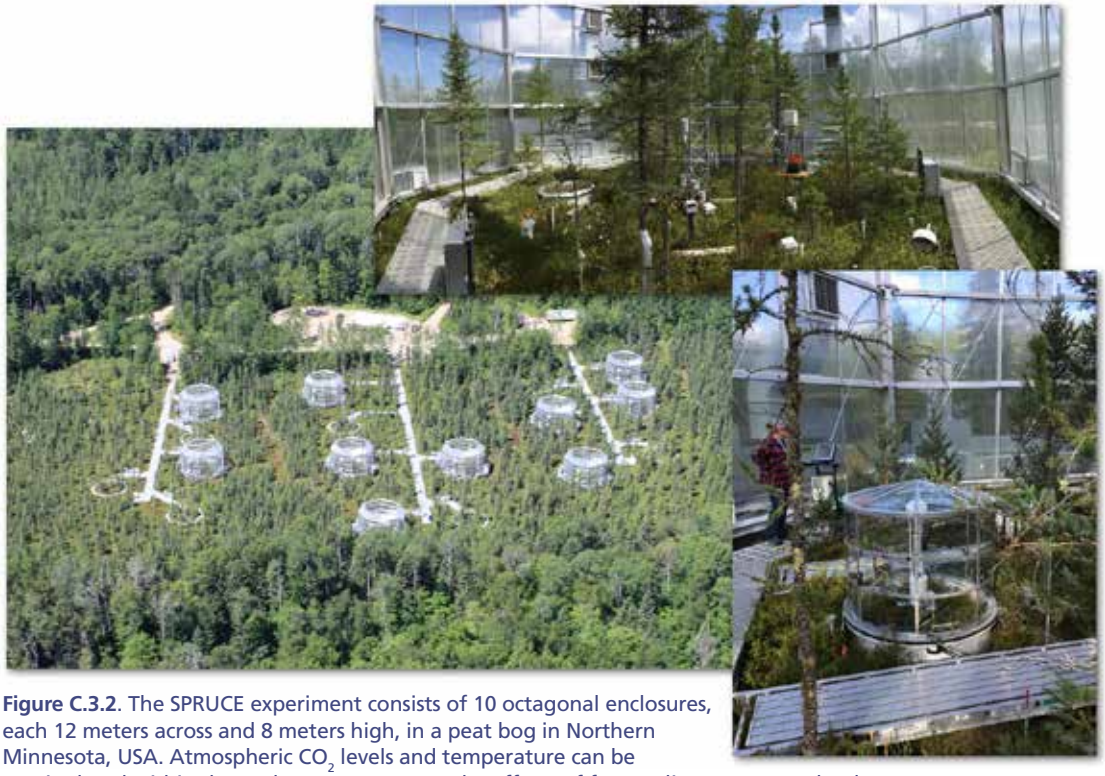


Figure C.3.2. The SPRUCE experiment consists of 10 octagonal enclosures, each 12 meters across and 8 meters high, in a peat bog in Northern Minnesota, USA. Atmospheric CO_2 levels and temperature can be manipulated within the enclosures to test out the effects of future climates on peatland ecosystems.

Figure C.3.1. Four rings at the Oak Ridge National Laboratory FACE experiment.

C.3.2 New Metrics and Benchmarking Approaches

Most benchmarking approaches for perturbation studies do not differ significantly from traditional benchmarking, though the focus is on comparing model sensitivity to response of the perturbation over the control, for the target variable and driver change. To date, most comparisons have not exploited this approach. For example, model comparisons to FACE data have often focused on capturing the mean net primary productivity (NPP) response over the experiment period (Hickler et al., 2008). This is problematic because models can be tuned to get the right answer, but arrive at it for the wrong reasons. Alternatively the FACE model–data synthesis (De Kauwe et al., 2013; 2014; Walker et al., 2014; Zaehle et al. 2014) used the experimental data to understanding how and why models differed from each other and the observed responses, providing a clear roadmap to model improvement (Medlyn et al., 2015).

There are a number of existing experiments that we identified which as yet have been under-exploited for model benchmarking. These include: (i) warming; (ii) drought/rainfall exclusion (see Smith et al., 2014 for a review); (iii) competition changes (species composition); and (iv) acclimation responses. It is likely that additional funding will be required to synthesis past experiments to define big picture responses we feel models should be capturing.

Due to the small scale of manipulation experiments, it may be that the best route for benchmarking ESMs remains in targeted offline model intercomparison projects. It may also be possible to use these results as a set of response surfaces to benchmark future climate model runs in an emergent constraint framework (Hoffman et al., 2014), by estimating processes such as tissue turnover rates, or recovery times from disturbance. Nevertheless, now that many of these experiments have been completed there is scope to define a series of cross-site responses that could be used to define a benchmark for ESMs (Walker et al., 2015). One successful example of this approach was the use of data across 301 N-fertilization experiments to confront global predictions from two land surface schemes (Thomas et al., 2013). Another example was the use of a long-term throughfall exclusion experiment in the Amazon to probe how well models captured responses during drought (Powell et al., 2013).

C.3.3 Observational Data Needs

While perturbation experiments often collect extensive field-level data, much of these data are difficult to acquire and integrate. Many data, such as those on leaf-level parameters, NPP are stored in diverse formats (e.g., xls vs. csv vs. netCDF), often not open-source , rarely directly machine-readable and on archives that may require permissions to access. Metadata and protocol documents may not specify treatment details in sufficient detail to properly replicate in a model. For example, exactly how much biomass was removed and what was done with this biomass: was it removed from the site, or deposited as litter; the distinction matters for models wishing to replicated experiments. There remains an outstanding need to provide funds to experiments so that the datasets produced are open, self-describing and useable by outside groups (e.g., Bond-Lamberty et al., 2016a). Even in experiments which have readily shared datasets (e.g., FACE), often the experiments do so via site-specific websites and these datasets often lag datasets used in recent publications. There is a clear need for central archival repositories for manipulation experiments. This would not only help in distributing datasets, but would most likely also raise awareness to the wider community.

Model–data synthesis initiated at the start of experiments (Medlyn et al., 2016; Norby et al., 2016) is an excellent means to identify and solve many of these potential issues before the experiment begins. These synthesis activities can also lead to the development of experiment modelling protocol which direct other modelling groups how to set models up for individual experiments (Walker et al., 2014; Medlyn et al., 2016).

C.3.4 Model Development and Output Requirements

Several challenges exist in attempting to apply models to perturbation experiments. Many ESMs operate at relatively large scales (>50 km^2), whilst experimental plots may be relatively small in spatial size (1–100 m^2). In particular, for the core global change processes of CO_2 fertilization, drought, and nitrogen addition, mechanisms are limited in variety. Thus parameterizing a model or scoring performance against these experiments when mechanisms are not nuanced enough to address the main responses remains a challenge. It is important for benchmarking to consider multiple aspects of a response (e.g., biomass growth, allocation changes, water use, mortality rates) to reliably score a model against such an experiment. Otherwise a model may perform well in one area, but without the proper mechanisms, incorrectly capture other dynamics.

Models need to be able to run control and perturbation studies and produce output on the difference between these two across multiple types of variables that are measured on the ground, including soil respiration, NPP, transpiration, allocation, and root growth. A specific challenge may be properly specifying the actual treatment. While some like CO_2 fertilization or N deposition are straightforward, others like soil warming or biomass removal may require model modification to properly simulate the experimental protocol, if it affects the response. Properly specifying initial conditions and species specific parameters is also critical to properly simulate plot-level and ecosystem-scale studies.

Benchmarking applications need to consider comparing not just time and space overlapping state variables, but also comparisons of responses grouped by ecosystem function or structure. The benchmarking community should work jointly with experimentalists to identify a set of shared priorities for evaluation and experiments best designed for addressing those.

C.4 High Latitude Processes

Charlie Koven, Kevin Schaefer, and Umakant Mishra

C.4.1 Scientific Challenges and Opportunities for Model Evaluation

Northern high latitude soils contain about twice as much carbon as in the atmosphere (Hugelius et al., 2014). This enormous carbon pool is vulnerable to accelerated losses through mobilization and decomposition under anticipated warming scenarios, with potentially large global carbon and climate impacts (Koven et al., 2011; Schaefer et al., 2011; Schuur et al., 2015). Many processes control the response of this carbon pool to changing environmental conditions. For example, active-layer dynamics, thermokarst formation, thermal erosion, shrub expansion, fire disturbance, soil moisture heterogeneity, and the overall rate of wetting and drying that will accompany warming. These processes impact the vulnerability of permafrost carbon pool through different mechanisms. Active layer thickness determines the volume of SOC available for microbial decomposition, and has been projected to go deeper under future warming. Thermokarst formation on the permafrost landscape enhances methane emissions to the atmosphere. Thermal erosion due to permafrost collapse can increase microbial decomposition and translocate large amounts of soil carbon to river networks. Increased wildfire occurrence has been projected under future warming scenarios; wildfires can directly combust the carbon in the surface organic layers and may alter the soil moisture dynamics. Similarly, many studies projected shrub expansion northwards under future warming, which can further destabilize the existing permafrost.

The CMIP5 generation of models were still deficient with respect to their ability to simulate these processes. None of these models included permafrost carbon pools, many had poor representation of crucial physical processes such as snow insulation of organic soil physical properties (Slater and Lawrence, 2013; Koven et al., 2013), and none included a careful treatment of subgrid-scale heterogeneity in landscapes driven by polygonal features. Since then, research on modeling high latitude dynamics and creating observational benchmarks for these models has resulted in significant progress in the field. Some ESMs, including CESM and ACME, now represent permafrost carbon and nutrient cycle processes (Koven et al., 2015), while others have focused on dynamic organic layers (Yi et al., 2009), or high-latitude-specific vegetation dynamics (Euskirchen et al., 2009). A MIP led by the Permafrost Carbon Network (PCN) compared different representations of the high latitude system to identify the effects of different structural representations on model predictions (McGuire et al., 2016). Activities like PCN have focused on synthesizing existing datasets on soil carbon stocks (Hugelius et al., 2014; Harden et al., 2012); permafrost carbon

decomposability under oxic, anoxic, and frozen conditions (Schädel et al., 2014, 2016; Schaefer et al., 2016); and appropriate benchmarks for testing the physical dynamics of the coupled atmosphere-snow-soil system (Slater et al., submitted). DOE's NGEE Arctic project has focused on understanding the heterogeneity of polygonal tundra ecosystems, developing approaches to represent that heterogeneity in ESMs, and creating benchmarks for testing land models to reduce uncertainties of permafrost-affected ecosystems under a changing climate.

C.4.2 New Metrics and Benchmarking Approaches

In addition to assembling key datasets to benchmark physical, vegetation, and biogeochemical predictions of land models, it is crucial to identify the relationships between these variables in order to test whether model predictions of these relationships are accurate. While this is true everywhere, it is particularly the case at high latitudes because the climate gradients are especially steep and the heterogeneity of model-generated and reanalysis climates in the region is very high. For example, active layer thickness is a highly emergent quantity that results from the complex interplay between soil properties, snow dynamics, local climate, and fine-scale hydrologic variation; what is needed to benchmark models is not just observations of active layer thickness, but how measured active layers vary across gradients of these underlying driving variables, in order to diagnose specific model processes that are contributing to biases relative to observations. Thus, where possible, data from observational networks should be combined, and the types of observations made at existing networks should be expanded to best utilize observations focused on different aspects of the terrestrial system.

C.4.3 Observational Data Needs

One can break down the key observational needs into three main groups: vegetation, soil biogeochemistry, and the physical system. Each of these requires a much more detailed treatment and testing than was possible with the CMIP5 generation of models. For many of these, data exists and needs to be synthesized and developed into metrics, whereas for others the data must be collected.

Table C.4.1. Observational requirements for benchmarking of high-latitude processes.

Domain	Status	Variables
Vegetation	Data exists and is being used for benchmarking	LAI, Baseline PFT maps, Productivity
	Data exists but must be synthesized and/or used for benchmarks	Biomass, non-vascular plant dynamics, fire disturbance frequency
	Data does not exist	Large-scale changes to vegetation distributions
Soil Biogeochemistry	Data exists and is being used for benchmarking	Soil carbon distributions; ecosystem responses to nutrient fertilization; Site-level CH_4 fluxes
	Data exists but must be synthesized and/or used for benchmarks	Oxic, anoxic, and frozen soil respiration rates, ecosystem warming experiments; extreme scarcity of synthesized soil carbon observations from Siberia
	Data does not exist	Pan-arctic organic layer thickness maps
Physical Snow–soil–hydrologic system	Data exists and is being used for benchmarking	Snow cover extent, site-level soil temperatures, site-level hydrology, basin-scale streamflow, gravity-based mass changes, site-scale ALT
	Data exists but must be synthesized and/or used for benchmarks	Large-scale soil moisture, Large-scale snow thickness, SWE
	Data does not exist	Large-scale maps of ALT, Changes to permafrost extent

C.4.4 Model Development and Output Requirements

The ESM community has made substantial progress since CMIP5 in representing ESM structures of key systems that govern climate feedbacks from high latitude ecosystems. These include: permafrost physical state, exchange of energy and mass between the land and atmosphere in high latitudes, permafrost biogeochemical dynamics, dynamic organic soil layers, and vegetation dynamics across the tundra–boreal forest ecotone. However, these have been done primarily one at a time in different models with no coupled ESMs that include a high level of sophistication in all aspects of the high latitude system. Furthermore, some aspects of the high latitude system remain poorly resolved in models, in particular the complex hydrology and associated fine-scale heterogeneity that exists at high latitudes.

Approaches to better sample models to enable benchmarking are also critically required. CMIP5 protocols were able to benchmark soil thermal dynamics, but only poorly represented soil hydrological dynamics, for example, in predictions of unfrozen moisture content or detailed snowpack information, and had very little information on soil biogeochemical dynamics. CMIP6 protocols request more detailed output variables across these domains, including vertically-resolved carbon stocks, nutrient dynamics, and more finely-resolved thermal and hydrological variables (Jones et al., 2016; van den Hurk et al., 2016), allowing a more effective and systematic benchmarking capability for ESMs.

C.5 Tropical Processes

Nathan G. McDowell, Paul Moorcroft, and Charles D. Koven

C.5.1 Scientific Challenges and Opportunities for Model Evaluation

Tropical ecosystems present many processes that overlap with those in other biomes, but also have additional complexity that makes modeling and benchmarking a distinct challenge from that experienced in other regions. These include challenges related to biodiversity and how to represent it in simulations, and understanding the role biodiversity plays in buffering ecosystem responses to perturbations. Much advanced modeling has been done in tropical forests and through these efforts we have unveiled many challenges, including the difficulty in representing the diverse variety of above and belowground traits as they relate to water acquisition and use, and carbon metabolism. Benchmarking has revealed these challenges through comparison to drought-experiments and atmospheric constraints, with previous and current MIP's providing great insight into the advantages and disadvantages of various numerical representations. While advances have been made, most work has pointed to the critical need for more extensive benchmarking of a range of processes at a range of scales, along with associated UQ and model development.

Representing these processes is particularly crucial as tropical forests are predicted by the CMIP5 generation of ESMs to be particularly important for both the carbon–climate and carbon–concentration feedbacks. This importance led to the focus of the NGEE Tropics project to develop and synthesize key datasets required to test the representations of tropical forest dynamics in ESMs, as well as to develop and integrate into ESMs novel modeling approaches for representing these processes. The activities described below, including synthesizing forest inventory data for benchmarking demographic models, collecting more highly process-resolved observations on plant hydraulics and nutrients, and introspecting models to allow for benchmarking, are core activities of the project, which will help with the goal of reducing model uncertainties in tropical forest dynamics as an Earth system feedback.

C.5.2 New Metrics and Benchmarking Approaches

New and novel datasets, including spatially distributed inventories of survival and mortality (e.g., RAINFOR and Forest-GEO) and ecosystem processes (e.g., FluxNET, GEM), are providing insight into how to improve model realism, but these have not been capitalized on for benchmarking. Such regionally and pan-tropically distributed datasets can enable advances in model benchmarking, which thus far has been primarily sub-regional in scale.

C.5.3 Observational Data Needs

Data availability is improving for species level traits of value for model parameterization, but evaluation datasets against manipulations (drought, CO_2, temperature) are extremely limited, and while inventory datasets are available, benchmarking against them has yet to be attempted. Remote sensing is promising using a variety of platforms that can provide ecosystem level benchmarking, but cannot yet provide species or individual resolution information. FLUXNET sites exist, but again are few and far between. Understanding physiological processes is one of the largest uncertainties in the tropics, again due to the diverse nature of forest composition and climate drivers both within and across sites. Thus a combination of data types, from inventory to process measurements to fluxes and remote sensing, provides the best possible suite of benchmarking in the tropics. This is true for all systems, but in the particularly complex tropics this is especially true. Large gaps exist in spatial coverage of critical regions, particularly in the perhumid western Amazon, tropical Africa, and in the island regions surrounding southeast Asia.

Key parameters that require investment for data collection include turnover, C allocation, whole tree hydraulics, phenology, LAI, reproduction, dispersal, and all of their controls. Belowground processes, including soil depth, soil moisture availability, and soil water acquisition for transpiration, are recognized as important. Multiple processes were identified as poorly understood, such as how mechanisms of seasonal drought tolerance transcend to anomalous drought survival, and interactions with mean annual precipitation, vapor pressure deficit, and fertility. The community agreed that looking at response surfaces for benchmarking from both observational and manipulative studies was extremely valuable.

Table C.5.1. Observational requirements for benchmarking of tropical processes.

Domain	Status	Variables
Vegetation	Data exists and is being used for benchmarking	Greenness indices; upscaled carbon flux data; static remotely-sensed biomass
	Data exists but must be synthesized and/or used for benchmarks	Inventory data: biomass, growth, mortality; plant trait covariation with climate; chlorophyll fluorescence; experimental climate manipulations
	Data does not exist	Large-scale biomass dynamics; tropical CO_2 fertilization experiments; pantropical carbon allocation datasets
Soil Biogeochemistry	Data exists and is being used for benchmarking	Soil carbon distributions, profiles, isotopic data
	Data exists but must be synthesized and/or used for benchmarks	Ecosystem process variation across soil fertility gradients
	Data does not exist	Pan-tropical peatland maps
Physical soil-plant-atmosphere system	Data exists and is being used for benchmarking	Upscaled ET flux data; terrestrial water storage; river runoff
	Data exists but must be synthesized and/or used for benchmarks	Plant stemwood trait variation
	Data does not exist	Vertical root water uptake profiles, sap flow datasets

C.5.4 Model Development and Output Requirements

Model development in water uptake, plant hydraulics, carbon allocation and metabolism, and mortality and survival strategies, all within a framework that accounts for hyper-diversity, has been targeted as urgent steps for next-generation models in the tropics. Nearly all aspects listed above as observational needs also are model development needs. Thus, what is needed for benchmarking purposes is a greater ability to test these novel processes and compare them against observations.

As the community shifts from unstructured to structured vegetation models, model outputs must move beyond gross stocks and fluxes and include information on the heterogeneity and structure of vegetation. This includes size distributions, distributions of plant traits in models that predict these, and more detailed heterogeneity associated

with LULCC. Furthermore, how these axes of heterogeneity covary with each other and with plant function is crucial to understand the role that diversity and heterogeneity play in these ecosystems.

More finely-resolved process models must also include sufficient outputs to benchmark these processes against observations. For example, models that trace hydraulic fluxes from the soil through the canopy must be testable against observations of sap flow, tissue water potential, and overall canopy fluxes, and thus must output this information for purposes of comparison. As nutrient-enabled models include more detailed representation of both nitrogen and phosphorus, key diagnostics at site, regional, and global scales are required to evaluate the representation of the nutrient cycling and the relationships between nutrient and carbon cycling in these models.

C.6 Remote Sensing

Shawn Serbin

C.6.1 Scientific Challenges and Opportunities for Model Evaluation

The large extent and high diversity of vegetation comprising Earth's biomes present a significant challenge for local to global-scale terrestrial ecosystem process modeling efforts, including benchmarking and evaluation of model projections. To provide the knowledge and understanding necessary to improve model parameterizations, representation and evaluation of alternative model structures and observations are needed at the relevant spatial and temporal scales for controlling processes. The general goal of remote sensing from leaf to global scales is to provide critical information on ecosystem dynamics (e.g., seasonality, response to perturbations), and states (e.g., composition, structure, biomass), as well as to scale, map, and monitor important ecosystem properties and processes across space and through time. Compared with other observational and model evaluation datasets (e.g., inventory, eddy covariance, manipulation, and global change experiments), remote sensing data provide the synoptic, continuous, and temporally frequent observations needed for site to global model benchmarking. Moreover, the relative magnitude of remote sensing datasets of various types and temporal extents has helped to usher in the current data-rich era in ecology and global modeling, providing large volumes of information across scales that could be leveraged within data assimilation frameworks for model calibration and development activities.

Remote sensing observations and products useful for model evaluation span a fairly broad range of scales (temporally and spatially) as well as biophysical properties such as leaf area index (LAI) and the fraction of photosynthetically active radiation absorbed by vegetation (e.g., Myneni et al., 2002; Baret et al., 2007), states such as biomass (e.g., Saatchi et al., 2011), soil or canopy moisture (Petropoulos et al. 2015; Schimel et al., 2015), energy balance products such as surface albedo (Shaaf et al., 2002), to process-level observations, including evapotranspiration (Mu et al., 2011), photosynthesis (e.g., Running et al., 2004; Ryu et al., 2011; Guanter et al., 2014; Serbin et al., 2015), and plant functional traits (e.g., Asner et al., 2015; Singh et al., 2015). Calibration of algorithms for the retrieval of measurements using remote sensing observations vary in approach and complexity but generally require some degree of the physical relationship as well as independent information from ground or other observations for evaluation prior to any scientific or modeling use. In addition to other smaller campaigns, past and ongoing global change manipulations (e.g., FACE, DOE SPRUCE), field experiments, and large-scale projects such as the DOE Next Generation Ecosystem Experiments (NGEE) in the Arctic and tropics as well as NASA's Arctic Boreal Vulnerability Experiment (ABoVE) provide ample opportunities to refine remote sensing methods and products for use within ILAMB and elsewhere (Schmid et al., 2015). Leveraging projects such as these enables the development and testing of existing approaches, new techniques, and the development of new observations or data products based on new instrumentation or the "fusion" of observations into new synthetic measurements. Here, we briefly review the use of past, present, and future uses of remote sensing data and new technologies for model evaluation within ILAMB, provide caution for proper use and avoiding pitfalls, and provide some guidance on ways to use observations within model–data integration or data assimilation frameworks.

Within the scope of benchmarking terrestrial ecosystem processes and climate–biosphere feedbacks with remote sensing observations, we explored the following key questions:

- What can be observed with remote sensing now and with additional research or development efforts? What is operational and what is experimental?
- What can be done with existing technologies?

- What new imaging technologies, approaches, or product development efforts are needed?
- How do we better leverage airborne platforms? Can we use sub-orbital data for local-scale benchmarking and couple this with larger scale activities using satellite observations?
- How do we sustain a suite of remote sensing observations for current and future MIPs and benchmarking activities given the typically ephemeral lifespan of most airborne and spaceborne platforms?
- How do we better incorporate uncertainties in remote sensing observational data products with benchmarking activities?

In addition, it is important to understand what processes and scales remote sensing data can contribute for model evaluation and development. Many approaches exist for developing observations and data products for efforts like ILAMB from the leaf to global scales. Critical for these activities are a careful consideration of the methods used for scaling observations, including algorithms and uncertainties, as well as the methods for evaluation such as point-to-point versus average response. These topics and others were discussed to develop a roadmap for data product development, evaluation, uncertainties and appropriate uses, and sustainability and evolution within ILAMB. It was agreed that this discussion was a critical activity for developing long-term products that can help constrain new process representations and model structures. Importantly, these new data products can be developed with the same iterative uncertainty quantification frameworks used for model–data experimentation (ModEx). This is critical for developing standard approaches for model evaluation and calibration through data assimilation, which is currently limited by availability of products and a dearth of information on product uncertainties.

C.6.2 New Metrics and Benchmarking Approaches

Remote sensing observations and derived data products fill a critical role in the evaluation of process models from the site to global scales (Schmid et al., 2015; Schimel et al., 2015). One of the key capabilities of remote sensing observations for model evaluation and benchmarking is the ability to capture the broad, synoptic context as well as relevant timescales (annual mean, seasonal cycle, interannual variability, trend) for comparisons with a wide array of model states. However, important considerations include the necessary spatial extent, spatial and spectral resolution, and whether individual tree/plant scale or larger watershed scales are relevant. Moreover, these considerations also depend on the model process under evaluation. Finally, remote sensing metrics and benchmarks can evolve with new instrumentation and/or help to guide further investments in observing platforms to improve benchmarking activities. Table C.6.1 highlights some of the new remote sensing benchmarks that could be leveraged or expanded within ILAMB.

Table C.6.1. New Metrics/Model Diagnostics/Benchmarks.

Topic	Proposed Approach	Details & Rationale	Spatial & Temporal Scales	Benchmarks
Ecosystem state	Active and time-series optical remote sensing, sensor fusion	Benchmark model output states, such as biomass, canopy structure or soil moisture. Do models capture the evolution and spatial patterns	1 m – 10 km, annual	RMSE, spatial patterns, vertical distribution
Vegetation/ soil properties, parameters, and functional diversity	Imaging spectroscopy, microwave, thermal, gravity	Evaluate model parameterization and emergent properties. Do models adequately capture patterns in plant functional properties/traits and soil moisture through time, resulting in accurate states for the right reasons?	1 m – 30 km, monthly to annual	RMSE, spatial and temporal patterns, Vertical distribution, functional relationships

Topic	Proposed Approach	Details & Rationale	Spatial & Temporal Scales	Benchmarks
Vegetation dynamics	Time-series active/ optical remote sensing, sensor fusion	Do models accurately represent plant demography and succession, growth/ mortality	1 m – 10 km, daily to monthly	Functional relationships, phase, RMSE
Vegetation seasonality & functional phenology	In situ, airborne, satellite time series, synthetic time series from multiple platforms, thermal, SIF	Evaluate model capacity to represent phenology from arctic to tropics, capture seasonality of C, water, EB cycling	1 m – 10 km, daily to weekly	Phase, temporal pattern, interannual variability, functional relationships
Canopy optical properties and energy balance	Canopy simulator: Simulate the spectral signature (SWIR, thermal, microwave) of various remote sensing instruments given a particular model state. Enable direct connection between RS data and model structure	Modify model canopy radiative transfer code to provide directly comparable outputs (e.g., surface reflectance, LiDAR waveform, thermal brightness radiance). Evaluate model structure and dynamics, facilitate direct data assimilation	1 m – 1 km, weekly to monthly	RMSE, seasonal cycle, evolution of canopy optical properties, functional relationships between optical properties and model processes (e.g., GPP)
Perturbations	Time series, active microwave and lidar, sensor fusion, thermal	Test ability for models to capture and correctly respond to various disturbance or change events	10 – 100 km, days to annual	RMSE, spatial patterns, temporal trajectory, phase

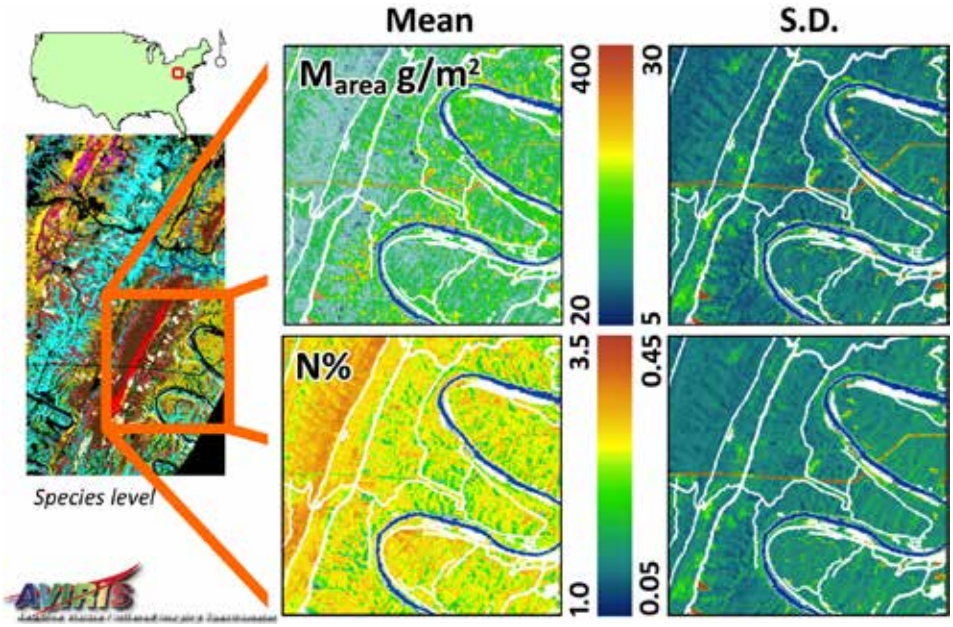


Figure C.6.1. Example maps of foliar morphology (leaf mass area, Marea) and nitrogen concentration (N%) derived with NASA AVIRIS imagery. Trait maps such as these can be used to benchmark prognostic model predictions of properties such as canopy / leaf nitrogen over space and time. However, the utility of these maps is dependent on providing appropriate uncertainty estimates to evaluate model spread versus data uncertainty. Adapted from Singh et al. (2015).

In general, ILAMB and model benchmarking could leverage new remote sensing techniques, technologies, and platforms, including airborne platforms, to expand the diversity and extent of observations for evaluating models (Schimel et al., 2015; Shugart et al., 2015; Schmid et al., 2015). For example, imaging spectroscopy (IS) enables the retrieval of canopy and soil functional traits at a range of scales (e.g., Ollinger et al., 2002; Ustin et al., 2004; Singh et al., 2015; Serbin et al., 2015; Figure C.6.1), which can be used to test model carbon and nutrient cycling. IS data can also quantify plant composition and functional diversity across landscapes, allowing for the characterization of patterns across climatic and topographic gradients, enabling the parameterization or validation of model response surfaces (Fisher et al., 2015). Active systems such as LiDAR (from airborne or spaceborne platforms) could be used to evaluate modeled changes in canopy structure through time or in response to disturbance, or to test model predictions of carbon storage, succession, or demography, together with spectroscopic information e.g., Kumar et al., 2015; Figure C.6.2). Thermal infrared (TIR) observations are particularly useful for evaluating model predictions of the surface energy balance, seasonality, and water cycle and can be coupled with measurements of soil moisture or storage. Together, observations from imaging spectroscopy, LiDAR, and TIR can be used to benchmark the representations of surface energy balance, albedo and canopy radiative transfer (Figure C.6.3). Finally, there are an increasing number of leaf to near-surface remote sensing datasets (e.g., leaf optical properties, phenology cameras, tower-mounted spectrometers) that could be used to benchmark and evaluate model leaf to canopy parameterization, predictions of seasonality, or scaling approaches.

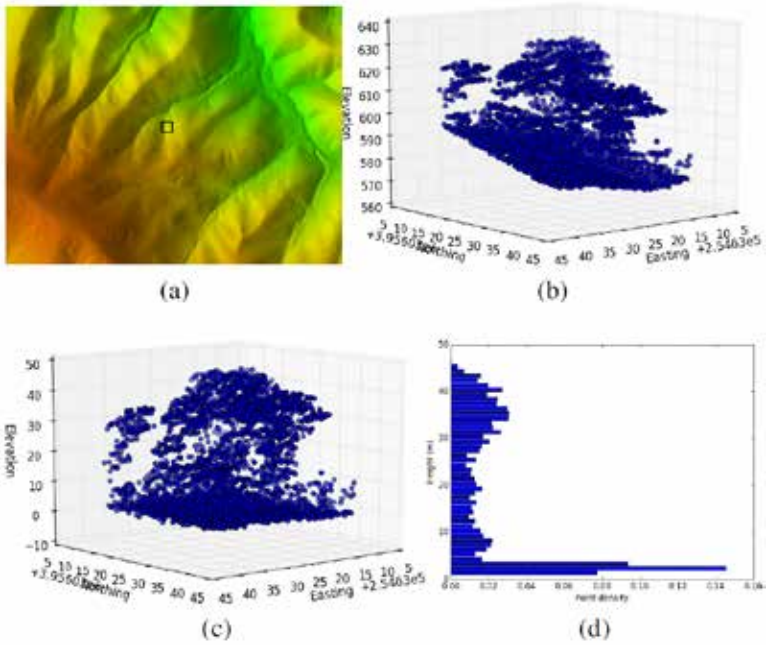


Figure C.6.2. (a) 3-D LiDAR point cloud at 30 m × 30 m region (black square) in a typical cove forest of the Great Smoky Mountains National Park. (b) The raw LiDAR point cloud (3,985 points), showing the imprints of the underlying cove topography. (c) LiDAR point cloud after topographic detrending and filtering (3,936 points) that converted the elevations to above ground level elevation. (d) Distribution of LiDAR point density along the vertical profiles in a cove forest dominated by tall trees and a dense understory. Adapted from Kumar et al. (2015).

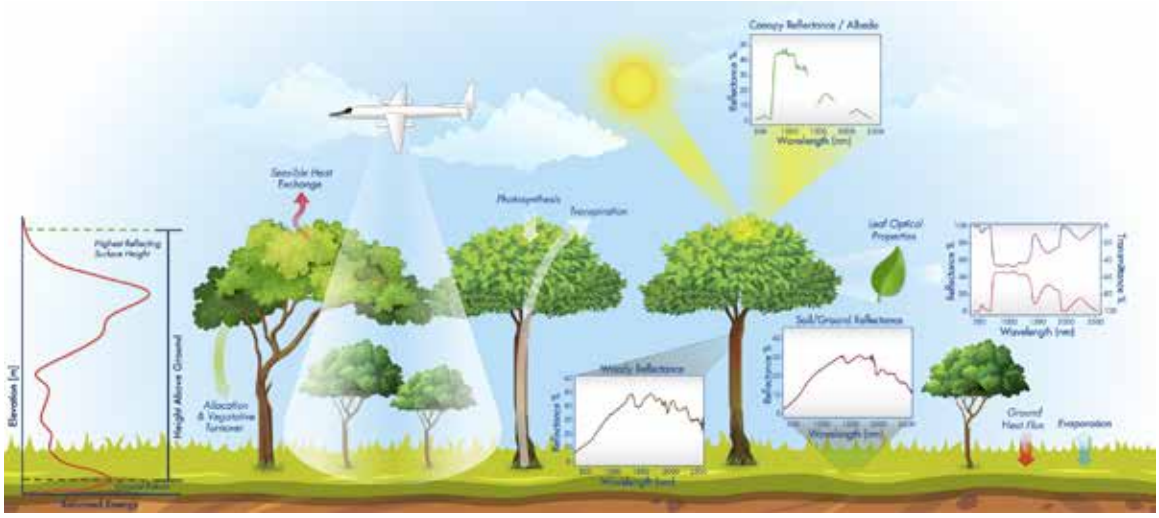


Figure C.6.3. The relationship between radiation, canopy structure, optical properties and key processes including metabolism, water and energy cycling, as well as C allocation and turnover. Optical and thermal data can inform model representation energy, C, and water fluxes while LiDAR remote sensing can provide critical information on canopy structure, turnover and disturbance. (Adapted from Serbin et al., in prep)

In addition to direct or seasonal comparisons, remote sensing data within ILAMB should be used as a metric of functional responses. For example, models often fail to adequately capture short-term perturbations, such as acute drought; however, remote sensing observations can often more completely characterize the ecosystem response and short- to long-term recovery (AghaKouchak et al., 2015). By comparing the observed functional responses through a suite of remote sensing measurements (e.g., Table C.6.2) we can test the model response in magnitude, timing, and extent. Moreover, we can mine remote sensing archives to find similar perturbations through the record of data to identify the typical response to a change event and test the model functional response. This serves as a means to both assess and provide functional constraints for models.

C.6.3 Observational Data Needs

The advantage as well as the disadvantage of remote sensing observations for model benchmarking is the diversity in scale, platforms, sensors, and approaches for collecting, scaling and providing data products for key terrestrial biophysical and functional properties. As such, a challenge for benchmarking with remote sensing is reconciling the typically ephemeral nature of many satellite or aircraft missions which make it challenging to provide consistent or wall-to-wall data products over long periods, scale mismatch, and embedded assumptions in data products. However, the diversity of past, present, and future missions also lends itself to the development of new observations and products for model evaluation such as comparison of model states (e.g., LiDAR canopy structure, imaging spectroscopy derived forest composition, and functional diversity) and process (e.g., SiF vs. model GPP). An additional challenge lies in the tendency to develop remote sensing data products which are themselves based on a model (e.g., global MODIS NPP, LAI, ET) and the need to reconcile the potential differences in the ways in which these observations/model outputs are defined. This often results in products that should not be used in benchmarking (see Section C.6.4 below). Below are the beginnings of a list of considerations for filling data needs:

Table C.6.2. Measurement Needs.

Topic	Measurement Approaches	Temporal Scale	Spatial Scale	Considerations
Aboveground structure & biomass	LiDAR, radar, repeat high-resolution optical imagery, sensor fusion	Annual	1 m – 10 km	LiDAR coverage is still limited and spatial coverage is typically small. Data availability varies. Access to high-resolution optical imagery to create canopy height maps is still limited. Microwave and interferometric SAR coverage is limited or pixel size is typically too large for detailed site scale assessment. Uncertainties with allometry and scaling approaches
Plant demography	LiDAR, optical time series, imaging spectroscopy	Monthly to annual	1 m – 10 km	LiDAR similar to above, limited access to IS data. Need to integrate remote sensing with ground observations
Detailed plant composition, land-cover change	LiDAR, optical time series, imaging spectroscopy, sensor fusion	Annual	1 m – 100 km	Beyond basic PFTs. Spatial scale, temporal resolution, phenological timing

Topic	Measurement Approaches	Temporal Scale	Spatial Scale	Considerations
Succession and mortality	LiDAR, optical time series, microwave, sensor fusion	Monthly to annual	1 m – 10 km	Attribution, timing of imagery
Carbon flux, photosynthesis, photosynthetic capacity	Optical time series, imaging spectroscopy, vertical column CO_2, SIF	Daily to monthly	10 m – 100 km	Measurements of C flux parameters/ photosynthetic capacity (e.g., Vcmax) are preferred over correlation with GPP. Leverage geostationary satellites, space station instrumentation (OCO-3). SIF still needs development to identify links to GPP at remote sensing scales
Water flux/ET, canopy moisture, balance, wetlands	Optical, thermal, microwave, gravity	Daily to annual	10 m – 100 km	Matching flux with storage, delineating seasonal and permanent wetlands
Surface energy balance	Thermal, imaging spectroscopy	Daily to monthly	10 m – 10 km	Higher temporal frequency TIR data at spatial scales of 30 – 100 meters is needed. Spaceborne IS is needed to get high-resolution surface albedo data globally
Vegetation seasonality, LAI, and functional phenology	Optical time series, imaging spectroscopy, thermal, SIF	Daily to monthly	1 m – 100 km	SIF retrieval of C flux still needs development
Vegetation functional traits, biochemistry	Imaging spectroscopy	Monthly to annual	1 m – 10 km	In situ datasets of key plant traits in critical biomes (e.g., Arctic, tropics) are needed to calibrate empirical scaling approaches. RTM approaches need additional development to incorporate a wider range of plant traits. Spaceborne IS is needed to gather global plant trait datasets
Canopy optical and thermal properties, architecture	LiDAR, optical imagery, imaging spectroscopy, thermal	Daily to monthly	1 m – 10 km	Spaceborne IS is needed. Higher temporal frequency TIR data at spatial scales of 30 – 100 meters is needed
Vegetation optical depth	Active/passive microwave	Weekly to monthly	1 km – 30 km	Data availability, spatial and temporal resolutions

In addition to identifying the broad sensor types listed in Table C.6.2 (e.g., spectroscopy, LiDAR, thermal), consideration of the methods for developing data products to meet observations needs together with detailed uncertainty assessment is required for any remote sensing data in ILAMB. Many of the data products can be generated with a range of approaches from empirical to modeled (i.e., variable driven or radiometric data-driven). In some cases empirical approaches are currently preferred (e.g., canopy traits) while others such as microwave vegetation optical depth can be retrieved effectively through radiative transfer modeling. The approach taken should minimize the number of assumptions as well as avoid the use of a model that differs significantly from that underlying the terrestrial biosphere model. In other words, minimize the use of a remote sensing observation that uses a different approach and assumptions or treat as a comparison rather than as a direct observation. In addition, the approach should provide the best estimate of data product uncertainties possible, preferably at the pixel scale. To the extent possible, the development of products and approaches should be shared across groups to standardize. Airborne data should be leveraged where possible to provide novel local to regional scale benchmark data products or target "hot spots" to challenge the models. To achieve these goals, ILAMB should expand its "cyber infrastructure" to enable on-the-fly remote sensing data retrieval (including the mining of airborne data), generation of benchmarks, and model evaluation. This may require some development of standard data pre-processing and preparation, but this should be based on the state-of-the-art in the literature or through discussions with experts in the field.

Finally, capturing the seasonal "functional" phenology instead of only observing the changes in leaf quantity (e.g., LAI), for example, should be explored for use as an ILAMB benchmark. Models may capture the broad leaf emergence/senescence patterns but often fail to capture the true seasonality of C, water, and energy fluxes because they rarely account for changes in canopy physiology through the growing season. Imaging spectroscopy, thermal IR, and SIF are all ways to explore patterns in vegetation functional diversity and seasonality (e.g., Serbin et al., 2015; Guanter et al., 2014). However, we must caution the use of SIF as a direct model constraint given there is still significant uncertainty in the signal observed by *in situ* and satellite based observations. Additional exploration of the SIF signal over time, in response to perturbations, and across sites is needed to better understand the connection between SIF and C flux.

C.6.4 Potential Pitfalls and Misuse of Remote Sensing in Model Benchmarking

A number of potential misuses and pitfalls exist when leveraging remote sensing observations as model benchmarks. As already mentioned, remote sensing data products that are derived from models should be treated as a comparison benchmark and not a direct observation. However, treating model benchmarks as an actual observation and tuning the process model to match the remote sensing benchmark could lead to inappropriate parameterization or unstable model outputs under new environmental conditions. Moreover, remote sensing driven light-use and water-use efficiency approaches should be used cautiously. Instead of direct benchmarks, these products should be used to assess the capacity of models to capture seasonal or inter-annual cycles given environmental conditions and as a comparison of spatial patterns, but the significant differences in model complexity make any direct evaluation challenging and inappropriate. In addition, remote sensing LUE/WUE approaches typically incorporate environmental downscaling on the efficiency term so that it is inappropriate to then evaluate model response to climate to test process model functional responses. However, remote sensing approaches that leverage the same fundamental photosynthetic schemes as the full process model (e.g., Ryu et al., 2011) could be used as an alternative. Whenever possible the allometric relationships used to derive remote sensing estimates of biomass, carbon, or canopy structure should match those used for the same PFTs within the model. Difference in assumptions, uncertainties, and detail of allometric relationships can lead to significant model–data mismatches. At a minimum these uncertainties need to be included in model benchmarking. We also suggest caution using SIF as a benchmark for model carbon flux and GPP. There is still significant uncertainty in the signal observed by *in situ* and satellite-based observations and the current regressions with tower-based GPP are insufficient as direct benchmarks for GPP, although they can be used for comparative purposes. In addition to methodological considerations, spatial and temporal scale of remote sensing data should be considered when developing model benchmarks to minimize having to aggregate/disaggregate data products to match model outputs.

Finally, the ILAMB framework needs a direct way to integrate uncertainties in model outputs and remote sensing benchmarks. Accounting for uncertainty will provide more accurate assessments of model predictions and error, as well as facilitate data assimilation to improve model calibration.

C.6.5 Model Development and Output Requirements

Remote sensing can not only help to evaluate models and submodels to guide new developments at various spatial and temporal scales but also could guide the development of new model outputs to further facilitate direct model–data comparisons (or assimilation) specifically focused on the use of remote sensing observations (Figure 6.3). Requirements for this are consistent spatial and temporal scales, similar variable definitions and units (less important), and explicit development and handling of remote sensing product uncertainties. An important model development recommendation and model output requirement identified within the group was a remote sensing "simulator" (e.g. Figure C.6.4; Viskari et al., in prep) to provide outputs that are more directly comparable with basic remote sensing measurements (e.g., spectral reflectance, thermal radiance, LiDAR waveform) rather than derived products (e.g., biomass). To facilitate this, models would need to update their canopy radiative transfer code base to provide a full canopy spectral response, based on the leaf optical properties and internal model structure, instead of the typical surface albedo in a few spectral regions (e.g., visible, near-infrared). Importantly, it would be beneficial for models to develop this simulator as code that can execute outside of the full model framework, but still compile against the full model code based and functions/libraries, to facilitate rapid execution by running only the functions needed to simulate canopy radiation transfer. This is dependant on the remote sensing instrument's ability to be simulated (e.g., Shiklomanov et al., 2016) to directly compare model predicted (dependent on modeled state) versus observed remote sensing patterns. This approach would also facilitate direct assimilation of remote sensing data to inform model parameterization and test alternative model structures.

In addition to the remote sensing simulator, model code should be adapted to leverage the latest high-performance computing environments to facilitate comparison of models against large remote sensing datasets across large spatial and temporal scales. In addition, on-the-fly retrieval and processing of airborne data will require distributed computing.

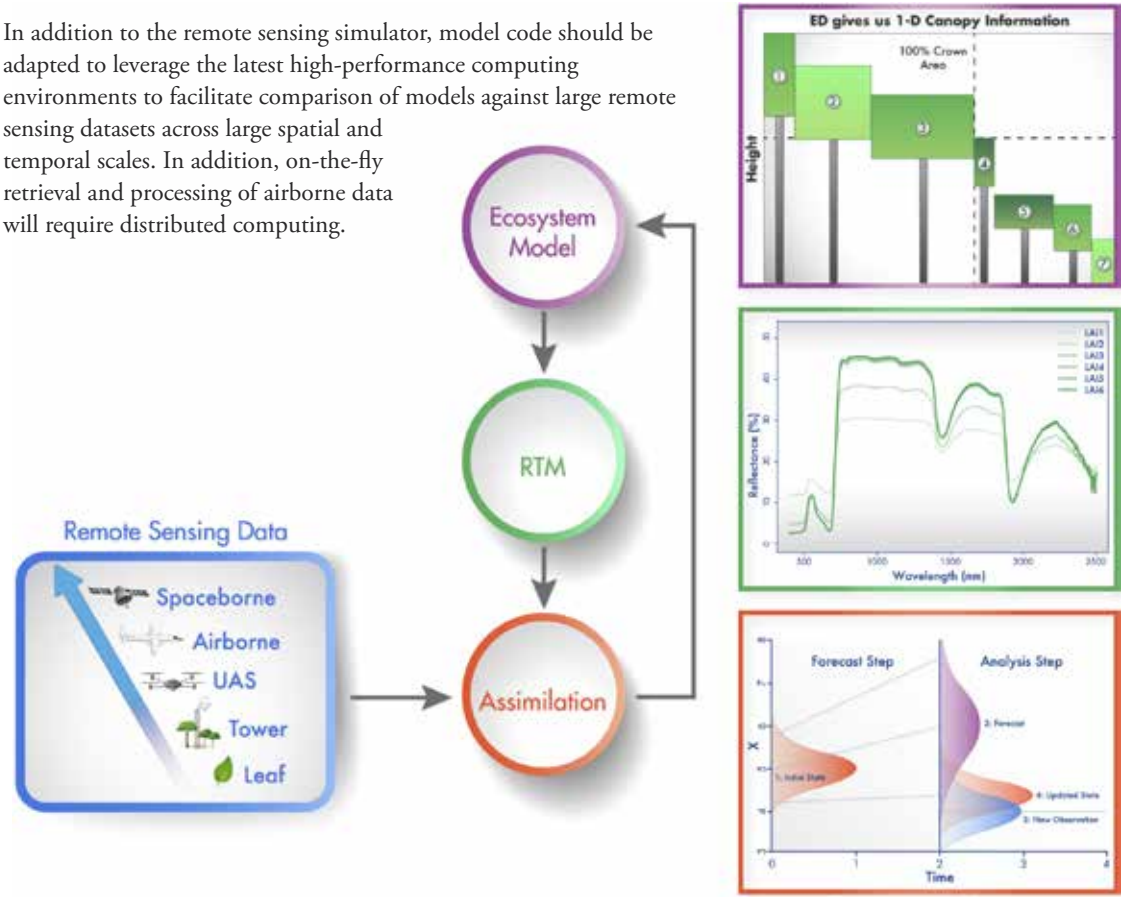


Figure C.6.4. Example of the use of an "sensor simulator" within a terrestrial biosphere model (TBM; in this case ED2) to facilitate direct assimilation of and/or benchmarking against remote sensing observations within the PEcAn framework (Shiklomanov et al., 2016; Viskari et al., in prep). In this approach the output TBM spectral signature is based on the internal model structure (i.e. canopy biomass, height, RT properties) and compared with comparable remote sensing observations (i.e., surface reflectance, albedo). This allows for direct comparison and evaluation of associated processes such as photosynthesis, energy balance, surface temperature and evapotranspiration as well as identify uncertainties and areas to target for model improvement.

C.6.6 Computational Needs and Requirements

Depending on the type of remote sensing observation, scale, algorithmic approach, and resultant data product, the computational needs will vary considerably. For example, an empirical model applied to a series of Landsat images (i.e. image stack) will be relatively quick to generate a new product; however, the use of a radiative transfer model (RTM) together with a highly dimensional dataset to develop a complex data product could take several hours to months to produce on a high-performance computing (HPC) environment. These considerations therefore determine the degree of availability of data products as well as the capacity to provide near real-time information for benchmarking models during short-term perturbations. There are means to reduce computational costs or improve the speed of product development such as look-up tables and learning algorithms (e.g., ANN). ILAMB should invest in a cyberinfrastructure to facilitate the use of airborne remote sensing data such as imaging spectroscopy and LiDAR, including data storage, handling, and preprocessing, as well as distributed or HPC to create or refine benchmark products. In addition, ILAMB should invest in efforts to assimilate remote sensing datasets to test, calibrate, and update model structures. Finally, all of these efforts need to consider observation uncertainties that may require computational approaches (e.g., resampling, ensembles) to estimate algorithm and pixel-scale error assessments.

C.7 Roles for Flux Networks

Dennis Baldocchi

C.7.1 FLUXNET: A Network of Eddy Covariance Flux Measurement Networks

Regional and global networks of eddy covariance flux towers, measuring fluxes of carbon, water and energy between terrestrial ecosystems and the atmosphere, are providing crucial data to the global carbon cycle science community (Baldocchi et al., 2001; Baldocchi et al., 2012; Reichstein et al., 2014).

Individual eddy covariance flux towers are capable of measuring mass and energy fluxes directly and quasi-continuously on time scales of hours, days, years, and now decades. And, by assembling networks of flux towers, one is able to deduce how carbon, water and energy fluxes vary spatially, across many of the Earth's climate and ecological spaces and disturbance/management regimes. Together, these flux data are being used to: 1) produce annual carbon, water and energy budgets (Baldocchi, 2008); 2) provide process level information about how ecosystem metabolism responds to biophysical perturbations (Biederman et al., 2016; Reichstein et al., 2014); 3) examine the role of trends in carbon fluxes to rising CO_2 (Keenan et al., 2013); 4) quantify the variability of carbon fluxes to extreme events in climate forcings (Frank et al., 2015; Reichstein et al., 2013); 5) validate and parameterize a spectrum of machine learning, process and remote sensing models that are predicting, interpolating and extrapolating carbon flux information in time and space (Beer et al., 2010; Xiao et al., 2014); and 6) provide priors for Bayesian data assimilation models (Bloom et al., 2016; Williams et al., 2009).

Between 1997 and 2012 the global FLUXNET project was funded in an *ad hoc* manner with successive grants from NASA and the National Science Foundation as well as from Microsoft Corporation. An effort to modernize and update the FLUXNET data system and expand the database is currently being supported by DOE in the 2014 to 2017 time-frame via collaboration with computer scientists from Lawrence Berkeley National Lab and University of Virginia along with our international partners.

The FLUXNET project and database are ready and ripe for use to advance carbon cycle synthesis by process-based, data assimilation and machine learning models and to address the next generation set of problems and questions; what is causing interannual variability in net and gross carbon fluxes?; are trends in carbon fluxes being induced by global change, and are these changes detectable?; how do fluxes respond to disturbance and management?; is ecosystem photosynthesis and water use efficiency responding to elevated CO_2?; how well do models perform with soil water deficits and over open and complex canopies?; how accurate are the global upscaling estimates of gross and net carbon fluxes?

The production and distribution of flux data, and its accompanying metadata, to the global carbon cycle community requires human effort to recruit data from different countries and cultures, to build a harmonized dataset that has been subjected to quality control and assurance and to have the software and staff to update the database as new

data are submitted. Distribution of these data to the users and production of value added products that are of use to the modeling community for benchmarking model simulations requires a coordinated and sustained effort. Today, the FLUXNET database has submissions from 450 sites, representing 2700+ site-years of data, and 200 variables on meteorological condition, water, carbon and energy fluxes. These data are distributed through **http://fluxnet.fluxdata.org/**, and the dataset continues to grow and expand. In addition, there are 77 sites with over a decade of data, giving the scientific community a new opportunity to study and model interannual variability, trends in fluxes and the effects of climate extremes on carbon, water and energy fluxes. Regional flux networks will continue into the future and new funding is needed to support the global FLUXNET activity to ensure these data are available and are in a useful form for new efforts on data-model inter-comparisons.

C.7.2 Current and Future Roles of FLUXNET for Carbon Cycle Synthesis

The eddy covariance method is currently the standard method used by biometeorologists to measure fluxes of trace gases between ecosystems and atmosphere. Fluxes are measured by computing the covariance between the vertical velocity and target scalar mixing ratios at each individual node (site). Key attributes of the eddy covariance method are its ability to measure fluxes directly, *in situ*, without invasive artifacts, at a spatial scale of hundreds of meters, and on time scales spanning from hours, days, years, and now, decades (Baldocchi, 2014).

Today, eddy covariance measurements of carbon dioxide and water vapor exchange are being made routinely on all continents. The flux measurement sites are linked across a confederation of regional networks in Americas, Europe, Asia, Africa, and Australia, into a global network called FLUXNET. This global network includes more than eight hundred registered and four hundred active measurement sites, dispersed across most of the world's climate space and representative biomes (Figure C.7.1). Within this larger network, smaller meso-networks target specific land use types, such as urban areas, inland water systems, within a region. Many of these locales serve as focal points or anchor sites for sets of ecosystem-scale 'manipulative' studies. Comparative flux measurements are being made at satellite-sites that may differ by plant functional type, biophysical attributes, biodiversity, time since disturbance (e.g., fire, logging, windthrow, flooding, or insect infestation), or management practices (e.g., fertilization, irrigation, or cultivation). Distinct scientific attributes of the flux network include its ability to detect emergent scale properties of ecosystem metabolism at local to regional and global scales and quantify temporal and spatial variability in carbon, water and energy fluxes.

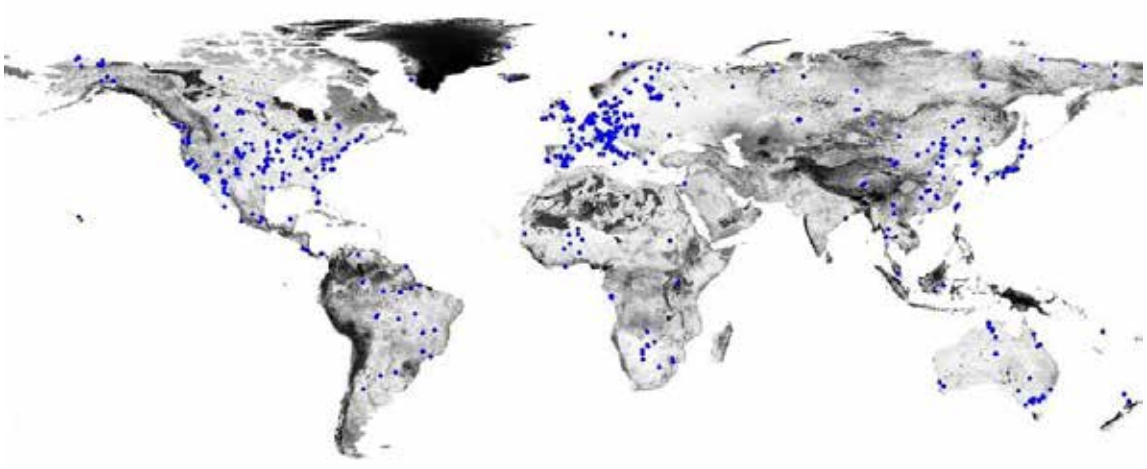


Figure C.7.1. The spatial representativeness of the FLUXNET network (existing towers labeled as blue dots), which is mapped relative to a set of quantitative ecoregions (white-to-black colors). Distance in data space to the closet ecoregion containing a site quantifies how well the FLUXNET network represents each ecoregion in the map. Environments in the darker ecoregions are poorly represented by this network. (Jitendra Kumar, Forrest M. Hoffman, William W. Hargrove, in prep.).

The flux network continues to grow and expand, giving the model community open and fair use access to over 2700 site years of flux data and complimentary meteorological and site information. The size and value of this database is unprecedented in the history of carbon cycle science and offers many unique opportunities for collaboration with model synthesis activities. So continued support for the operation of FLUXNET is a necessary and warranted cost if we are to achieve the scientific goals mandated to the carbon cycle science modeling community.

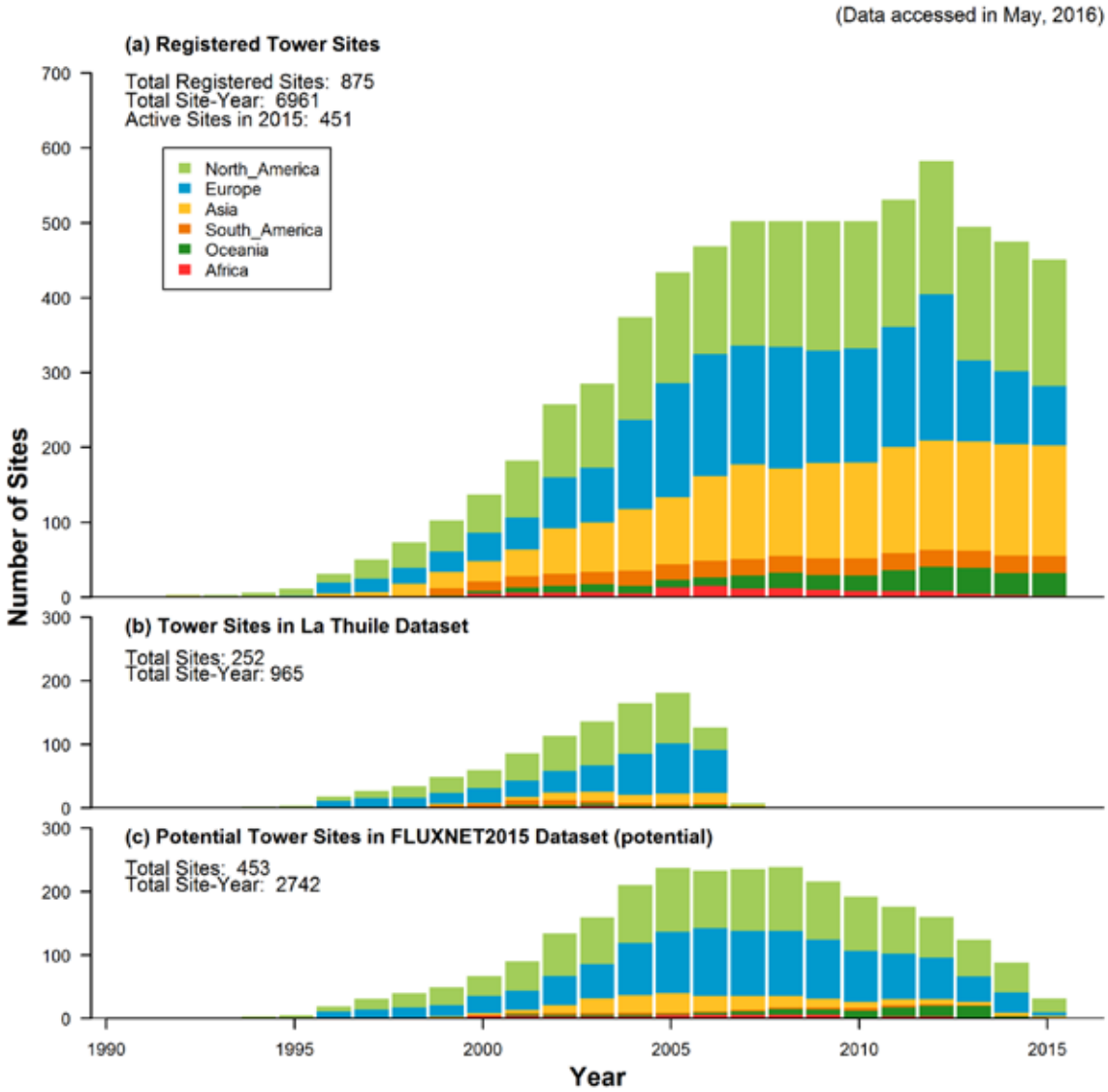


Figure C.7.2. Time series of flux network size by continent. Panels are for potential sites registered in the network, the previous 2007 La Thuile dataset and the potential size of the 2015 FLUXNET dataset, which is being processed, quality assured and corrected.

With regards to modeling work, the flux network is highly representative of most of the world's ecosystems and climate spaces (Figure C.7.3). And statistically, the sparse tower network is representative of much wider regions and landscapes than the individual distinct tower footprints, as shown in Figure C.7.1 (Sundareshwar et al., 2007).

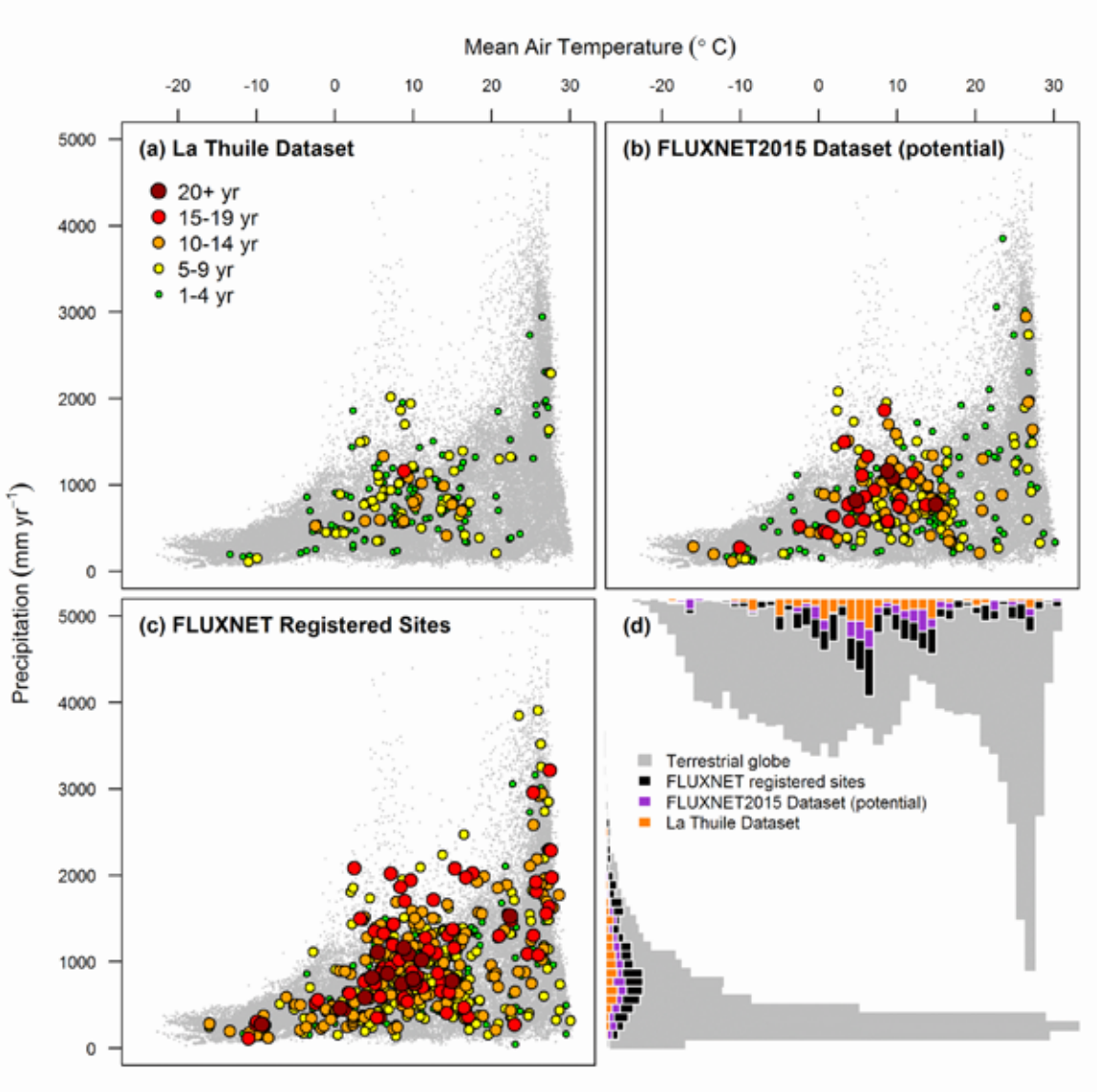


Figure C.7.3. The correspondence between FLUXNET sites and the climate space (precipitation and temperature) of the Earth.

Appendix D. Model Intercomparison Project (MIP) Benchmarking Needs and Evaluation Priorities

D.1 CMIP6 Historical and DECK

David M. Lawrence, Veronika Eyring, and Charles D. Koven

D.1.1 Scientific Challenges and Opportunities for Model Evaluation

The core of the CMIP6 process is a series of experiments, called the Diagnostic, Evaluation, and Characterization of Klima (DECK) (Eyring et al., 2016b). These runs formalize the set of standard climate model configurations that have historically been used both by the modeling centers and by previous CMIP activities, and comprise four experiments: (1) a land–atmosphere only model forced by reconstructed historical sea surface temperatures (i.e., Atmospheric Model Intercomparison Project (AMIP)), (2) a coupled land–atmosphere–ocean preindustrial control, (3) an abrupt quadrupling of CO_2, and (4) an idealized 1% per year CO_2 increase. Because of the idealized nature of these experiments, they are expected to be conducted in all future CMIP activities. In addition to the DECK experiments, all participating CMIP6 models are expected to perform a transient coupled land–atmosphere–ocean historical experiment driven by time-varying greenhouse gas concentrations (Historical) and, for ESMs with a fully prognostic carbon cycle, a second transient coupled land–atmosphere–ocean historical experiment driven by CO_2 emissions rather than concentrations (esmHistorical). The DECK experiments form the hub of all CMIP6 activities (Figure D.1.1), and all other CMIP6 experiments may optionally be performed by modeling centers.

The Historical and esmHistorical experiments have provided the strongest basis for benchmarking of models, because of their correspondence to the period of scientific observation. In the first version of ILAMB (Mu et al., in prep), all benchmark diagnostics for the CMIP5 models were performed on either the Historical or esmHistorical (Hoffman et al., 2014) experiments, both of which are in the CMIP6 DECK experiments. These include a broad suite of remote sensing data, upscaled data such as soil maps, and system-integrative data such as atmospheric CO_2 concentrations.

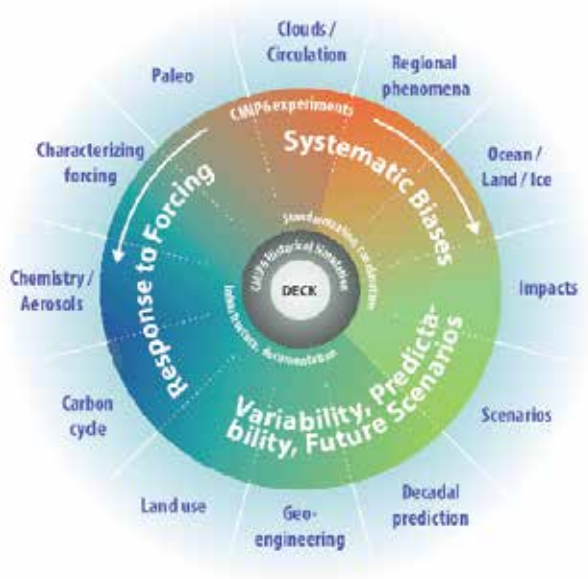


Figure D.1.1. Overview of the CMIP6 structure. All modeling centers will perform the DECK experiments and may optionally perform any other MIPs. Adopted from (Eyring et al., 2016b).

D.1.2 New Metrics and Benchmarking Approaches

As with CMIP5 and the first version of ILAMB, we expect the DECK experiments to form the fundamental basis for model benchmarking approaches. One novel application in applying the ILAMB system to the CMIP6 DECK experiments will be to benchmark the AMIP experiments in addition to the Historical and esmHistorical experiments. This will allow the diagnosis of land model fidelity as a function of ESM complexity, as that complexity changes from the relatively constrained AMIP experiments to the less physically-constrained Historical to the less biogeochemically constrained esmHistorical experiments.

D.2 C^4MIP

Forrest M. Hoffman, Charles D. Koven, and James T. Randerson

D.2.1 Scientific Challenges and Opportunities for Model Evaluation

For the coupled climate–carbon cycle model intercomparison project (C^4MIP; Friedlingstein et al., 2006, 2014a; Jones et al., 2016), several aspects of the experiments create unique opportunities and challenges with respect to benchmarking and model evaluation. A key goal of C^4MIP is to assess model-to-model variations in the strength of carbon–climate and carbon–concentration feedbacks. This is accomplished through a factorial experimental protocol that separates the radiative effects of CO_2 from the biogeochemical effects of CO_2. The use of benchmarks to discriminate among strong and weak feedback parameters such as beta-land (β_L) and gamma-land (γ_L) may contribute to the development of new models that yield more realistic scenarios of carbon dioxide and temperature change during the latter part of the 21st century. These models, in turn, may be able to provide more accurate estimates of allowable emissions necessary to stabilize greenhouse gases at a particular level, thereby achieving a desired maximum temperature change target.

In this context, the development of "emergent constraint" benchmarks is potentially valuable and important. In past work, emergent constraint benchmarks have been developed for gamma-land (γ_L) using interannual variability in atmospheric carbon dioxide and temperature (Cox et al., 2013; Keppel-Aleks et al., 2014) and for the combined magnitude of beta-land (β_L) and beta-ocean (β_O) by assessing model biases relative to the long-term secular trend of atmospheric CO_2 at Mauna Loa (Hoffman et al., 2014; Figure B.3.1). Additional work has suggested that the magnitude of NPP responses to atmospheric CO_2 may be overestimated in the models because they do not properly account for influences on growth from nitrogen and phosphorus limitation (Wieder et al., 2015b). Further quantitative assessment of significance to nutrient limitation for contemporary forest responses to global change is needed, although model overestimates of the strength of leaf area trends in many areas as compared with satellite observations provide evidence for a positive bias in the sensitivity of NPP to CO_2 enrichment (Smith et al., 2016b).

For the C^4MIP simulations planned as a part of CMIP6, new simulations forced with historical and future "business as usual" CO_2 concentrations from 1850 to 2300 will permit exploration of the consequences of contemporary biases in the representation of soil processes for the strength of the permafrost-mediated carbon–climate feedback. In CMIP5, none of the models had made investments in the representation of permafrost carbon stocks, and the idealized 140 year 1% per year CO_2 increase (1pctCO2) simulations were not designed to allow for a quantitative assessment of soil thaw processes that take several centuries to develop.

Modeling centers that will contribute simulations to CMIP6 are expected to use ESMs that have improvements in the representation of several processes, including permafrost (Koven et al., 2011), nitrogen dynamics, fires (Li et al., 2013; Kloster et al., 2010), and hydrological processes (Swenson et al., 2012). Some of the models will have a new representation of dynamic vegetation, and some improvements are expected in the ability of these models to capture observed trends in shrub and tree cover. Furthermore, it is expected that existing aspects of the models will be much more highly constrained by existing observations than in prior versions. For example, observations that were not available at the time of CMIP5, such as globally-upscaled FLUXNET-based GPP (Beer et al., 2010), can allow the models a clearer observationally determined current state of the biosphere to use as their target for development.

D.2.2 New Metrics and Benchmarking Approaches

The crucial requirement for enhancing predictive capability is the ability to tie the transient behavior of the models over the future period to currently-observable quantities. A promising approach here is the identification of possible emergent constraints, as discussed above, for both system-integrative measures such as atmospheric CO_2 concentration or growth rate, and more process-resolved emergent constraints on different aspects of the Earth system. Identifying these and evaluating their domain of applicability is crucial to developing a more predictive capacity for understanding terrestrial carbon cycle feedbacks.

D.2.3 Observational Data Needs

Within the last 5 years, considerable progress has been made in quantifying aboveground live biomass stocks. Estimates by Saatchi et al. (2011), and Baccini et al. (2012) have effectively combined optical, LiDAR, and microwave remote sensing techniques with plot-level field observations to create pan-tropical estimates of aboveground biomass. These estimates point to a considerable reduction in the magnitude of aboveground carbon stocks in intact tropical forests compared with earlier estimates from the International Geosphere-Biosphere Program in the 1970s and other approaches.

In parallel, new estimates of soil carbon have become available in permafrost areas (Hugelius et al., 2014) and globally from analysis of plot-level soil profile observations.

Important gaps that remain include accurate quantification of litter and coarse woody debris pools, wood and litter turnover times, and the representation of organic soil layers. In several biomes, including boreal forests, aboveground and belowground litter is mixed with a living moss layer, live roots, and coarse woody debris in organic soil layers above the mineral surface. Some ambiguity remains with respect to the representation of organic soils and moss layers in existing soil carbon datasets.

Another critical issue is that many of the aboveground live biomass products have been developed for forests. Depending on the methodology, tree and shrub biomass may not be included, making it challenging to compare with grid cell averages from models that reflect contributions from a combination of different plant functional types. Also, this means that aboveground biomass estimates in savannas and shrublands have higher uncertainties. By disaggregating stocks for different plant functional types, C^4MIP may enable more accurate comparisons in the future.

Apart from stocks, important carbon cycle analysis has explored the change in forest inventories to estimate rates of carbon accumulation (Pan et al., 2011). One important next step that could increase the value of the inventory observations is the development of coarsely gridded (~0.5°) carbon change products that do not compromise privacy of landowners, yet enable effective model comparisons and validation using remote sensing imagery. Another important goal is to harmonize the global stock estimates with carbon fluxes derived from national inventories.

Higher quality datasets of land cover change, changing human population density, roads, and other measures of landscape fragmentation are needed to better quantify disturbance dynamics and migration rates within the models.

So far, evaluation of model dynamics at hourly and diurnal time scales has not advanced as rapidly as evaluation using monthly means. This allows model biases that are evident at this timescale (e.g., Ghimire et al., 2016) to persist. This deficiency could be addressed by outputting a set of model fluxes that most highly correspond to measured eddy covariance data (NEP, GPP, Re, LH, SH) at sub-daily frequency over the period of flux tower observations (approximately 1995–present), for direct comparison.

Ultimately, a global carbon stock data assimilation system that integrates inventory and plot-level data to create maps of stocks and accumulation/degradation rates would be extremely valuable to the ESM community. Key requirements for such a system would be the need for wall-to-wall coverage of carbon in all vegetation types and accurate accounting of the continuum of carbon among living vegetation (separate above and belowground components), litter (separate above and belowground components), coarse woody debris, and soil organic and soil mineral pools. Extensive validation would be essential for creating a useful system. Such a system could be forced with "fast" response variables such as assimilated GPP, but a flux-driven system also could be a parallel activity because the timescales and types of observational constraints that are useful are so different.

D.2.4 Model Development and Output Requirements

Currently, the terrestrial components in ESMs have major limitations that may bias carbon cycle feedback projections, and further model development is required to alleviate these shortcomings. A crucial limitation is the current representation of nutrient cycles, which may provide a strong limitation to growth under elevated CO_2, while stimulating growth in response to increased soil decomposition in a warmer climate (McGuire et al., 2001). Whereas some terrestrial components of ESMs have begun including nitrogen and/or phosphorus cycles (Thornton et al., 2007; Wang et al., 2010, Zaehle and Friend, 2010; Yang et al., 2016), uncertainty in these processes is extremely high and requires much more consistent benchmarking with observations. Vegetation models currently in use also primarily represent woody biomass as a uniform pool with a set turnover time, which barely changes under the global change pressures of the 21st century (Koven et al., 2015), whereas in reality wood turnover is a highly emergent property resulting from the recruitment, growth, and mortality of individual tree stems. Models that relax this "big wood" assumption (Wolf et al., 2011) require more detailed output of forest size distributions and output of process variables resolved along an axis of plant size for comparison with observations. Furthermore, vegetation models have typically represented vegetation with fixed PFT traits and either fixed PFT geographic distributions or highly-parameterized DGVM submodels. The changes of plant traits and their geographic distributions in emerging novel climates are highly uncertain and require much more detailed representation of the processes that govern plant functional diversity and biogeography. Other vegetative processes that are poorly represented in current models include water transport from roots to stomates and allocation of plant resources to multiple plant tissues.

In addition to the above weaknesses in the representation of vegetation processes, soil carbon cycling processes are also highly uncertain and likely biased in current models. Current terrestrial models assume linear soil carbon tendencies, a poorly-founded assumption given that decomposition is driven by microbial activity exhibiting highly nonlinear dynamics. Some modeling centers are developing nonlinear soil models (e.g., Sulman, 2014; Wieder et al., 2015a), but the jump in complexity and associated parametric and structural uncertainty of these models (Wang et al., 2014; Wang et al., 2016) must be met with greatly increased benchmarking datasets. Second, the assumption that the near-surface soil environment is a good proxy for whole-soil decomposition is poorly founded, particularly for the high latitude soil carbon pool where steep gradients in the soil environment—driven by transport, cryoturbation, and bioturbation processes—result in enormous stocks of carbon at depth. Resolving these gradients leads to a sign change in the response of the high latitude system with warming (Koven et al, 2011), and it is thus imperative for models to systematically resolve these vertical gradients (He et al., 2016) and output biogeochemical variables along the vertical axis for benchmarking purposes. Lastly, terrestrial models have typically focused on mineral soils, despite the importance of peatlands in both high latitude and tropical ecosystems. Resolving the processes responsible for organic soil dynamics, and benchmarking these models with synthesized datasets, is crucial to remove this bias from model projections.

D.3 LS3MIP

Hyungjun Kim, Jiafu Mao, and Andrew G. Slater

LS3MIP (van den Hurk et al., 2016), another set of optional CMIP6 experiments, contains a series of coupled and off-line land surface experiments designed to illuminate feedback processes as well as provide information about model structure and parameters. It is a coordinated effort among the Global Soil Wetness Project 3 (GSWP3), Global Land–Atmosphere Coupling Experiment (GLACE) and Earth System Model Snow Model Intercomparison Project (ESM-SnowMIP). Each project may have experiments additional to the LS3MIP core. Metrics for LS3MIP models are likely to include standard verification methods, items aimed at assessing feedbacks, and methods for understanding process representation in models. As with all data used for model assessment or data assimilation, understanding the uncertainties (both observational error and representativeness error) is required. LS3MIP is largely concerned with snow and soil processes, their (often seasonal) timescales, and their impact on the greater climate system. To that end, the discussion here revolves around snow and soil.

Verification involves simply comparing model output to observations using standard scores such as bias, root mean squared error, mean absolute difference, etc. These metrics are designed to demonstrate the skill of the model simulation, while not necessarily attributing cause and/or effect. Measures of feedback strength have been proposed for soil moisture (Koster et al., 2004), snow (Xu and Dirmeyer, 2011) and albedo (Qu and Hall, 2006, 2007). The

relationship between large scale snow variables and atmospheric circulation indices such as the Arctic Oscillation have been used (Furtado et al., 2015). Deciphering specific land model weaknesses may best be achieved by making assessment independent of forcing data and/or initial conditions; understanding functional relationships between variables provides a likely path.

A further consideration is that of model output time and spatial scale. ILAMB has primarily used monthly mean data from land models, often interpolated from their native grid to a standard grid, which can lead to limitations. Future output may consider derived diagnostic variables that are integrative or decipher finer timescale processes—for example, daily runoff from river basins would allow for hydrograph recession analysis that gives more insight to surface vs. groundwater runoff processes, or model systems may store the final day-of-year of the seasonal snowpack, which might be defined as a 60-day or more continuous snow cover (Slater et al., 2013) or something similar.

Snow cover extent data include the NOAA Climate Data Record based on the Rutgers historic snow extent (Robinson et al., 1993) dating back to 1967. More recent, higher-resolution records of snow cover are available: NOAA's Interactive Multisensor Snow data (4 km since 1997) and the NASA EOS era of data (1999–present) using MODIS sensors at 500 m resolution (Hall et al., 2006, 2010). At the global level, our knowledge of snow cover is fairly good at least in a relative sense (one year compared to another), though exact timing (to the day) in marginal snow cover and mountainous regions still contain uncertainty. An example of analysis of snow extent using the CMIP5 models was performed by Brutel-Vuilmet et al. (2013), where the emphasis was whether models capture the observed multi-decadal trend of decreasing extent in the spring season. Along with area covered, the date of final melt can be indicative of melt rate relative to available energy. The choice of metric for assessing snow extent can be important; for example, Toure et al. (2016) use the Nash-Sutcliffe Efficiency (NSE) score and correlation coefficient (r) for evaluating the time series of snow cover in CLM4. However, for a time series containing a cyclic component, NSE and r will return high values so long as the seasonal cycle is reproduced, therefore not elucidating model capability.

SWE at the global scale, in the opinion of the author (Slater), remains an unknown quantity for the purposes of rigorously verifying models. Station-based interpolations (Brown and Brasnett, 2010) and products applying remote sensing techniques (e.g., GlobSno [Pulliainen, 2006]) give broad estimates and may give indications of model results that might be largely erroneous but that are not yet of the standard to suggest what it "correct"—this remains a gap in our knowledge. Because of poor SWE information at large scales and in mountain regions, there is a long-term initiative underway among the snow community to improve this situation, including the advance of satellite sensors and coordinated data assimilation systems to NASA's Decadal Survey. Spatial heterogeneity of snow depth and SWE should urge caution when comparing point measurements to gridcell averages; poor comparisons can be made, for example with SNOTEL data (Toure et al., 2016).

Functional relationships have been used to assess model abilities. To separate the influence of surface meteorology forcing from model structural or parameter error, Slater and Lawrence (2013) used a simple empirical model of permafrost driven by surface meteorology of respective CMIP5 models. The relative position and trajectory of permafrost diagnosed directly compared to the empirical model can inform whether land models are too warm or cold. Similarly, the impact of modeled snow insulation was assessed by looking at differences in air and soil temperatures (Koven et al., 2013) and extended to account for snow depths and relative climates (Slater et al., submitted).

The International Soil Moisture Network (**http://ismn.geo.tuwien.ac.at/ismn/**) curates a variety of *in situ* and satellite derived estimates of soil moisture which can be used for assessing modeled water budgeting; e.g., partitioning between runoff, evaporation, and storage. Standard comparisons of moisture levels are useful (Xia et al., 2015b), although innovative methods to understand sources of simulation uncertainty are even more desirable (Nearing et al., 2016). Total water storage from gravity anomalies (GRACE) can provide a broader integrative view of model abilities (Kim et al., 2009; Swenson and Lawrence, 2015).

Shallow soil temperature data (< 5 m, often < 1 m) suffers from heterogeneity issues, and is often sporadic, poorly distributed, and sometimes not reported as a standard variable even when measured. As an example, historic soil data from Russia is often measured at disturbed agricultural plots that are not representative of the local vegetation type. The rate of heat uptake over time within the terrestrial surface and actual temperatures at depths greater than 10 m are available only from a variety of boreholes across many different climate zones, from tropical to polar (Cuesta-Valero et al., 2016).

Albedo retrievals from satellites, such as the NASA-sponsored MOD43 series of products (Schaaf et al., 2002), have proved useful in assessing prognostic albedo in models as well as detecting weaknesses, including poor representation of canopy snow interception (Thackeray et al., 2015).

Also, LS3MIP includes additional experiments using four alternative meteorological forcing data sets: GSWP3 (Kim et al., in preparation), the Princeton forcing (Sheffield et al., 2006), WFD and WFDEI combined (allowing for offsets as needed [Weedon et al., 2014]) and the CRU-NCEP forcing used in TRENDY (Sitch et al., 2015). The model outputs will allow assessment of the sensitivity of land-only simulations to uncertainties in forcing data. Kim (2010) utilized a similarity index (Ω; Koster et al., 2000) to estimate the uncertainty derived from an ensemble of precipitation observation data sets relative to the uncertainty from an ensemble of model simulations for evapotranspiration and runoff. It was found the uncertainty of forcing precipitation propagates in a relatively reduced way to evapotranspiration and an amplified way to runoff (Kim, 2010; Figure D.3.1).

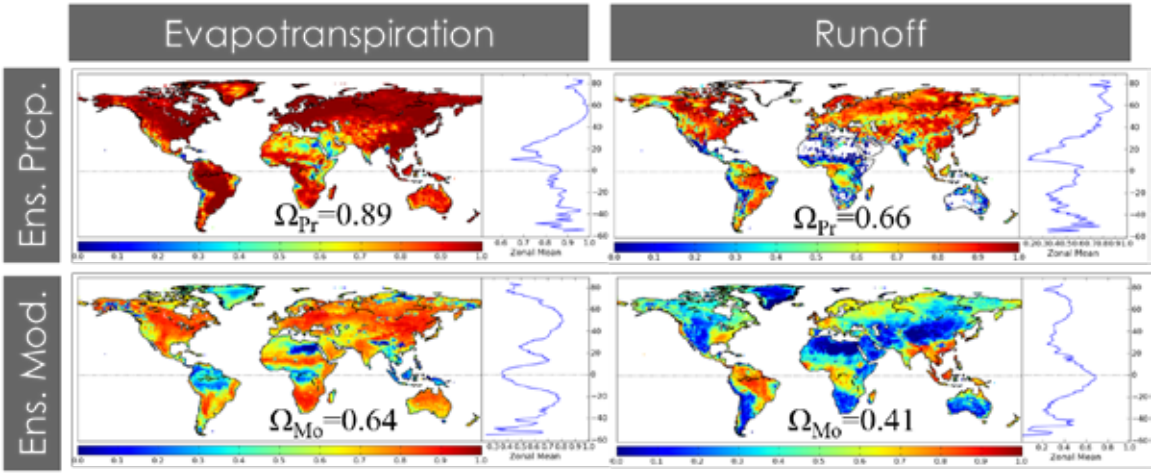


Figure D.3.1. Uncertainty in simulated evapotranspiration and runoff introduced by different land surface schemes in GSWP2 are larger than precipitation uncertainty-induced uncertainty by 28% and 40% in the similarity index (Ω) globally.

D.4 LUMIP

David M. Lawrence, Elena Shevliakova, and Atul K. Jain

The challenge of evaluating effects of land-use and land-cover change in the CMIP6 Land Use Model Intercomparison Project (LUMIP; Lawrence et al., 2016) is threefold:

1. Land use and land cover change (LULCC) is an external forcing that many CMIP6 experiments (e.g., DECK, historical, future scenario, and LUMIP) will be using, but the forcing data itself is complex, uncertain, and challenging to interpret and use with climate models and ESMs. Analysis of CMIP6 experiments should begin with an evaluation of the consistency between the CMIP6 LULCC scenario and its implementation in different ESMs (e.g., agricultural areas, extent of different crops, area and amount of wood harvested). Additional benchmarks need to be developed for stand-alone LUMIP experiments focused on effects of management on physical and biogeochemical states. Evaluation or benchmarking the CMIP6 LULCC reconstruction itself is crucial in attributing sources of ESM uncertainty/biases, particularly in regions with a long history or intensification of LULCC.

2. ESMs have dramatically different LULCC components, including types of land-use and land-management practices. Many LULCC parameters are not informed by data and do not capture historical patterns and practices (e.g., fraction of harvested residue and its fate). Furthermore, the relative importance of different types of land use and land management (e.g., wood harvest, prognostic crops, irrigation, fertilization, shifting cultivation, pasture representation, tilling, etc) and representation of response to disturbances are not fully understood from either observational or modeling perspectives.

3. LULCC affects many land processes and properties. Detection and attribution of LULCC effects are the major challenges for both models and observations, including impacts on
   - atmospheric CO_2,
   - ecosystem processes and states,
   - hydrology,
   - soil carbon and nutrient biogeochemistry,
   - vegetation dynamics, and
   - surface energy and BGC fluxes.

D.4.1 Land-use Metrics

A goal of LUMIP is to establish a useful set of model diagnostics that enables a systematic assessment of land use-climate feedbacks and improved attribution of the roles of both land and atmosphere in terms of generating these feedbacks. The need for more systematic assessment of the terrestrial and atmospheric response to land-cover change is one of the major conclusions of the LUCID studies. Boisier et al. (2012) and de Noblet-Ducoudré et al. (2012) argue that the different land use–climate relationships displayed across the LUCID models highlight the need to improve diagnostics and metrics for land surface model evaluation in general and the simulated response to LULCC in particular. These sentiments are consistent with recent efforts to improve and systematize land model assessment. LUMIP will promote a coordinated effort to develop biogeophysical and biogeochemical metrics of model performance with respect to land-use change that will help constrain model dynamics. These efforts dovetail with expanding emphasis in CMIP6 on model performance metrics.

Several recent studies have utilized various methodologies to infer observationally based historical change in land surface variables impacted by LULCC or divergences in surface responses between different land-cover types (Boisier et al., 2013, 2014; Lee et al., 2011; Lejeune et al., 2016; Li et al., 2015; Teuling et al., 2010; Williams et al., 2012). For example, Boisier et al. (2013) took MODIS albedo at 0.05° resolution and derived monthly albedo climatologies for croplands and four other land cover types. They then reconstructed the changes in surface albedo between preindustrial times and present-day by combining these climatologies with the land cover maps of 1870 and 1992 used in individual land models that participated in LUCID. The reconstructed albedo changes were then compared with the simulated albedo changes in the models. Because the same land cover change map is used in the reconstruction and in the simulations, one can infer that the differences in albedo change can be attributed to limitations in the parameterization of albedo in the models.

Another promising area for LULCC metrics development is with paired tower site analyses. Paired sites typically have one flux tower located in a forest and one in nearby open land (grassland, cropland, or open shrub). Differences in fluxes and states for these paired sites can be taken as representative of the impacts of local land cover change (deforestation in these cases). Lee et al. (2011), Chen and Dirmeyer (2016), and Burakowski et al. (2016) have all utilized paired sites to assess the impact of LULCC on surface temperature and to identify what processes are driving changes in surface temperature. Two important findings from these analyses are that daytime and nighttime responses differ, even in terms of their sign and that at different sites, the impact of LULCC can be attributed to different causes or combinations of causes (e.g., changes in roughness, albedo, and Bowen Ratio).

Several sources of data and methods with promise for LULCC metric development have been identified, including the following:

- Paired tower sites with known LULCC activities
- Food and Agriculture Organization of the United Nations (**http://www.fao.org/statistics/databases/en/**) and national (e.g., USDA Forest Service data, National Agricultural Statistics Service data) statistics
- Inferred impacts derived from any global dataset (e.g., albedo; see Boisier et al., 2013; Lejeune et al., 2016); compare nearby pixels that are mostly forest to mostly open land
- Water storage (Landerer and Swenson, 2012) and discharge from perturbed (managed) and unperturbed basins (Milly et al., 2014)
- Land use carbon fluxes and their components from bookkeeping models (Houghton, 2013; Richter and Houghton, 2011), global carbon project data sets (Le Quéré et al., 2015), RECCAP synthesis project (**http://www.globalcarbonproject.org/reccap**)

- Impact of LULCC in South and Southeast Asia (Adachi et al., 2011; Cervarich et al., 2016; Tao et al., 2013; Piao et al., 2012)
- Impact of LULCC on soil carbon and nitrogen; Review Analysis (Smith et al., 2016a)
- Global aboveground carbon estimates for both forest and non-forest biomes during the past two decades from satellite passive microwave observations (Liu et al., 2015)
- Fire emissions (van der Werf et al., 2010)

D.4.2 Land-only Versus Coupled Model Assessment

Importantly, the availability of both land-only and coupled historic simulations in CMIP6 will enable a more systematic assessment of the roles of the land and atmosphere in simulated responses to LULCC. With both coupled and uncoupled experiments with and without land-use change, LUMIP will be able to systematically disentangle the simulated LULCC forcing (i.e., changes in land surface water, energy, and carbon fluxes due to land-use change) from the response (i.e., changes in climate variables such as temperature and precipitation that are driven by LULCC in surface fluxes).

D.4.3 Subgrid Data Reporting and Analysis

New output data standardization for LUMIP will also enrich and expand analysis of model experiment results. Particular emphasis within LUMIP is on archiving subgrid land information in CMIP6 experiments, including LUMIP experiments and other relevant experiments from ScenarioMIP, C^4MIP, and the CMIP historical simulation. In most land models, physical, ecological, and biogeochemical land state and surface flux variables are calculated separately for several different land surface types or land management "tiles" (e.g., natural and secondary vegetation, crops, pasture, urban, lake, glacier). Frequently, including in the CMIP5 archive, tile-specific quantities are averaged and only grid-cell mean values are reported. Consequently, a large amount of valuable information is lost with respect to how each land-use type responds to and interacts with climate change and direct anthropogenic modifications of the land surface. LUMIP has developed a protocol and associated data request for CMIP6 for selected key variables on multiple land-use tiles (i.e., primary and secondary land, crops, pastureland, and urban).

Several recent studies have demonstrated that valuable insight can be gained through analysis of subgrid information. For example, Fischer et al. (2012) used subgrid output to show that not only is heat stress higher in urban areas compared to rural areas in the present day climate, but also that heat stress is projected to increase more rapidly in urban areas under climate change. Malyshev et al. (2015) found a much stronger signature of the climate impact of LULCC at the subgrid level (i.e., comparing simulated surface temperatures across different land-use tiles within a grid cell) than is apparent at the gridcell level. Subgrid analysis can also lead to improved understanding of how models operate. For example, Schultz et al. (2016) showed, through subgrid analysis of CLM, that the assumption that plants share a soil column and therefore compete for water and nutrients has the side effect of an effective soil heat transfer between vegetation types, which can alias into individual vegetation type surface fluxes. Furthermore, reporting carbon pools and fluxes by tiles will enable assessment of land-use carbon fluxes not only with the standard method of differencing land-use and no land-use experiments, but also within a single land-use experiment, utilizing bookkeeping approaches (Houghton et al., 2012), which allows a more direct comparison of observed and modeled carbon inventories.

D.5 MsTMIP

Christopher R. Schwalm

The North American Carbon Program (NACP) Multi-scale Synthesis & Terrestrial Model Intercomparison Project (MsTMIP) is a coordinated model intercomparison and evaluation effort designed to improve the diagnosis and attribution of carbon sources and sinks at local to global scales (Huntzinger et al., 2013). MsTMIP is distinct from CMIP because it focuses on the land component of ESMs. There are currently about 20 participating state-of-the-art LSMs in MsTMIP; each executed in offline mode using a standardized protocol (Wei et al., 2014a,b). This key design tenet of MsTMIP mandates that forcing data, boundary conditions, steady-state spin up, and sensitivity simulations are uniform across all models. Thus, inter-model spread is attributable solely to process representations, which permits

a skill-to-structure mapping. That is, since biophysical and biogeochemical representations are the only differences across models, changes in model skill can be attributed to model structures (Huntzinger et al., 2014).

MsTMIP is divided into two phases. The now-complete Phase I (Huntzinger et al., 2013; Wei et al., 2014a,b) was based on a set of retrospective semi-factorial runs where historical time-varying climate, CO2 concentration, land cover, and nitrogen deposition are sequentially "turned on" after steady-state. Each model completed a set of five runs with the final run having all factors enabled. Phase I runs were global (0.5° spatial resolution) from 1901 to 2010 at a monthly time step. Forcing data were based on the CRU-NCEP product with sub-monthly scale variability from the NCEP reanalysis merged with the CRU monthly fields (Wei et al., 2014a,b). The Phase I results from 15 LSMs are available online (Huntzinger et. al., 2016). In Phase I each model run was performed by the corresponding modeling team.

Phase II differs from Phase I in several ways. It focuses on the future, from present to the end of the 21st century (2011 to 2100), and forcing data are based on downscaled ESM meteorological fields from CMIP5. Each LSM is forced with 10 plausible climate futures using all possible combinations of two RCPs (4.5 and 8.5) and five ESMs (CMIP5 historical runs) chosen to reflect a range of temperature changes. The Phase I and Phase II forcing data boundary is smoothed to remove any discontinuities and to provide for a single time trajectory from 1901 to 2100. As the same set of semi-factorial runs is preserved in Phase II, 40 runs are required for each model. There are no additional steady-state runs; Phase II transient runs are initialized with the 2010 states from Phase I. In addition, models are executed centrally on the NASA JPL Model Farm, which contains a subset of all MsTMIP LSMs run with both standardized protocol and output code. The Model Farm offers greater flexibility than relying on separate teams to run their models, and it reduces financial, logistical, and interoperability challenges.

To date, MsTMIP simulations were used to (1) diagnose global patterns of soil organic carbon (Tian et al., 2015), (2) understand climatic vs. anthropogenic controls on evapotranspiration (Mao et al., 2015), (3) aggregate individual model results with benchmark-driven model ensemble integration (Schwalm et al., 2015), (4) quantify the net climate effect of the terrestrial biosphere (Tian et al., 2016), and (5) evaluate the impact of climate extremes on carbon cycling (Zscheischler et al., 2014). With retrospective Phase I and predictive Phase II simulations, MsTMIP can serve as a unified platform to evaluate how model structural differences, key controls of carbon metabolism, and plausible climate futures alter future carbon dynamics.

D.6 PLUME-MIP

Anders Ahlström

PLUME-MIP addresses the responses of vegetation and land surface models to environmental drivers under current and future projections, and attempts to advance the state-of-the-art in attributing modeled carbon cycle responses to underlying mechanisms, as represented in the models.

The project is divided into two main tiers as follows:

- Tier 1 involves standard transient simulations using bias-corrected CMIP5 climate outputs for the recent past and future under a set of CO_2 concentration pathways. The outcomes will be used to evaluate the different responses of the terrestrial C cycle to climate projections and CO_2 pathways.
- Tier 2 adopts a recently developed transient version of the Traceability Framework (TF) (Xia et al., 2013) to identify underlying causes of model differences in their responses to current and future climate forcing. The framework is designed to facilitate model intercomparisons by tracing components and their differences across models. Using the TF, Tier 2 will focus on locating the main carbon cycle processes that are responsible for causing differences among models and between models and data.

Currently Tier 1 simulations are nearly complete, and Tier 2 simulations are being performed or prepared. Methodology for applying the transient TF has been developed, tested, and partially published (Ahlström et al., 2015). In our analysis, we aim to answer two main research questions: (1) what is the relative role of main ecosystem processes in inter-model differences today and in the future? and (2) which processes are responsible for model–data inconsistencies and biases?

To answer (1), we will utilize results from Tier 1 and Tier 2 in combination with empirical data products on plant productivity, carbon pools, and turnover in a novel and transparent analysis. The transient TF can be used as an emulator that perfectly represents the flows of carbon between carbon pools while maintaining the model's structure. This way we can exchange processes (fluxes) between models (e.g., NPP, vegetation turnover, soil respiration rates) to identify what processes contribute to inter-model differences.

For (2), we will utilize TF to replace simulated processes with empirically derived data products (e.g., NPP, turnover, and respiration rates) and evaluate the resulting carbon pools against empirical datasets. The TF allows us to replace one or several processes on which independent data exists and recalculate carbon pools while preserving model structure and functioning from remaining processes. The resulting carbon pools will be evaluated against independent data using ILAMB benchmarking resources with the overall aim to find the processes (fluxes) and functions (e.g., soil respiration rates) responsible for model–data inconsistencies while identifying potential compensation between processes.

Both tiers and analysis steps (1) and (2) contribute to the goal of isolating the processes responsible for differences between models and their future projections and between models and data, using a transparent and systematic methodology.

Appendix E. Integration with Uncertainty Quantification Frameworks

E.1 An Uncertainty Quantification Framework Designed for Land Models

Maoyi Huang, Zhangshuan Hou, Jaideep Ray, Laura Swiler, L. Ruby Leung

Current-generation land models, such as the Community Land Model (CLM) and the Accelerated Climate Modeling for Energy Land Model (ALM), include numerous sub-models and associated parameters. The high-dimensional parameter space presents a formidable challenge for quantifying uncertainty and improving Earth system model predictions needed to assess environmental changes and risks. In practice, many parameters in land surface models are expected to vary from site to site and are poorly estimated or subjectively assigned. There is a strong need to calibrate/optimize the parameter values; however, with the high-dimensional parameter space, systematic calibration at numerous field sites is mission impossible because of the computational demand and the ill-posed nature of the inverse problems.

There have been attempts to calibrate LSMs. Because of their computationally expensive nature, ongoing efforts also target the construction of emulators (surrogate models) that map LSM's outputs to its inputs. The emulators can then be used (instead of the LSM itself) in sensitivity analysis, parameter estimation, propagation of parametric uncertainty and other many-query applications. Sargsyan et al. (2014) attempted to construct surrogates for five variables of interest from CLM4 with prognostic carbon and nitrogen modules turned on (i.e., CLM4-CN) using Bayesian compressive sensing (BCS) in combination with polynomial chaos expansions (PCEs). Müller et al. (2015) used an RBF to create a surrogate of the data–model mismatch and estimated 11 parameters of the CLM4.5's methane module using a global optimization method called DYnamic COordinate search using Response Surface models (DYCORS) (Regis and Shoemaker, 2007). Gong et al. (2015) used adaptive surrogate-based optimization to perform parameter estimation of 12 independent parameters in the CLM deterministically using six observables jointly.

Probabilistic methods, based on Monte Carlo simulations, have been used to calibrate LSMs. (Lo et al., 2010) used Monte Carlo techniques to estimate hydrological parameters of Community Land Model (CLM) 3.0, while Prihodko et al. (2008) calibrated Simple Biosphere Model version 2.5. Järvinen et al. (2010) and Solonen et al. (2012) used multi-chain Markov Chain Monte Carlo (MCMC) methods to address the formidable computational cost of calibrating the parameters of a climate model, while Zeng et al. (2013) used the same approach to calibrate the parameters of a crop module in CLM version 3.5. Bilionis et al. (2015) used a sequential Monte Carlo method to calibrate 10 parameters of the Crop module in CLM4.5. Tian and Xie (2008) used an unscented Kalman filter to calibrate CLM 2.0.

Significant progress has been made toward quantifying uncertainty associated with hydrologic parameters in the CLM and calibrating those parameters using an uncertainty quantification (UQ) framework. The framework features importance sampling, exploratory data analyses, HPC-enabled numerical simulations, classification of a complex system into a few relatively homogeneous regions, and Bayesian inversion using Markov Chain Monte Carlo techniques. The UQ framework has been applied to flux towers and watersheds under different climate and site conditions in the contiguous United States.

By performing numerical simulations using an efficient stochastic sampling-based sensitivity analysis approach, linear, interaction, and high-order nonlinear impacts of hydrologic parameters in CLM on simulated surface water and energy fluxes are analyzed via statistical tests and stepwise backward removal parameter screening at 13 selected flux tower sites (Figure E.1.1) and 431 river basins (Figure E.1.2) from the Model Parameter Estimation Experiment (MOPEX) in the United States (Hou et al., 2012; Huang et al., 2013; Ren et al., 2016). Based on this analysis, a subset of hydrological parameters (4 out of 10 being analyzed) have been identified to have significant impacts on latent heat, sensible heat, and runoff generation, and the results are consistent across all sites, as shown in Figure E.1.3. The reduction in parameter space through such an analysis establishes the foundation for inverse modeling, or parameter calibration. As a first attempt, Sun et al. (2013) implemented a single-chain Markov-Chain Monte Carlo (MCMC) inversion procedure with CLM and demonstrated that it was feasible to invert CLM hydrologic parameters at the site level, when observed fluxes and streamflow are used to constrain the parameters. However, the computational expense of CLM makes a single-chain MCMC method not plausible, as the simulations have to be conducted sequentially.

Figure E.1.1. Geographic locations of the selected flux towers. Adopted from Hou et al. (2012).

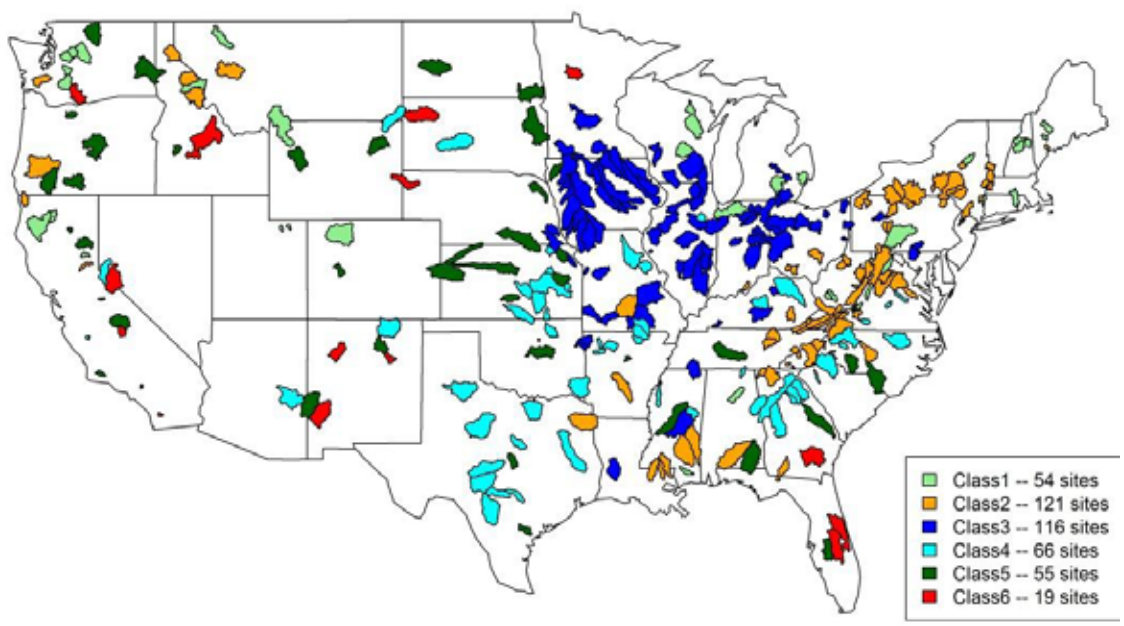


Figure E.1.2. Classes of the 431 MOPEX basins classified using parameter sensitivity scores with runoff as the response variable in the General Linear Model sensitivity analysis. Adopted from Ren et al. (2016).

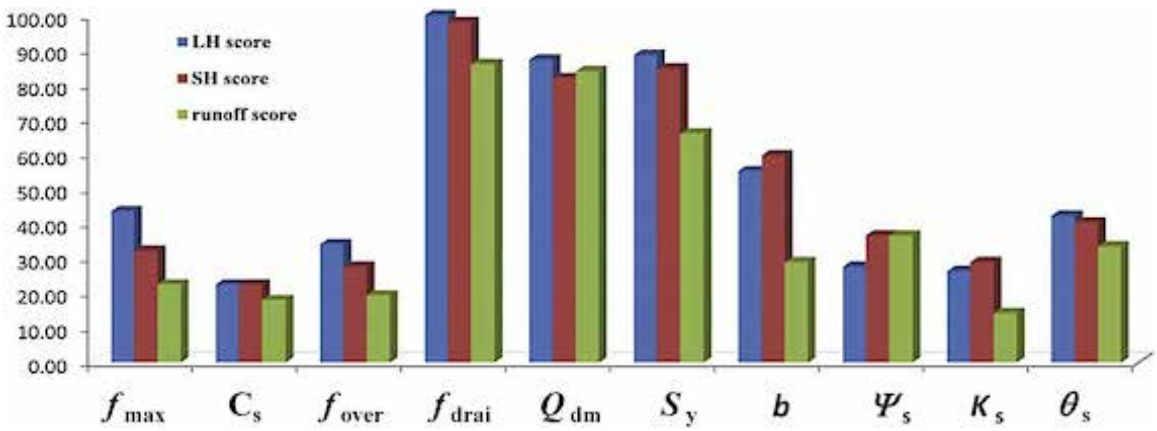


Figure E.1.2. Classes of the 431 MOPEX basins classified using parameter sensitivity scores with runoff as the response variable in the General Linear Model sensitivity analysis. Adopted from Ren et al. (2016).

To address this issue, there is a need to reduce computational costs and utilizing high-performance computing infrastructure. A Surrogate-based Markov chain Monte Carlo (MCMC)-Bayesian inversion approach has been developed for CLM and tested at 12 flux tower sites (Huang et al., 2016; Ray et al., 2015). The procedure starts with building surrogates using CLM4 simulations driven by perturbed parameter sets using a space-filling quasi-MC sampling approach. The surrogates, after careful validation and selection, are then used as computationally efficient alternatives to the CLM numerical simulator, for improving the estimates of the hydrological parameters, and therefore LH predictions, with quantified uncertainties. Given the large number of MOPEX basins and their wide geographic extent, parameter significance scores are used to classify the basins into different classes by grouping basins with similar parameter significance patterns into six unique classes. Each MOPEX basin can be assigned to a unique class, and then appropriate unknown parameters are to be included in the calibration. The unknown parameters are a reduced subset which makes the model calibration/optimization feasible (Ren et al., 2016). Efforts to further alleviate computational burdens to the model optimization efforts are on-going by evaluating similarity/transferability of parameters within each class.

However, it has been recognized that surrogate-based inversion is intrinsically subject to errors as a result of approximating a complex model using simplified functions, not to mention the potential risk of failures in building the surrogates due to the complex relationships between model parameters and outputs of interest (Huang et al., 2016). To address this limitation, a Scalable Adaptive Chain Ensemble Sampling (SAChES) method has been developed that seeks to collect the samples required to construct the probability density functions by combining the scalability of Differential Evolution Monte Carlo (DE-MC), a genetic algorithm, with the sampling efficiency of adaptive Metropolis-Hastings sampling. The core hypothesis is that the parameter space can be efficiently searched using a large number of loosely coupled Markov chains. SAChES has been integrated with CESM1.2 (the code foundation of ACME) (Swiler et al., 2015). The capability of SAChES to invoke a large number of chains simultaneously has its obvious attraction in high-dimensional inversions, i.e., when a gridded field, rather than a few model parameters, has to be estimated. Some studies have begun to explore whether SAChES could be used to estimate saturation fields using ground penetrating radar measurements, as well as to estimate saturation and porosity fields using seismic and electromagnetic response observations (Bao et al., 2016), with potential applications to highly spatially resolved models such as the coupling between CLM and the reactive transport code PFLOTRAN (Hammond et al., 2014).

To summarize, the global sensitivity analysis and Bayesian inversion procedures are useful tools for parameter estimation with uncertainty bounds, as well as for identifying potential model structural errors by extensively exploring the parameter space and comparing discrepancies between model predictions and observations. To successfully integrate such tools with land models, model reduction techniques are critically needed to make the problem tractable. Integrating such tools with the benchmarking datasets available in the International Land Model Benchmarking (ILAMB) framework (e.g., data from AmeriFlux network, streamflow gages, data products from the Moderate Resolution Imaging Spectroradiometer), would help the community to better constrain land model parameters and identifying model structural and parametric uncertainties. Although only being integrated with the CLM, the tools are general and therefore portable to other land models.

E.2 Use of Emulators in Uncertainty Quantification

George S. Pau

Quantifying uncertainties in land surface models (LSMs) is an important aspect of benchmarking exercises. Since observation data is inherently uncertain, one potential robust verification approach involves comparing the probability density functions of the observation data and the model outputs. The difficulty of quantifying the uncertainties in the observation data has been addressed elsewhere in this report. Here we focus on the task of quantifying the probability density functions of the model outputs. In particular, we consider the case where the nonlinearity in the model response necessitates the use of robust uncertainty quantification (UQ) techniques, especially Monte Carlo (MC) methods. Accurate statistical descriptions of model outputs also allow for more informative comparison between different LSMs.

MC methods require many evaluations of a LSM, each of which can be computationally challenging if modeled at the scale of the observation data. Brute force application of MC methods is typically infeasible even with existing high-end computing ecosystems because of the significant computational resources required. There is thus a need to develop MC methods that do not require a large number of LSM evaluations. Fortunately, there are many recent advances in MCMC methods and particle-based MC methods. Some new efficient methods include implicit particle filter (Chorin and Tu, 2009), stochastic Newton MCMC method (Martin et al., 2012), and MCMC methods that use Gibbs samplers (Kuczera et al., 2010), differential evolution samplers (Laloy and Vrugt, 2012), affine invariant ensemble samplers (Goodman and Weare, 2010) and surrogate-based samplers (Goodwin, 2015; Ray et al., 2015). These methods have varying degrees of parallelism that affect their efficient deployments on supercomputers. Apart from the surrogate-based samplers, the number of LSM evaluations is still typically very large.

In surrogate-based MC methods, surrogate models, built based on a limited number of LSM evaluations, are used as efficient emulators of the LSM. An offline-online computational framework allows UQ analyses to be performed efficiently at the desired spatial and temporal scales using surrogate models (online stage) through an amortization of the construction cost of these models (offline stage). The offline stage is computationally intensive because of the need to obtain outputs from a large number of LSM evaluations. The construction of the surrogate models from these outputs can also be computationally and memory intensive. An additional advantage of this computational framework is its efficient utilization of heavily shared high-performance computing resources. By executing the offline stage during the off-peak cycles, we are able to execute the online stage even during peak cycles. We can also execute the online stage on smaller machines with smaller user base and thus better throughput.

There are many approaches to constructing an appropriate surrogate. However, this task differs from the data mining challenges in the industry. First, we are emulating computationally expensive numerical models that are typically deterministic. We need a strong theoretical framework for using statistical emulators to describe results from these numerical models. Second, since we are emulating physical systems, outputs from the surrogate models must obey the constraints inherent in the physical systems. Third, we are typically data-limited; although each high-resolution numerical simulation produces a lot of data for a given scenario, the number of scenarios that we simulated is relatively small.

Several promising surrogate-modeling methods are currently being used to emulate the output of LSMs. For scalar quantities, popular methods include Gaussian process regression (Drignei et al., 2008; Edwards et al., 2011; Holden et al., 2010; Olson et al., 2012; Ray et al., 2015; Rougier et al., 2009), and polynomial chaos expansion (Liu et al., 2016b; Ray et al., 2015; Sargsyan et al., 2014). However, these methods cannot be directly applied to emulate field solutions due to the sheer number of outputs from high-resolution LSMs. A typical approach combines dimensional reduction techniques, such as proper orthogonal decomposition, with the scalar approaches mentioned above (Higdon et al., 2008; Liu et al., 2016a; Wilkinson, 2011). However, statistical models may have difficulties capturing the complex and nonlinear behavior of a LSM. In these cases, a coarse resolution numerical model can be used as a surrogate model. Downscaling techniques are then used to downscale the resulting outputs onto a high-resolution grid (Pau et al., 2014, 2016; Walton et al., 2015).

The use of surrogate models within an UQ framework poses several challenges. In particular, the required accuracy of a surrogate model depends on the chosen UQ method. For example, a two-stage Monte Carlo method (Ma et al., 2008) allows the use of a surrogate model with lower fidelity since it is only used to guide the selection of the parameters for performing a full model evaluation. However, a poorly constructed surrogate model can lead to a large number of full model evaluations, severely reducing the benefit of using a surrogate model. Directly substituting the full model in a MC method by a surrogate model will provide greater efficiency gain. However, the analysis can be meaningless if the surrogate model failed to adequately and consistently give accurate predictions within the range of uncertainty of the parameters (Goodwin, 2015). Increasing the number of training samples can increase the accuracy of the surrogate models but it reduces the net computational gains. A potential strategy is to choose a MC method that better constrains the parameter space in which the surrogate model needs to be accurate, thus reducing the number of training samples required (Liu et al., 2016b).

In conclusion, surrogate models have potential to reduce the computational cost of a MC method. However, more research is still needed to ensure the use of surrogate models within a MC method is robust, efficient, and theoretically sound.

E.3 Uncertainty Quantification in the ACME Land Model: Summary

Daniel M. Ricciuto, Khachik Sargsyan, Dan Lu, Jiafu Mao, Peter Thornton

Uncertainty about land surface processes contributes to a large spread in model predictions about the magnitude and timing of climate change in the 21st century. LSM's incorporate a diverse array of processes across various temporal and spatial scales, and they include a large number of uncertain parameters. Traditionally, land surface model output uncertainty has been estimated using multimodel ensembles such as CMIP5 (Friedlingstein et al., 2014b) or MsTMIP (Huntzinger et al., 2013), which combine uncertainties related to model structure, boundary conditions, and parameters. Improved understanding about the sensitivity of model outputs to specific parameters and processes, as well as the contribution of parametric uncertainties to overall prediction uncertainty, is of critical importance not only for directing future model development and measurements, but also for increasing the accuracy of future predictions.UQ methods that perform such analyses have advanced considerably in the last decade and may be successfully applied to complex LSMs. Ultimately, land-surface observations and benchmarks, including those from ILAMB, could be included in a UQ framework to optimize model parameters and further improve model predictions.

Global sensitivity analysis (GSA) or variance-based decomposition is a popular method to quantify the effects of model parameter uncertainties on specific quantities of interest (QoIs). Although a number of GSA methods exist (e.g., Sobol, 1993; Saltelli et al., 2006), many simulations are generally required, which is rapidly becoming computationally infeasible as the number of parameters increases. In complex land surface models, simpler one at a time (OAT) approaches, which vary parameters around nominal values of variables and do not require very large ensembles, have been applied (e.g., Zaehle et al., 2010). However, these results can be misleading if parameter interactions are important or if sensitivities vary significantly over the full multidimensional parameter space (Saltelli et al., 2004). Surrogate models, which use a set of basis functions to reproduce the behavior of a given model for a given QoI, can be used to estimate sensitivities with low computational cost. These surrogate models are often constructed using polynomial chaos (PC) expansions, which have gained popularity recently as convenient machinery for uncertainty representation and propagation (Ghanem and Spanos, 1991; Le Maitre and Knio, 2010), allowing analytical extraction of both single-parameter and joint-interaction sensitivities. However, for high-dimensional problems with many model parameters, the construction of the surrogate still requires an infeasible number of model evaluations because the number of basis terms is prohibitively large. This problem is resolved in this study by using a new algorithm that iteratively searches for the best set basis terms. The new algorithm, Weighted Iterative Bayesian Compressive Sensing (WIBCS), builds upon earlier PC surrogate-based sensitivity analysis (Sargsyan et al., 2014).

Here we apply this new method to perform GSA at 96 FLUXNET sites (Figure E.3.1) using the initially committed version 0 of the DOE Accelerated Climate Model for Energy (ACME), the land component of which is largely based on the CLM 4.5 (Oleson et al., 2013). These 96 sites cover a large range of climates, plant functional types, and other land surface characteristics. A total of 65 model parameters related to biogeophysics and biogeochemical cycling were varied randomly within uniform ranges justified by literature or expert judgment. In order to construct site-specific surrogate models, 3000 model simulations were performed for each site on the Titan supercomputer at the Oak Ridge Leadership Computing Facility, examining 5 QoIs: gross primary productivity, latent heat flux, net ecosystem exchange, vegetation biomass and soil organic matter carbon. We find for all PFTs, generally 15 or fewer parameters drive most of the variance in the outputs. Within a PFT for a given output, generally the same parameters appear as sensitive at each site while differences in parameters are evident among PFTs and different outputs (Figure E.3.2). The sensitivities of some parameters vary as a function of climate variables such as temperature or precipitation. This sensitivity analysis will serve as the basis for more focused, lower-dimensional studies leading to parameter calibration and improved land-surface model prediction at global scales.

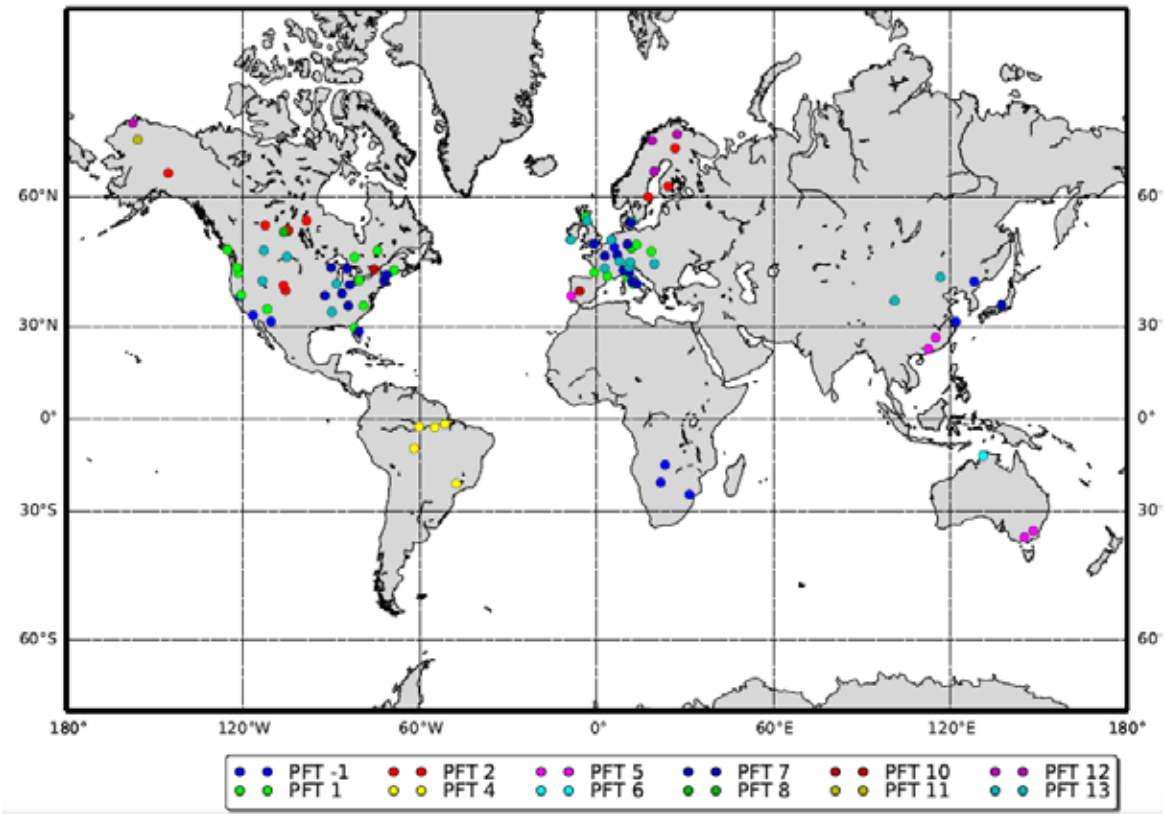


Figure E.3.1. Sites used in the global sensitivity analysis for the ACME land model at FLUXNET sites. Plant functional types at each site as used in the model are indicated.

Initial efforts to calibrate the ACME land model have been specific to individual eddy covariance or experiment sites, focus on a limited number of parameters, and do not estimate posterior uncertainties. We found that, by using 1 year of net ecosystem exchange (NEE) data from the Missouri Ozark flux site to optimize 14 model parameters, we were able to achieve a 30% reduction in root mean squared error in NEE over 2 subsequent years. However, when the calibrated parameters were used at the 2 similar deciduous forest sites Morgan Monroe State Forest and University of Michigan Biological Station, there was no increase in predictive skill compared to the model default parameters. However, when multiple QoIs are used in a calibration framework, the results are more promising (Mao et al., 2016; Ricciuto et al., 2011). Using the ILAMB framework, which contains a diverse set of data and benchmarks, for model calibration may significantly enhance the predictive skill of land surface models and begin to help explain or resolve the differences among models.

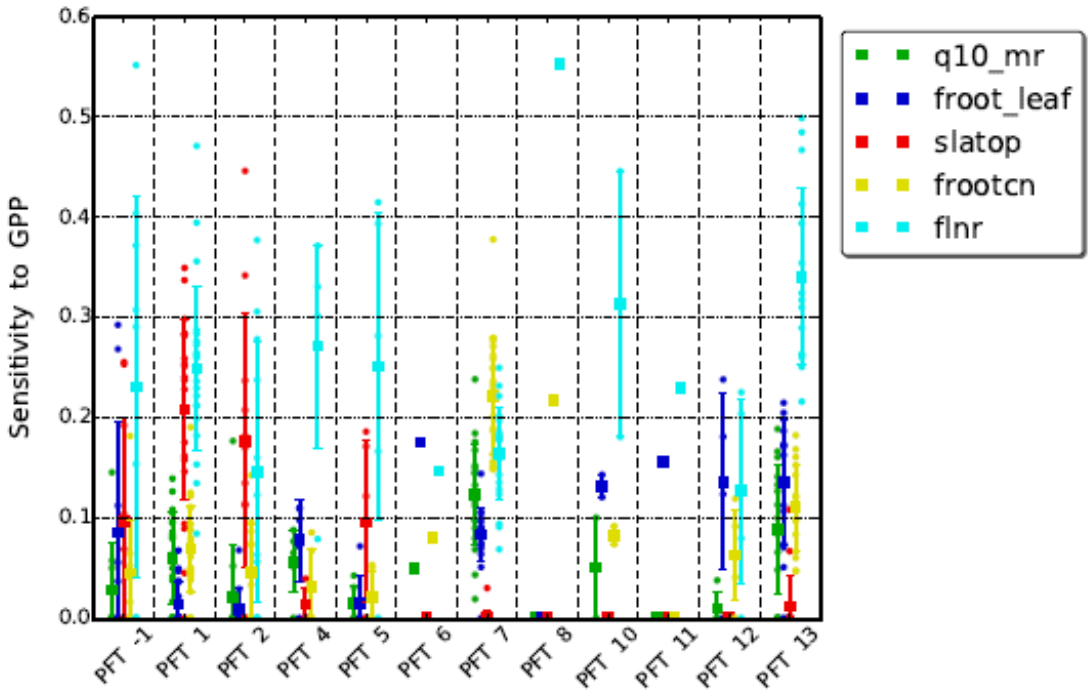


Figure E.3.2. Main effect sensitivity indices as a function of plant functional type (PFT) for gross primary productivity (GPP) for the five most sensitive parameters: the temperature sensitivity of maintenance respiration (*q10_mr*), the fine root to leaf allocation ratio (*froot_leaf*), the specific leaf area at the top of the canopy (*slatop*), the fine root carbon:nitrogen ratio (*frootcn*), and the fraction of leaf nitrogen in RuBisCO (*flnr*). Error bars indicate the standard deviation of the sensitivity index across multiple sites within a plant functional type.

E.4 The Predictive Ecosystem Analyzer (PEcAn): A Community Tool to Enable Land Model Synthesis, Evaluation, and Forecasting

Shawn Serbin, Michael Dietze, and the PEcAn Team

Process models are our primary tool for synthesizing our understanding of terrestrial ecosystems and projecting the impact of global change on ecosystem services associated with carbon, energy and water fluxes, and storage. Recently the use of models as a scaffold for data-driven synthesis has expanded and there is increasing interest in formal model–data experimentation (ModEx) frameworks to quantify uncertainties, evaluate models, enable the integration of observations, and guide model developments (Dietze et al., 2013). However, models remain inaccessible to most ecologists, in large part due to the informatics challenges of managing the flows of information in and out of such models. Moreover, the ecological sciences have witnessed an explosion in the amount and types of data that can be brought to bear on the potential responses of the terrestrial C, water, and energy cycles and biodiversity to global change. Many of the most pressing questions about global change are not limited by the need to collect new data as much as by our ability to synthesize and efficiently use existing data (Luo et al., 2011).

Because no one measurement provides a complete picture of terrestrial ecosystems, multiple data sources must be integrated in a sensible manner. Process-based models represent an ideal framework for integrating these data streams because they represent multiple processes at different spatial and temporal scales in ways that capture our current understanding of the causal connections across scales and among data types. Three components are required to bridge this gap between the available data and the required level of understanding: 1) state-of-the-art ecosystem models, 2) a workflow management system to handle the numerous streams of data, and 3) a data assimilation statistical framework to synthesize the data with the model.

Managing the communication between models and data involves three distinct challenges: 1) dealing with the volume of big data, 2) processing unstructured and uncurated long tail data, and 3) managing uncertainties in model–data comparisons and formal data–model assimilation. Finally, model development has long been an academic cottage industry, with different models lacking compatible formats for inputs, outputs, and settings. This has lead to redundant efforts and minimal reproducibility. As a result, the pace of model improvement has typically been slow. To address these challenges in modeling and model evaluation our group has developed the Predictive Ecosystem Analyzer (PEcAn, **http://pecanproject.org/**), a scientific tool box designed to automate many of the tasks and challenges required for conducting model–data ecoinformatics, which makes ecosystem modeling more accessible, analyses more automated and repeatable, and facilitates the evaluation of model projections, uncertainties, data-model fusion, forecasting, and decision support (Figure E.4.1). Model uncertainty quantification and propagation are a central part of PEcAn's design, which takes a Bayesian approach of treating model parameters and predictions as probability distributions and updating these distributions as new information becomes available (LeBauer et al., 2013; Dietze et al., 2014).

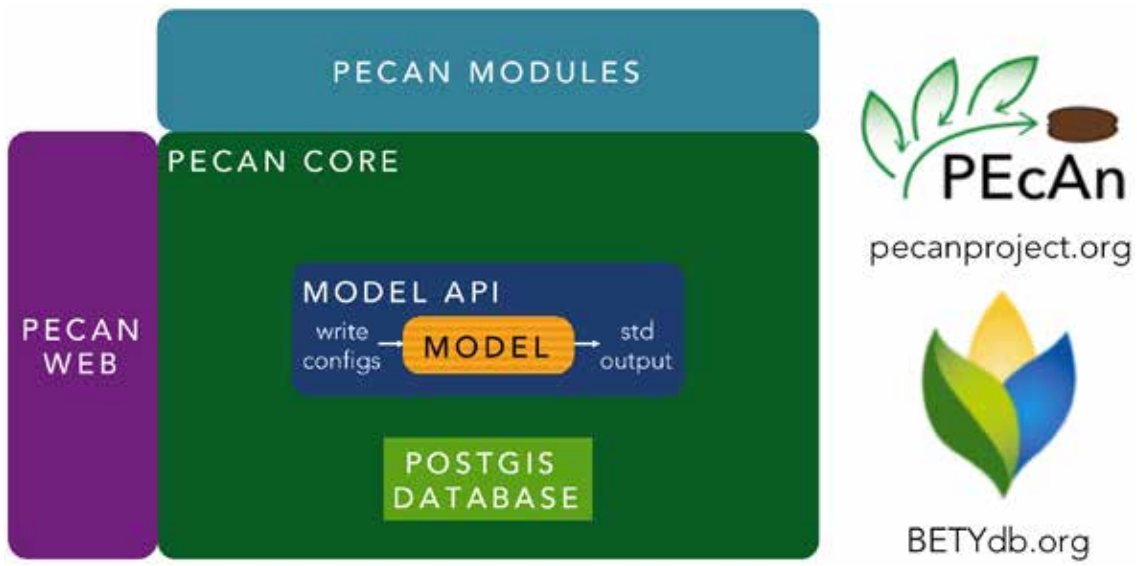


Figure E.4.1. Schematic representing the PEcAn framework for model–data integration and uncertainty quantification (LeBauer et al., 2013; Dietze et al., 2014). PEcAn provides a number of tools for standardization of model inputs and outputs, provenance tracking to enable repeatable and transparent analyses, distributed network and web accessible interface, and well as general reusable tools for extraction, analysis and visualization.

PEcAn users interact with models through an intuitive Google-Map-based interface (Figure E.4.2) and standardized file formats for model inputs (e.g., meteorological drivers, initial conditions), benchmarks, and outputs. Standardization allows the development of common, reusable tools for processing inputs, visualizing outputs, and automating the suite of analyses available within PEcAn. In addition, PEcAn includes state-of-the-art Hierarchical Bayesian tools for model parameterization, data assimilation, UQ and variance decomposition (VD). In addition to these tools, PEcAn leverages a PostGIS database network (Figure E.4.3; **https://www.betydb.org/**) to track all inputs, outputs, and model runs, greatly increasing reproducibility and reliability. Within the PEcAn network, the database syncs all results and facilitates file sharing to allow models to talk to each other and enables the community to effectively analyze many models distributed across a global network, thereby increasing the ability to conduct multi-model, multi-institutional model comparisons, synthesis, and evaluation activities.

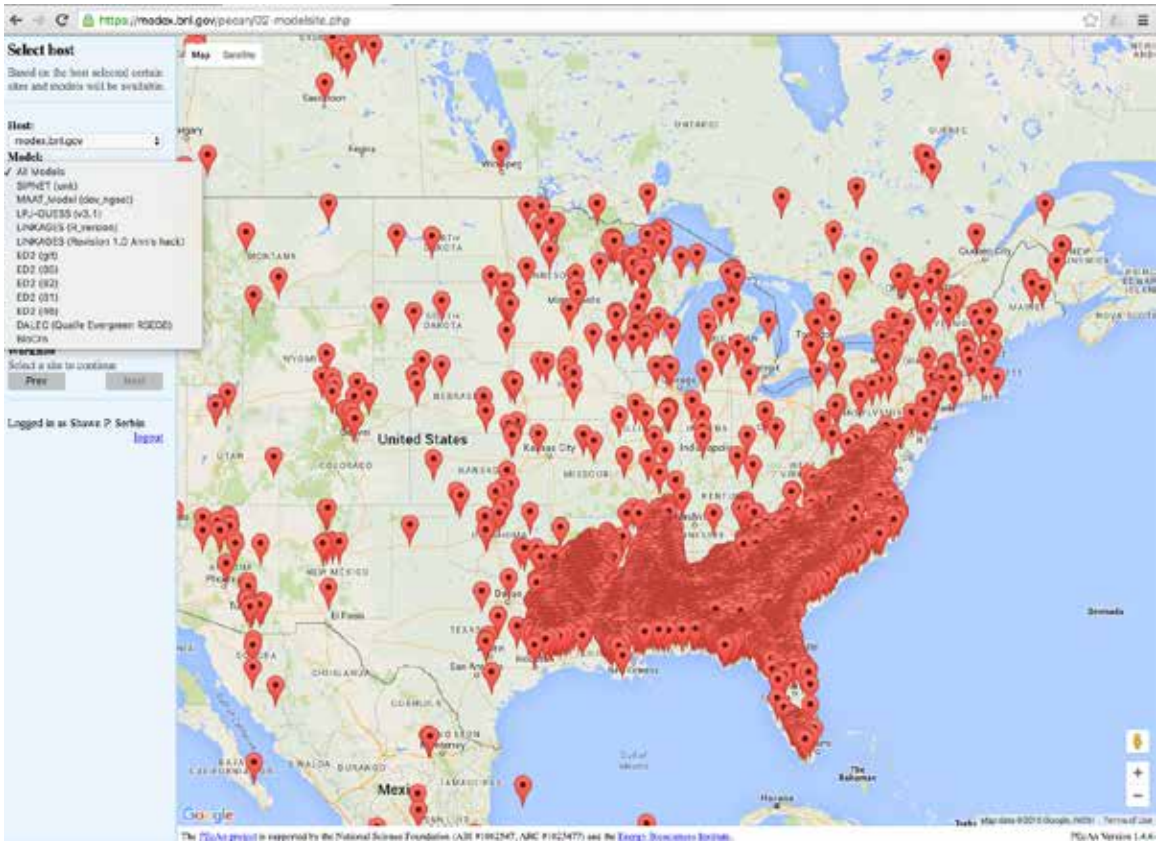


Figure E.4.2. The PEcAn framework provides a simple web-based graphical user interface (GUI) that leverages Google maps and PHP to link to the core PEcAn tools and PostGIS database (Figure E.4.1). Each node of the PEcAn framework (Figure E.4.3, this example from **https://modex.bnl.gov/**) serves up this interface which also serves to link model runs and results across the network. From this interface users can select sites, models, inputs, analyses (e.g., ensemble, UQ, data assimilation) and examine outputs with built in diagnostic plots or through an interactive R Shiny interface.

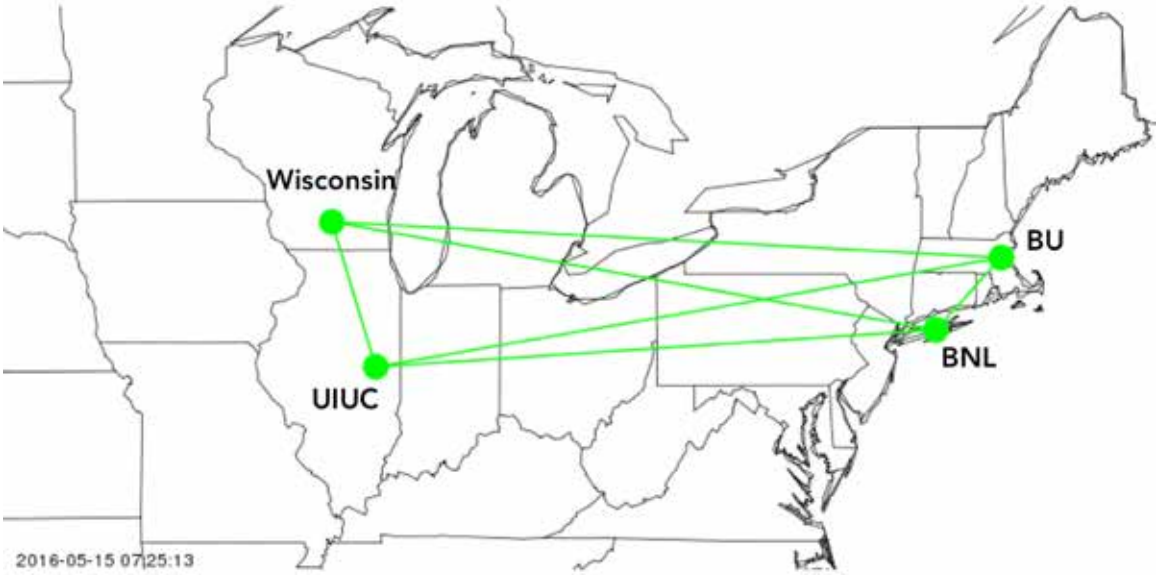


Figure E.4.3. An example status map (availible at https://pecan2.bu.edu/pecan/status.php) showing the current PEcAn network. Each node of the network shares data within the institutions database, model run history, and results.

Core components of the PEcAn framework include model parameterization and the quantification, propagation, and analysis of uncertainties (LeBauer et al., 2013; Dietze et al., 2014). These tools facilitate the efficient parameterization of models combining expert knowledge, trait observations, and field data to constrain plant functional types (PFTs). Within PEcAn the model uncertainty analysis workflow follows three automated steps: 1) a hierarchical Bayesian meta-analysis to summarize observational trait data and constrain ecosystem model parameters (Figure E.4.4), 2) a parameter sensitivity analysis, and 3) a variance decomposition analysis that uses the outputs from the first two steps to partition predictive uncertainty into the contributions from different model parameters. The workflow can also be repeated, without the first step, after iterative rounds of parameter data assimilation to assess the contribution of different data constraints to uncertainty reduction. A detailed description of this workflow can be found in LeBauer et al. (2013).

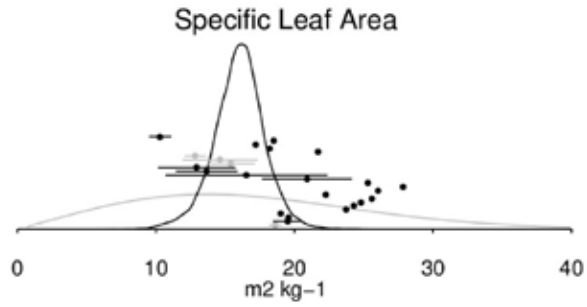


Figure E.4.4. Example PEcAn Bayesian meta-analysis result for specific leaf area (SLA, m^2 kg^{-1}). (adapted from LeBauer et al., 2013). The curves show the prior (gray) and posterior (black) distributions of SLA as selected from the PEcAn database (**https://www.betydb.org/**) for the perennial C_4 grass switchgrass (*Panicum virgatum*). Data from plants grown under an experimental treatment are presented in gray while data from field-grown plants under control treatments are in black. The posterior distribution is then used in the PEcAn uncertainty analysis to to generate the ecosystem model posterior based on the selected trait quantiles (Figure E.4.5).

Following the meta-analysis step, the PEcAn model sensitivity analysis consists of perturbations to the model parameters to evaluate how a specific model output (for example net primary productivity) changes as the parameter changes. The model perturbations are based on the quantiles of the parameter's posterior distribution, such that each parameter is moved in proportion to its uncertainty (Figure E.4.5). The quantiles are flexible and can be chosen by the user. The response function (i.e., model output as a function of a parameter value) for each parameter within each PFT is then approximated using a spline.

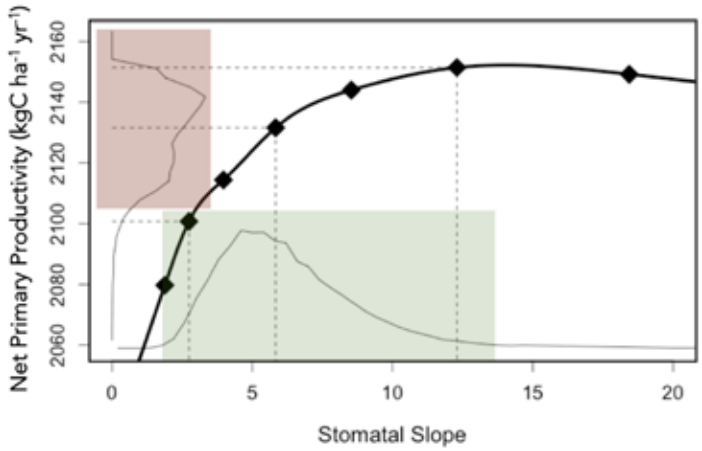


Figure E.4.5. Adapted from Dietze et al., (2014). Example uncertainty analysis for the 10 year mean NPP response of a typical temperate mid-successional hardwood plant functional type to the Ball-Berry stomatal slope parameter (Leuning, 1995). The probability density on the x-axis (green shaded area) captures the uncertainty in the stomatal slope parameter as estimated by the PEcAn Bayesian meta-analysis (Figure E.4.4). The solid diamonds represent the sensitivity analysis, depicting NPP projections using the Ecosystem Demography model (ED v2.2; Medvigy et al., 2009) for different values of stomatal slope, and the solid line is a spline fit to these points. The predictive uncertainty in NPP due to stomatal slope is represented by the probability density on the y axis (red shaded area), which is generated automatically within PEcAn by transforming the parameter distribution through the spline sensitivity function. Within PEcAn the partial variance is the variance of this predictive distribution divided by the sum of the variances across all parameters.

The PEcAn variance decomposition analysis estimates the uncertainty in model predictions (outputs) associated with each model parameter (inputs). A Monte Carlo generalization of the Delta method is used by transforming the posterior parameter distribution through the spline sensitivity function (Figure E.4.5). Because the predictive uncertainty is directly a product of parameter uncertainty and model sensitivity, these quantities are also automatically provided within the PEcAn UQ workflow (e.g., Figure E.4.6). To allow easier comparisons among variables, parameter variance and model sensitivity are expressed in dimensionless form as the posterior coefficient of variation and elasticity (sensitivity normed by both the parameter and output means), respectively. Moreover, PEcAn provides the predictive uncertainties associated with

each model parameter as the proportion that each variable contributes to the overall model predictive variance to enable direct comparisons across models, model parameters, and different model outputs.

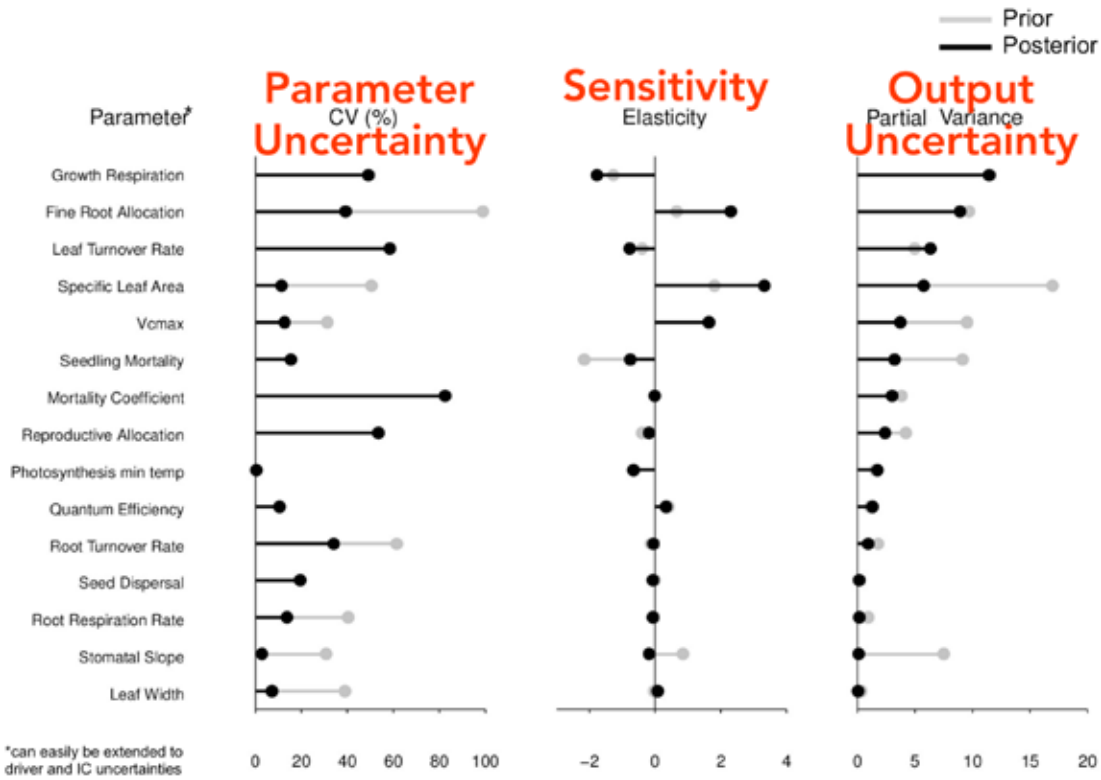


Figure E.4.6. Example PEcAn variance decomposition results presented for model runs before (gray) and following (black) the updating of model parameter estimates with species-level data from a PEcAn meta-analysis. Parameter Uncertainty: Uncertainty associated with each parameter is presented as the coefficient of variation and the degree to which some parameters have been constrained by species-level data is indicated by the reduction in CV in the black compared to the gray bars. Sensitivity: The sensitivity of modeled output to select parameters is presented as elasticity (normalized sensitivity; an elasticity of 1 indicates that model output will double when the parameter value doubles). Output Uncertainty: The contribution of each parameter to model uncertainty. This is a function of both the parameter variance and sensitivity. Parameters with both large CV and elasticity have the highest uncertainty.

Importantly, the results of the PEcAn uncertainty analysis workflow provide an understanding of the dominant drivers of uncertainty for outputs of interest (e.g., NPP). The information provided by PEcAn can be used to guide data synthesis, field campaigns, and Bayesian calibration. For example, an uncertainty analysis of the Ecosystem Demography model (ED2; Medvigy et al., 2009) across seventeen PFTs (Dietze et al., 2014), identified consistent patterns in the parameters driving model uncertainty (Figure E.4.7). In addition, the UQ/VD tools within PEcAn have been used to explore the impact of uncertainties in canopy radiative transfer on the projections of ED2 carbon, water, and energy fluxes and storage (Figure E.4.8; Viskari et al., in prep). This ongoing work is highlighting the need for better constraint on the representation of canopy radiative transfer within models to reduce uncertainties in associated processes such as photosynthesis.

An additional core component of the PEcAn framework, which is highly relevant to ILAMB and other model evaluation, benchmarking, and calibration activities, are the formal model–data assimilation workflows. Within PEcAn, users can make use of both parameter and state data assimilation with a range of approaches and algorithms. Parameter data assimilation (PDA) is used to update prior model parameter distributions based on a Likelihood function that quantifies how the error between model outputs and observed data changes as parameters are varied (Shiklomanov et al., 2016). By contrast, model state-variable data assimilation (SDA) uses observations (with their associated uncertainties) to constrain model states (e.g., vegetation composition, leaf area index, carbon stocks; e.g., Viskari et al., 2015) instead of model parameters. The core of SDA is the forecast/analysis cycle. In the forecast step the model states are predicted forward with uncertainties. In the analysis step, the model forecast is treated as the prior and updated based on the Likelihood of new observations (Figure E.4.9). Following the integration of data

the total forecast uncertainty is lowered than that from either the model or data alone. In addition, when conducted over a region, locations without observations are updated based on their covariances with measured locations. Similarly, covariances among modeled states are also used to update unobserved model state variables (e.g., the relationship between canopy cover, a remotely sensed property, and aboveground biomass). Taking a Bayesian approach to data assimilation within PEcAn allows for an iterative approach to both parameter and state assimilation, where analyses can be updated when new data is added without having to rerun analyses from scratch.

A priority highlighted in this report is the capacity to benchmark against and directly assimilate remotely sensed observations, such as surface reflectance. Remote sensing observations can be used to track seasonal and inter-annual changes in vegetation structure and function (Schmid et al., 2015). While existing benchmarks focus on comparing model outputs to derived data products, an important alternative is for models to output a full spectral signature. This "sensor simulator" approach (e.g., Figure E.4.10) would enable the direct comparison of model output to remote sensing observations (from leaf to regional scales) which also assures a consistency between the terrestrial biosphere model (TBM) output and the data, as data derived from remote sensing products (e.g., LAI) inevitably involves assumptions that are rarely identical to the assumptions of the TBM. Moreover, this approach facilitates more rapid inclusion of new data as it becomes available since it does not require the generation of derived data products (and the associated uncertainties that are often difficult to adequately quantify) and can easily be applied to sensors as they come online. Importantly, PEcAn already has this functionality for the ED2 model (Figure E.4.10; Viskari et al, in prep) which could be expanded to include other TBMs as needed. Coupling this functionality with ILAMB would further enable the coordination of model benchmarking and synthesis activities that

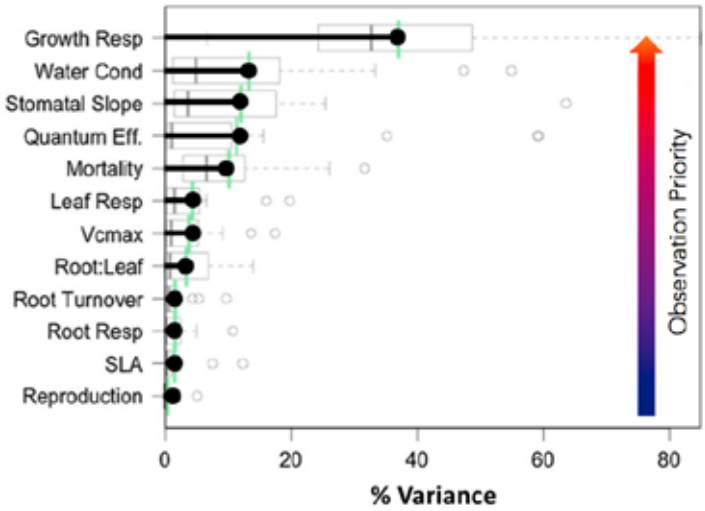


Figure E.4.7. Example ED2 multi PFT multi biome UQ synthesis conducted within the PEcAn framework (Adapted from Dietze et al., 2014). This example illustrates what parameters still dominate model uncertainty in NPP following a trait meta-analysis to constrain model parameters. It was found that the priority for improved model representation and parameterization was growth respiration, but also bulk water conductance from the soil, leaf stomatal slope, the quantum efficiency of photosynthesis, and plant mortality also dominated the model uncertainty across the PFTs.

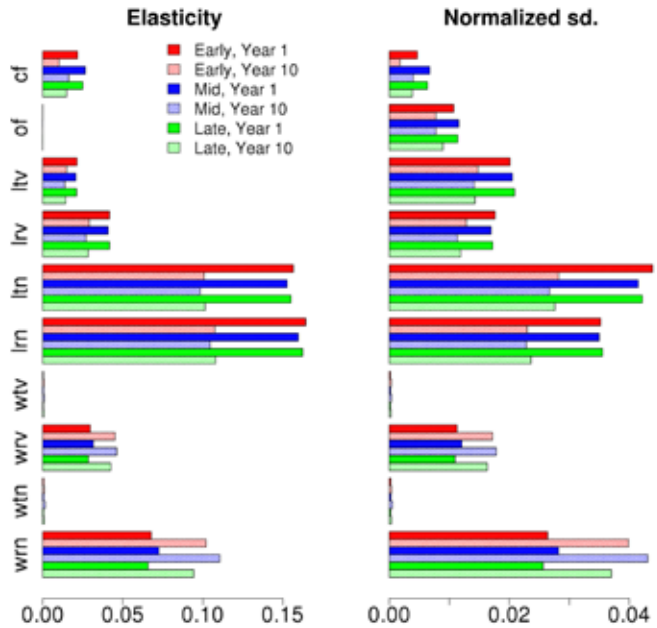


Figure E.4.8. Example PEcAn variance decomposition of ED2 canopy albedo showing the impact of uncertainty in model radiative transfer parameterization including leaf and stem optical properties, orientation, and clumping factors for early, mid, and late hardwood broadleaf PFTs in the first year (full) and tenth year (shaded) of the simulation. These results show the importance of evaluating, benchmarking, and constraining underlying processes and structures such as light harvesting and utilization as well as the more commonly explored outputs such as plant growth, dynamics, and seasonality / LAI. Adapted from Viskari et al., (in prep) and funded by NASA TE #NNX14AH65G.

have been identified as a critical need by the modeling community.

There are several important ways the ILAMB and PEcAn projects could collaborate and share tools, resources, and workflows for analyzing and benchmarking models at the site and regional scales. A key strength of the PEcAn package is the strong focus on the cyberinfrastructure, scientific workflows, provenance tracking, and on-demand multi-model synthesis capabilities. For example, the PEcAn network contains greater than ten ecosystem models that can be run locally or through the web interface one-at-a-time or together to produce a custom model intercomparison project (MIP). Furthermore, a number of additional models are in the process of being integrated, which consists of developing the software wrappers to manage and standardize the flows of information into and out of each model. This allows end-users the ability to easily run site-level/multi-site model/multi-model simulations and perform experiments that typically require significant investments in software, hardware and personnel. On the other hand, ILAMB has strong model benchmarking, diagnostics, and model evaluation tools that could be leveraged by other tools such as PEcAn. In addition, visualization tools within ILAMB are useful outside of the ILAMB package. Furthermore, the tools within ILAMB to quantify changes in model output due to code updates, initial conditions, or meteorological drivers are key for frameworks such as PEcAn since they provide the capability to understand different sources of uncertainty beyond model parameters and structure. Therefore, coupling ILAMB and PEcAn into a synthetic virtual framework would serve to significantly expand the model evaluation capabilities available to the community and avoid any potential redundancies in software development. Importantly ILAMB has historically been focused on the Earth system models (ESMs) at the centennial scale but is shifting focus to include regional and site-level evaluation with a more process-oriented focus, which was highlighted as an important need at this workshop. The ability to leverage tools within ILAMB and PEcAn would provide a framework for conducting shorter timescale but focused model benchmarking activities, including the leveraging of the existing and/or proposed ILAMB metrics such as functional relationships. Finally, a key recommendation for ILAMB was to provide model–data assimilation capabilities to facilitate observationally constrained model hindcasting in order to produce the best initial conditions for future forecasts. PEcAn already contains a suite of tools for parameter and state variable assimilation that could be leveraged in the future.

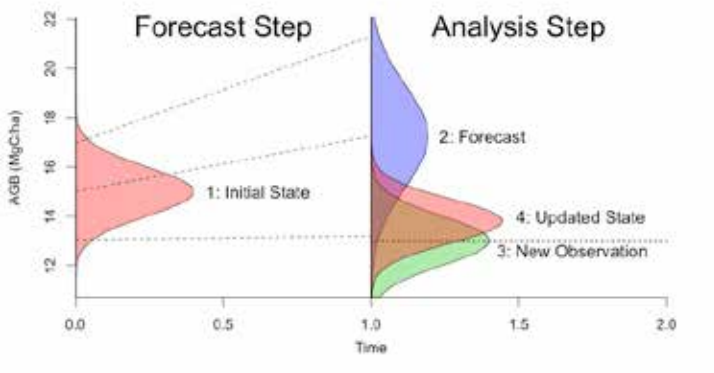


Figure E.4.9. Simplified example of the PEcAn state data assimilation (SDA) forecast/analysis cycle used to inform model projections within PEcAn. Adapted from Dietze 2017 Ecological Forecasting.

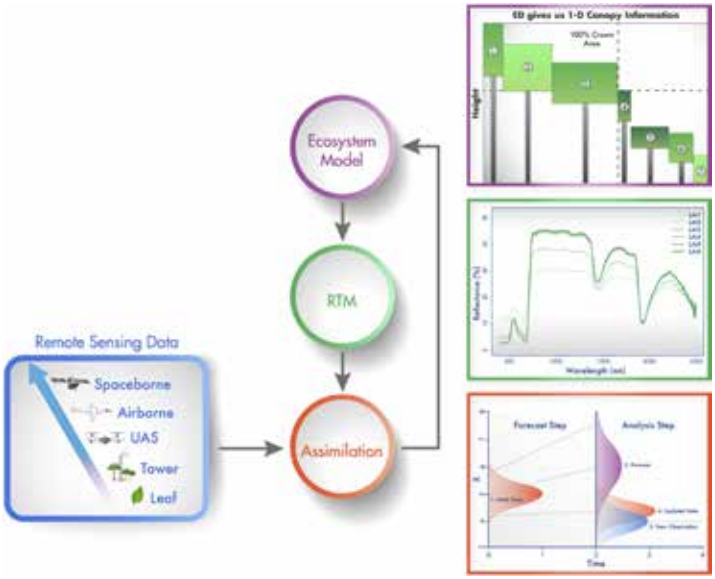


Figure E.4.10. Example of the use of an "sensor simulator" within a TBM (in this case ED2) to facilitate direct assimilation of and/or benchmarking against remote sensing observations within the PEcAn framework (Viskari et al., in prep). In this approach the output TBM spectral signature is based on the internal model structure (i.e. canopy biomass, height, RT properties) and compared with comparable remote sensing observations (i.e. surface reflectance, albedo). This allows for direct comparison and evaluation of associated processes such as photosynthesis, energy balance, surface temperature and evapotranspiration as well as identify uncertainties and areas to target for model improvement.

Appendix F.
ILAMB 2016 Workshop Materials

F.1 Agenda

May 16–18, 2016, DoubleTree by Hilton Hotel Washington DC
1515 Rhode Island Avenue, NW, Washington, DC 20005-5595, USA

Monday, May 16, 2016

7:00	**Breakfast**	Ballroom Lobby
8:00	**Welcome, Introductions, and Safety** – *Renu Joseph*	Terrace Ballroom
	8:00 Welcome and Safety – *Renu Joseph and Dorothy Koch*	
	8:05 U.S. Dept. of Energy (DOE) Research – *Sharlene Weatherwax*	
	8:15 DOE Climate Research Priorities – *Gary Geernaert*	
	8:25 DOE RGCM Program – *Renu Joseph*	
	8:35 DOE ESM Program – *Dorothy Koch*	
	8:45 Biogeochemistry–Climate Feedbacks SFA – *Forrest M. Hoffman*	
	8:55 Accelerated Climate Modeling for Energy (ACME) – *William J. Riley*	
	9:05 Workshop Charge and Reporting – *James T. Randerson*	
9:10	**Plenary Presentations on Benchmarking Tools** – *David M. Lawrence*	Terrace Ballroom
	9:10 P.1 Protocol for the Analysis of Land Surface models (PALS) – *Gab Abramowitz*	
	9:20 P.2 PLUMBER: PALS Land sUrface Model Benchmarking Evaluation pRoject – *Martin Best*	
	9:30 P.3 Towards efficient and systematic model benchmarking in CMIP6 – *Peter Gleckler*	
	9:50 P.4 Land surface Verification Toolkit (LVT): A formal benchmarking and evaluation framework for land surface models – *Sujay Kumar*	
	10:10 P.5 The International Land Model Benchmarking (ILAMB) Package – *James T. Randerson, Forrest M. Hoffman, and David M. Lawrence*	
10:30	**Morning Break**	Ballroom Lobby
11:00	**Plenary Discusson on Model Evaluation** – *Gretchen Keppel-Aleks*	Terrace Ballroom
	11:00 Summary of Evaluation Methods at Modeling Centers – *Gretchen Keppel-Aleks*	
	11:15 Discussion on Model Evaluation – *David M. Lawrence*	
11:50	**Plenary Presentations on Emergent Constraints and Evaluation Metrics I**	Terrace Ballroom
	11:50 P.6 Evaluation of vegetation cover and land-surface albedo – *Victor Brovkin*	
	12:10 P.7 Judging the dance contest – Metrics of land–atmosphere feedbacks – *Paul Dirmeyer*	
12:30	**Working Lunch**	Ballroom Lobby
13:30	**Metrics Breakout Group Meetings I** – *James T. Randerson*	
	Ecosystem Processes and States – *Nancy Y. Kiang and Ben Bond-Lamberty*	Terrace Ballroom
	Hydrology – *Randal Koster and Hongyi Li*	Directors Room
	Atmospheric CO_2 – *Gretchen Keppel-Aleks and William J. Riley*	Congressional Room
15:00	**Afternoon Break**	Ballroom Lobby
15:20	**Metrics Breakout Group Meetings II** – *Forrest M. Hoffman*	
	Soil Carbon and Nutrient Biogeochemistry – *Gustaf Hugelius and Jinyun Tang*	Terrace Ballroom
	Surface Fluxes (Energy and Carbon) – *Scott Denning and Dan Ricciuto*	Directors Room
	Vegetation Dynamics – *Rosie Fisher and Chonggang Xu*	Congressional Room

16:50	**Breakout Group Reports** (1–3 datasets, 1–3 new metrics, and bibliographies)	Terrace Ballroom
	16:50 Ecosystem Processes and States	
	16:55 Hydrology	
	17:00 Atmospheric CO_2	
	17:05 Soil Carbon and Nutrient Biogeochemistry	
	17:10 Surface Fluxes (Energy and Carbon)	
	17:15 Vegetation Dynamics	
17:20	**Poster Lightning Presentations**	Terrace Ballroom
18:00	**Poster Session and Reception**	
	Posters A.1 through A.8	Terrace Ballroom
	Posters B.1 through B.8	Directors Room
	Posters C.1 through C.8	Congressional Room
20:00	**Adjourn for the Day**	

Tuesday, May 17, 2016

7:00	**Breakfast**	Ballroom Lobby
8:00	**Keynote Presentation**: P.8 Role of flux networks in benchmarking land atmosphere models – *Dennis Baldocchi*	Terrace Ballroom
8:30	**Plenary Presentations on MIP Benchmarking Needs** – *William J. Riley*	Terrace Ballroom
	8:30 P.9 Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) Experimental Design and Organisation – *David M. Lawrence*	
	8:45 P.10 Assessing feedbacks for the Coupled Climate–Carbon Cycle Modeling Intercomparison Project (C^4MIP) – *Forrest M. Hoffman*	
	9:00 P.11 The Land Surface, Snow and Soil moisture Model Intercomparison Project (LS3MIP) and Global Soil Wetness Project Phase 3 (GSWP3) – *Hyungjun Kim*	
	9:15 P.12 Landuse and landcover change model performance metrics for LUMIP – *David M. Lawrence*	
	9:30 P.13 Multiscale Synthesis & Terrestrial Model Intercomparison Project: From cohort to insight – *Christopher R. Schwalm*	
	9:45 P.14 Processes Linked to Uncertainties Modelling Ecosystems (PLUMEMIP) – *Anders Ahlström*	
	10:00 Discussion – *Peter Gleckler*	
10:30	**Morning Break**	Ballroom Lobby
11:00	**Plenary Presentations on Emergent Constraints and Evaluation Metrics II**	Terrace Ballroom
	11:00 P.15 New benchmarks for northern high latitudes – *Charles D. Koven*	
	11:15 P.16 Permafrost Benchmarking System (PBS) – *Kevin Schaefer*	
11:30	**Breakout Groups on CMIP6 Evaluation Priorities (pre-lunch)** – *Gretchen Keppel-Aleks*	
	C^4MIP – *James T. Randerson and Charles D. Koven*	Terrace Ballroom
	LS3MIP – *Jiafu Mao and Andrew Slater*	Directors Room
	LUMIP – *Elena Shevliakova and Atul K. Jain*	Congressional Room
12:30	**Working Lunch**	Ballroom Lobby
11:00	**Breakout Groups on CMIP6 Evaluation Priorities (post-lunch)** – *Gretchen Keppel-Aleks*	
	C^4MIP – *James T. Randerson and Charles D. Koven*	Terrace Ballroom
	LS3MIP – *Jiafu Mao and Andrew Slater*	Directors Room
	LUMIP – *Elena Shevliakova and Atul K. Jain*	Congressional Room

14:00	**Breakout Group Reports** (1–3 datasets, 1–3 new metrics, and bibliographies) 14:00 C^4MIP 14:10 LS3MIP 14:20 LUMIP	Terrace Ballroom
14:30	**Keynote Presentation**: P.17 Theory-enabled model evaluation and improvement – *Yiqi Luo*	Terrace Ballroom
15:00	**Global Synthesis Discussion** – *Sha Zhou and Chris Lu*	Terrance Ballroom
15:15	**Afternoon Break**	Ballroom Lobby
15:45	**ILAMB v1 Package Demonstration and Application** – *Mingquan Mu*	Terrace Ballroom
16:45	**ILAMB v2 Package Tutorial / Training Session** – *Nathan Collier*	Terrace Ballroom
18:00	**Dinner on your own**	Downtown DC

Wednesday, May 18, 2016

7:00	**Breakfast**	Ballroom Lobby
8:00	**Plenary Presentations on Emergent Constraints and Evaluation Metrics III** 8:00 P.18 Evaluating the simulations of global nutrient cycles: Available observations and challenges – *Ying-Ping Wang* 8:20 P.19 Empirically derived sensitivity of vegetation to climate as a possible functional constraint for process based land models – *Gregory Quetin* 8:40 P.20 Some suggestions on emergent constraints and metrics on model evaluations over land – *Xubin Zeng* 9:00 P.21 Decomposition of CO_2 fertilization effect into contributions by land ecosystem processes: Comparison among CMIP5 Earth system models – *Kaoru Tachiiri*	Terrace Ballroom
9:20	**Breakout Groups on Next Generation Benchmarking Challenges and Priorities I** – *James T. Randerson* Process-specific experiments (litterbags, ^{14}C) – *Mathew Williams and Jianyang Xia* Metrics from extreme events – *Hyungjun Kim and Maoyi Huang* Design of new perturbation experiments – *Martin De Kauwe and Ankur Desai*	 Terrace Ballroom Directors Room Congressional Room
10:30	**Morning Break**	Ballroom Lobby
11:00	**Breakout Groups on Next Generation Benchmarking Challenges and Priorities II** – *David M. Lawrence* High latitude processes – *Kevin Schaefer, Charles D. Koven, and Umakant Mishra* Tropical processes – *Nathan McDowell and Paul Moorcroft* Global remote sensing – *David Schimel and Shawn Serbin*	 Terrace Ballroom Directors Room Congressional Room
12:10	**Breakout Group Reports** (1–3 datasets, 1–3 new metrics, and bibliographies) 12:10 Process-specific experiments 12:15 Metrics from extreme events 12:20 Design of new perturbation experiments 12:25 High latitude processes 12:30 Tropical processes 12:35 Global remote sensing	Terrace Ballroom
12:40	**Working Lunch**	Ballroom Lobby

13:40	**Plenary Presentations on Uncertainty Quantification (UQ) Methods** – *Forrest M. Hoffman*	Terrace Ballroom
	13:40 P.22 An uncertainty quantification framework designed for land models – *Maoyi Huang*	
	13:50 P.23 Use of emulators in uncertainty quantification – *George Pau*	
	14:00 P.24 Uncertainty quantification in the ACME land model – *Dan Ricciuto*	
	14:10 P.25 PEcAn: A community tool to enable synthesis, evaluation & forecasting – *Shawn Serbin*	
14:20	**Prioritizing Next Steps** – *James T. Randerson*	Terrace Ballroom
14:40	**Workshop Report Organization and Writing Assignments** – *Forrest M. Hoffman*	Terrace Ballroom
15:00	**Afternoon Break**	Ballroom Lobby
15:30	**Parallel Sessions on the ILAMB Packages and a Global Synthesis**	
	ILAMB v2 Package Tutorial / Training Session – *Nathan Collier*	Terrace Ballroom
	Global Synthesis Discussion (Continued from Tuesday) – *Yiqi Luo*	Directors Room
	ILAMB v1 Package Demonstration and Application – *Mingquan Mu*	Congressional Room
17:00	**Adjourn the Meeting**	

F.2 Plenary Presentation Abstracts

F.2.1 Benchmarking Tools

P.1 Protocol for the Analysis of Land Surface models (PALS)

Gab Abramowitz[1;2;†]

[1]University of New South Wales, Sydney NSW 2052, Australia
[2]Australian Research Council Centre of Excellence for Climate System Science (ARCCSS), Sydney NSW 2052, Australia
[†]Author to whom correspondence should be addressed; e-mail: gabsun@gmail.com

An increasing number of land surface model evaluation packages are becoming available, including ILAMB, LVT, EMSValTool and others. The first phase of the PALS web application also represented a something of a limited attempt at a standardised evaluation package, but was restricted to site-based evaluation and benchmarking. PALS facilitated the PALS Land sUrface Model Benchmarking Evaluation pRoject (PLUMBER; a MIP), also discussed at this meeting, and in particular promoted the use of empirical benchmarking as a way of defining model performance expectations.

With the arrival of the more comprehensive evaluation packages listed above, what have we learnt from PALS that is still of use? This presentation will focus in particular on the benefits of bringing tools such as these into an online web-based environment. These benefits include:

- ability to quickly and easily compare results internationally
- potential for better capture of simulation provenance information, increasing reproducibility
- simplicity and speed of creating MIPs
- MIPs can continue indefinitely, since they can be automated
- the ability to keep evaluation datasets for evaluation only (i.e. not calibration)
- identification of systematic performance issues across different models internationally
- new analyses can be applied to retrospectively to past simulation submissions
- ability to access archived historical model performance information
- increased transparency

Difficulties include sufficiently rigid i/o standards to enable automated analysis of model outputs, as well as intellectual property and security issues. Development of a second phase of a PALS-like environment that could incorporate a range of different analysis packages will also be discussed.

P.2 PLUMBER: PALS Land sUrface Model Benchmarking Evaluation pRoject

Martin Best[1;†], Gab Abramowitz[2;3], and Andy Pitman[2]

[1]UK Met Office, Exeter, EX1 3PB, UK
[2]University of New South Wales, Sydney NSW 2052, Australia
[3]Australian Research Council Centre of Excellence for Climate System Science (ARCCSS), Sydney NSW 2052, Australia
[†]Author to whom correspondence should be addressed; e-mail: martin.best@metoce.gov.uk

Many studies make the claim of undertaking model benchmarking. Unfortunately, there is often confusion about what "benchmarking" means; some undertake true benchmarking, others are undertaking the more traditional evaluation or comparison activities. In this presentation we will attempt to clarify the differences between the three approaches and demonstrate how the interpretation of model results can differ depending on which of the three measures of model performance are used. To enable this, data from the land surface benchmarking experiment PLUMBER (PALS Land sUrface Model Benchmarking Evaluation pRoject) are used.

In addition, a brief overview of the PLUMBER experimental protocol will be presented along with the key findings from the experiment to date. All land surface models had a consistent performance compared to the set of benchmarks when using standard statistical measures. These results demonstrated that the current day models perform better than older physical models, hence as a community we have progressed our knowledge over the last few decades. However, none of the models out performed the empirical benchmarks, with the models worse than a three variable piecewise linear regression for latent heat flux, but worse than even a single variable linear regression with downward shortwave radiation for the sensible heat flux!

Analysis using distribution statistics resulted in the land surface models having differing performance compared to the set of benchmarks. This result is inconsistent with the standard statistical measures and suggests that the models have been optimised for statistics such as mean bias error, standard deviation and correlation coefficient.

The conclusions from this study challenge our traditional view of the surface energy balance. In addition, the results suggest that improvements can be made to these models without the introduction of complexity, but by making better use of the currently available information content in the atmospheric forcing.

P.3 Towards efficient and systematic model benchmarking in CMIP6

Peter J. Gleckler[1;†] and Veronika Eyring[2],

[1]Lawrence Livermore National Laboratory, Livermore, California, USA
[2]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Oberpfaffenhofen, Germany
[†]Author to whom correspondence should be addressed; e-mail: gleckler1@llnl.gov

A more routine benchmarking and evaluation of models is envisaged to be a central part of CMIP6. One purpose of the DECK and CMIP historical simulations is to provide a basis for documenting model simulation characteristics. A few analysis packages currently under development will be routinely executed whenever new model experiments are contributed to the CMIP archive. The foundation that will enable this to be efficient and systematic is the community-based experimental protocols and conventions of CMIP, including their extension to obs4MIPs, which serves observations in parallel to the CMIP output on the ESFG. Examples of available tools that target routine evaluation in CMIP will be highlighted in this talk including the PCMDI Metrics Package (PMP) and the Earth System Model Evaluation Tool (ESMValTool). The PMP is built on DOE supported tools and emphases the implementation of a diverse suite of summary statistics to objectively gauge the level of agreement between model simulations and observations. ESMValTool includes a variety of diagnostics and metrics, including reproduction of the analysis in the IPCC AR5 model evaluation chapter. Both capabilities are open source, have a wide range of functionality, and are being developed as community tools with the involvement of multiple institutions. Collectively, the PMP, ESMValTool and ILAMB packages offer valuable capabilities that will be crucial for the systematic benchmarking of the wide variety of models and model versions contributed to CMIP6. This evaluation activity can, compared with early phases of CMIP, more quickly and openly relay to analysts and modelling centers the strengths and weaknesses of the simulations including the extent to which long-standing model errors remain evident in newer models. This talk will highlight the opportunities and challenges these capabilities provide as well as possible pathways to advance the coordination between them. It will also explain how this community-based benchmarking can accelerate the pace at which climate models can be used to further scientific understanding of climate change.

P.4 Land surface Verification Toolkit (LVT): A formal benchmarking and evaluation framework for land surface models

Sujay V. Kumar[1;†] and Christa D. Peters-Lidard[1]

[1]NASA Goddard Space Flight Center, Greenbelt, Maryland, USA
[†]Author to whom correspondence should be addressed; e-mail: Sujay.V.Kumar@nasa.gov

Though there is a vast amount of literature on land surface model development, model simulation studies and multi-model intercomparison projects, the evaluation methods and metrics used in them tend to be specific for individual case studies and mostly deterministic. These studies have not typically converged on standard measures of model performance for evaluating different LSMs. In this presentation, we describe the development and capabilities of a formal system for land surface model evaluation and benchmarking called the Land surface Verification Toolkit (LVT). LVT is designed to provide an automated, consolidated environment for model evaluation and includes approaches for conducting both traditional deterministic and probabilistic verification. LVT employs observational datasets in their native formats, enabling the continued use of the system without requiring additional implementation or data re-processing. Currently a large suite of in-situ, remotely sensed and other model and reanalysis datasets are implemented in LVT. Aside from the accuracy-based measures, LVT also includes metrics to aid model identification, such as entropy, complexity and information content. These measures can be used to characterize the tradeoffs in model performance relative to the information content of the model outputs. In addition to model verification, LVT also provides an environment for model benchmarking, where benchmark values for each metric is established a priori. The development of such benchmarks is facilitated in LVT, using regression and machine learning techniques. Finally, LVT also includes uncertainty and ensemble diagnostics based on Bayesian approaches that enable the quantification of predictive uncertainty in land surface model outputs. These capabilities provide novel ways to characterize LSM performance, enable rapid model evaluation efforts, and are expected to help in the definition and refinement of a formal benchmarking and evaluation process for the land surface modeling community. A suite of examples of using LVT for the evaluation of land surface model and data assimilation integrations will be presented.

P.5 Development of the International Land Model Benchmarking (ILAMB) System version 1 and its application to CMIP5 Earth system models and the Community Land Model

James T. Randerson[1;†], Mingquan Mu[1], Gretchen Keppel-Aleks[2], Charles D. Koven[3], William J. Riley[3]
Dave M. Lawrence[4], and Forrest M. Hoffman[5]

[1]University of California Irvine, Irvine, California, USA
[2]University of Michigan, Ann Arbor, Michigan, USA
[3]Lawrence Berkeley National Laboratory, Berkeley, California, USA
[4]National Center for Atmospheric Research, Boulder, Colorado, USA
[5]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[†]Author to whom correspondence should be addressed; e-mail: jranders@uci.edu

New approaches for evaluating earth system models (ESMs) are needed to improve the quality of simulations of future global environmental change and to speed model development. Here we describe the development of the International Land Model Benchmarking (ILAMB) software system. Version 1 of the ILAMB system (ILAMBv1) provides a framework for comparing model simulations with observations for 25 land surface variables. This set encompasses 9 carbon cycle and ecosystem, 5 hydrological and turbulent energy, 6 surface radiation, and 5 driver variables. For many variables, more than one dataset has been integrated within the system, enabling comparisons with data products that have different regional coverage or methodology. For each data set, scoring metrics and graphical output allow the user to explore model behavior within different regions and across seasonal, interannual, and (when appropriate) decadal time scales. Another set of variable to variable comparisons enables investigation of functional relationships, and limits the influence of climate system biases. We use the ILAMBv1 to evaluate ESMs participating in Phase 5 of the Coupled Model Intercomparison Project (CMIP5) and several versions of the Community Land Model. Analysis of historical simulations (1850-2005) from CMIP5 that had prognostic atmospheric carbon dioxide revealed several biases in the multi-model mean that may help guide future model development.

F.2.2 Emergent Constraints and Evaluation Metrics I

P.6 Evaluation of vegetation cover and land-surface albedo

Victor Brovkin[1,†], Lena Boysen[1], Thomas Raddatz[1], Veronika Gayler[1], Alexander Loew[1], and Martin Claussen[1]

[1]Max Planck Institute for Meteorology, Hamburg, Germany
[†]Author to whom correspondence should be addressed; e-mail: victor.brovkin@mpimet.mpg.de

In recent generation Earth System Models (ESMs), land-surface grid cells are represented as tiles covered by different plant functional types (PFTs) such as trees or grasses. Here, we present an evaluation of the vegetation cover module of the MPI-ESM for present-day conditions. The vegetation continuous fields (VCF) product [Hansen et al., 2003] that is based on satellite observations in 2001 is used to evaluate the fractional distributions of woody vegetation cover and bare ground. The model performance is quantified using two metrics: a square of the Pearson correlation coefficient, r^2, and the root-mean-square error, rmse. On a global scale, r^2 and rmse of modeled tree cover are equal to 0.61 and 0.19, respectively, which we consider as satisfactory values. The model simulates tree cover and bare ground with r^2 higher for the Northern Hemisphere (0.66) than for the Southern Hemisphere (0.48-0.50). We complement this analysis with an evaluation of the simulated land-surface albedo using the difference in net surface radiation. On global scale, the correlation between modeled and observed albedo is high during all seasons, while the main disagreement occurs in spring in the high northern latitudes. This discrepancy can be attributed to a high sensitivity of the land-surface albedo to the simulated snow cover and snow-masking effect of trees. In contrast, the tropics are characterized by very high correlation and relatively low rmse (5.4–6.5 W/m^2) during all seasons. The proposed approach could be applied for an evaluation of vegetation cover and land-surface albedo simulated by different ESMs.

P.7 Judging the dance contest – Metrics of land–atmosphere feedbacks

Paul A. Dirmeyer[1,†] and Liang Chen[1]

[1]Center for Ocean-Land-Atmosphere Studies (COLA), George Mason University, Manassas, Virginia, USA
[†]Author to whom correspondence should be addressed; e-mail: pdirmeye@gmu.edu

The Global Energy and Water Exchanges project (GEWEX), part of the World Climate Research Programme, has supported the investigation of processes involved in the local coupling between land and atmosphere and how they are simulated in models. From this effort, a compilation of coupling metrics has been produced that quantify both legs of the feedback from land to atmosphere: how biophysical land surface states affect surface fluxes, and what effect changes in surface fluxes have on the overlying atmosphere. A key consideration emerges from this approach – namely, that in climate models, both dance partners (land and atmosphere) must execute their steps correctly for the feedbacks to be realized. This requires there to be sufficient sensitivity in the links of the feedback chain, variability of the drivers of the feedbacks and memory of anomalies that excite feedbacks. Some metrics of land-atmosphere coupling are predicated on unobservable characteristics (e.g., the behavior of ensemble statistics in model simulations) but recent emphasis has turned towards metrics based on observable quantities and climate model variables, which provide a means for univariate and multivariate validation of coupled land-atmosphere behavior in models. Examples will be presented to prompt further discussion of potentials for benchmarking.

F.2.3 Ecological Sampling Networks

P.8 Role of flux networks in benchmarking land atmosphere models

Dennis Baldocchi[1,†]

[1]University of California Berkeley, Berkeley, California, USA
[†]Author to whom correspondence should be addressed; e-mail: baldocchi@berkeley.edu

Fluxnet is an international network of long term flux measurements of carbon dioxide, water vapor, heat and momentum fluxes. The network spans the globe in terms of climate and ecological spaces. Plus many locales have clusters of sites that address land use, land use change, disturbance and management. The network has been in operation since 1997 and many sites have more than a decade of data.

These flux data are proving to be useful to validate and parameterize light use efficiency models that are used by the satellite remote sensing community, to identify important processes that must be captures by land modules in climate models and as priors for the new generation of data model fusion methods. Site metadata are proving critical for providing initial conditions for models.

Lessons learned from the network and opportunities for the two communities to collaborate will be discussed.

F.2.4 MIP Benchmarking Needs

P.9 Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) Experimental Design and Organisation

Veronika Eyring[1;†], Sandrine Bony[2], Gerald A. Meehl[3], Cath Senior[4], Bjorn Stevens[5], Ronald J. Stouffer[6], and Karl E. Taylor[7]

Presented by David M. Lawrence[3]

[1]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Oberpfaffenhofen, Germany
[2]Laboratoire des Sciences du Climat et de l'Environnement, Gif sur Yvette Cedex, France and Universit Pierre et Marie Curie, Paris, France
[3]National Center for Atmospheric Research, Boulder, Colorado, USA
[4]UK Met Oce, Exeter, EX1 3PB, UK
[5]Max Planck Institute for Meteorology, Hamburg, Germany
[6]Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA
[7]Lawrence Livermore National Laboratory, Livermore, California, USA
[†]Author to whom correspondence should be addressed; e-mail: veronika.eyring@dlr.de

From Eyring et al., GMDD (2015): By coordinating the design and distribution of global climate model simulations of the past, current and future climate, the Coupled Model Intercomparison Project (CMIP) has become one of the foundational elements of climate science. However, the need to address an ever-expanding range of scientific questions arising from more and more research communities has made it necessary to revise the organization of CMIP. After a long and wide community consultation, a new and more federated structure has been put in place. It consists of three major elements: (1) a handful of common experiments, the DECK (Diagnostic, Evaluation and Characterization of Klima experiments) and the CMIP Historical Simulation (1850 – near-present) that will maintain continuity and help document basic characteristics of models across different phases of CMIP, (2) common standards, coordination, infrastructure and documentation that will facilitate the distribution of model outputs and the characterization of the model ensemble, and (3) an ensemble of CMIP-Endorsed Model Intercomparison Projects (MIPs) that will be specific to a particular phase of CMIP (now CMIP6) and that will build on the DECK and the CMIP Historical Simulation to address a large range of specific questions and ll the scientific gaps of the previous CMIP phases. The DECK and CMIP Historical Simulation, together with the use of CMIP data standards, will be the entry cards for models participating in CMIP. The participation in the CMIP6-Endorsed MIPs will be at the discretion of the modelling groups, and will depend on scientific interests and priorities. With the Grand Science Challenges of the World Climate Research Programme (WCRP) as its scientific backdrop, CMIP6 will address three broad questions: (i) How does the Earth system respond to forcing?, (ii) What are the origins and consequences of systematic model biases?, and (iii) How can we assess future climate changes given climate variability, predictability and uncertainties in scenarios? This CMIP6 overview presents the background and rationale for the new structure of CMIP, provides a detailed description of the DECK and the CMIP6 Historical Simulation, and includes a brief introduction to the 21 CMIP6-Endorsed MIPs.

Reference: Eyring, V., Bony, S., Meehl, G. A., Senior, C., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2015), Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organisation, *Geosci. Model Dev. Discuss.*, 8:10539-10583, doi:10.5194/gmdd-8-10539-2015.

P.10 Assessing feedbacks for the Coupled Climate–Carbon Cycle Modeling Intercomparison Project (C^4MIP)

Forrest M. Hoffman[1,†], James T. Randerson[2], Charles D. Koven[3], and the C^4MIP SSC and members

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[2]University of California Irvine, Irvine, California, USA
[3]Lawrence Berkeley National Laboratory, Berkeley, California, USA
[†]Author to whom correspondence should be addressed; e-mail: forrest@climatemodeling.org

The objective of the Coupled Climate–Carbon Cycle Modeling Intercomparison Project (C^4MIP) is to design, document, and analyze carbon cycle feedbacks and nutrient interactions in climate simulations for the sixth phase of the Coupled Model Intercomparison Project (CMIP6). These biogeochemical feedbacks are uncertain and potentially large, and they play a strong role in determining future atmospheric CO_2 levels in response to anthropogenic emissions and attempts to avoid dangerous climate change. Our recent paper (Jones et al., 2016) describes the simulations that will complement and extend the carbon cycle simulations included the CMIP6 core experiments known as the DECK. The key science motivations of these simulations are to 1) quantify and understand the carbon-concentration and carbon-climate feedback parameters, which capture the modeled response of land and ocean biogeochemistry components to changes in atmospheric CO_2 and the associated changes in climate, respectively; 2) evaluate models by comparing historical simulations with observation-based estimates of climatological states of carbon cycle variables, their variability and long-term trends; 3) assess the future projections of components of the global carbon budget for different scenarios. Model benchmarking efforts being undertaken for ILAMB are particularly important for the second of these motivations. In this presentation, we will briefly describe the experimental design of the CMIP6 historical and C^4MIP experiments and link these to model evaluation objectives that may be addressed by ILAMB benchmarking tools.

Reference: Jones, Chris D., Vivek Arora, Pierre Friedlingstein, Laurent Bopp, Victor Brovkin, John Dunne, Heather Graven, Forrest M. Hoffman, Tatiana Ilyina, Jasmin G. John, Martin Jung, Michio Kawamiya, Charles D. Koven, Julia Pongratz, Thomas Raddatz, James T. Randerson, and Sönke Zaehle (2016), The C^4MIP experimental protocol for CMIP6, *Geosci. Model Dev. Discuss.*, doi:10.5194/gmd-2016-36.

P.11 The Land Surface, Snow and Soil moisture Model Intercomparison Project (LS3MIP) and Global Soil Wetness Project Phase 3 (GSWP3)

Hyungjun Kim[1,†], Bart van den Hurk[2], Gerhard Krinner[3], Sonia I. Seneviratne[4], Chris Derksen[5], and Taikan Oki[1]

[1]University of Tokyo, Bunkyo-ku, Tokyo, Japan
[2]Royal Netherlands Meteorological Institute (KNMI), NL-3731 GA De Bilt, Netherlands
[3]Laboratoire de Glaciologie et Gophysique de l'Environnement (LGGE), Grenoble, France
[4]Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
[5]Environment Canada, Waterloo, Ontario, Canada
[†]Author to whom correspondence should be addressed; e-mail: hjkim@iis.u-tokyo.ac.jp

The solid and liquid water stored at the land surface has a large influence on the regional climate, its variability and its predictability, including effects on the energy and carbon cycles. Notably, snow and soil moisture affect surface radiation and flux partitioning properties, moisture storage and land surface memory. Recently, the Land Surface, Snow and Soil moisture Model Intercomparison Project (LS3MIP) was initiated as an intercommunity effort between Global Energy and Water Cycle Exchanges Project (GEWEX) and Climate and Cryosphere (CliC) to contribute to the 6th phase of Coupled Model Intercomparison Project (CMIP).

The experiment structure of the LS3MIP was designed to provide a comprehensive assessment of land surface, snow, and soil moisture feedbacks on climate variability and climate change, and to diagnose systematic biases in the land modules of current Atmospheric-Ocean General Circulation Models and Earth System Models with the following objectives:

- evaluate the current state of land processes including surface fluxes, snow cover and soil moisture representation in CMIP6 DECK runs;
- estimate multi-model long-term terrestrial energy/water/carbon cycles, using the surface modules of CMIP6 models under observation constrained historical (land reanalysis) and projected future (impact assessment) conditions considering land use/land cover changes;

- assess the role of snow and soil moisture feedbacks in the regional response to altered climate forcings, focusing on controls of climate extremes, water availability and high-latitude climate in historical and future scenario runs;
- assess the contribution of land surface processes to the current and future predictability of regional temperature/precipitation patterns. The outcomes of the LS3MIP will eventually contribute to the improvement of climate change projections by reducing the systematic biases and representing better feedback mechanisms in coupled models.

Further, the impacts of climate change on hydrological regimes and available freshwater resources including extreme events, such as floods and droughts, will be assessed based on multi-model ensemble estimates of long-term historical and projected future changes in energy, water, and carbon cycles over land surfaces. Those achievements will contribute to the next cycle of the Intergovernmental Panel on Climate Change.

P.12 Land-use and land-cover change model performance metrics for LUMIP

Dave M. Lawrence[1,†], George Hurtt[2], and LUMIP SSC and members

[1]National Center for Atmospheric Research, Boulder, Colorado, USA
[2]University of Maryland, College Park, Maryland, USA
[†]Author to whom correspondence should be addressed; e-mail: dlawren@ucar.edu

The main science questions that will be addressed by LUMIP (Lawrence et al. 2016), in the context of CMIP6 are:

- What are the global and regional effects of land-use and land-cover change on climate and biogeochemical cycling (past-future)?
- What are the impacts of land management on surface fluxes of carbon, water, and energy and are there regional land management strategies with promise to help mitigate and/or adapt to climate change?

In addressing these questions, LUMIP will also address a range of more detailed science questions to get at process level attribution, uncertainty, data requirements, and other related issues in more depth and sophistication than possible in a multi-model context to date. There will be particular focus on (1) the separation and quantification of the effects on climate from land-use change relative to fossil fuel emissions, (2) separation of biogeochemical from biogeophysical effects of land-use, (3) the unique impacts of land-cover change versus land management change, (4) modulation of land-use impact on climate by land-atmosphere coupling strength, and (5) the extent that direct effects of enhanced CO_2 concentrations on plant photosynthesis (changes in water-use efficiency and/or plant growth) are modulated by past and future land use.

One of the activities of LUMIP is to develop a set of metrics and diagnostic protocols quantify model performance, and related sensitivities, with respect to land use. De Noblet-Ducoudr et al (2012) identified the lack of consistent evaluation of a land model's ability to represent a response to a perturbation such as land-use change as a key contributor to the large spread in simulated land-cover change responses seen in the LUCID project. As part of this activity, benchmarking data products will be identified to help constrain models. Several recent studies have utilized various methodologies, including paired tower sites and reconstructed change maps from satellites, to infer observationally-based historical change in land surface variables impacted by LULCC or divergences in surface response between different land-cover types (Boisier et al. 2013, 2014; Lee et al. 2011; Lejeune et al. 2016; Li et al. 2015; Teuling et al. 2010; Williams et al. 2012).

P.13 Multi-scale Synthesis & Terrestrial Model Intercomparison Project: From cohort to insight

Christopher R. Schwalm[1,†], Deborah N. Huntzinger[2], Anna M. Michalak[3], Yuanyuan Fang[3], Kevin M. Schaefer[4], Andrew R. Jacobson[5], Joshua B. Fisher[6], Robert B. Cook[7], and Yaxing Wei[7]

[1]Woods Hole Research Center, Falmouth, Massachusetts, USA
[2]Northern Arizona University, Flagstaff, Arizona, USA
[3]Carnegie Institution for Global Ecology, Stanford University, Stanford, California, USA
[4]National Snow and Ice Data Center, University of Colorado, Boulder, Colorado, USA
[5]National Oceanic and Atmospheric Administration (NOAA), Boulder, Colorado, USA
[6]NASA Jet Propulsion Laboratory, Pasadena, California, USA
[7]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[†]Author to whom correspondence should be addressed; e-mail: schwalm.christopher@gmail.com

Earth system models (ESMs) are indispensable for extrapolating local observations and process level understanding of land–atmosphere exchange in both time and space. ESMs have and will continue to serve as predictive tools to understand carbon–climate interactions and global change. The North American Carbon Program (NACP) Multi-scale synthesis and Terrestrial Model Intercomparison Project (MsTMIP) is a formal intercomparison and evaluation effort focused on the land component of ESMs, i.e., land surface models (LSMs). MsTMIPs overarching goals are (1) to improve the diagnosis, attribution and prediction of carbon exchange at regional to global scales; and (2) to diagnose causes and consequences of inter-model variability. A key design tenet of MsMTIP is its standardized protocol. Forcing data, steady-state spin-up, and boundary conditions are uniform across all participating models. Modeler discretion is constrained to allow a mapping of skill to structure. The MsTMIP effort formally consists of two phases: Phase I (now complete) assembled a cohort of ca. 20 modeling teams and has released results from 15 LSMs. These results cover the 1901–2010 time period (half-degree resolution, monthly time step) and are based on a semi-factorial set of simulations; time-varying climate, land cover/land use change, carbon dioxide, and nitrogen deposition are sequentially enabled. Phase II (currently underway) extends Phase I models runs to 2100 using downscaled CMIP5 model output (5 ESMs and 2 RCPs [4.5 and 8.5]) as forcing data. With these predictive/forecast simulations MsTMIP can now serve as a platform to evaluate of how model structural differences, key controls of carbon metabolism, and plausible climate futures alter predictions of future carbon dynamics.

P.14 Processes Linked to Uncertainties Modelling Ecosystems (PLUME-MIP)

Anders Ahlström[1,2,†], Benjamin Smith[2], Almuth Arneth[3], Yiqi Luo[4], Jianyang Xia[5], and Michael Mishurow[2]

[1]Stanford University, Stanford, California, USA
[2]Lund University, Lund, Sweden
[3]Karlsruhe Institute of Technology, Karlsruhe, Germany
[4]University of Oklahoma, Norman, Oklahoma, USA
[5]East China Normal University, Shanghai, China
[†]Author to whom correspondence should be addressed; e-mail: anders.ahlstrom@nateko.lu.se

PLUME addresses DGVM/LSM responses to environmental drivers under current and future projections and attempts to advance the state-of-the-art in attributing modelled carbon cycle responses to underlying mechanisms, as represented in the models.

The project is divided into two main tiers.

Tier 1 involves standard transient simulations using CMIP5 recent past and future climate as forcing. The outcomes will be used to evaluate the different responses of the terrestrial C cycle to climate projections and CO_2 pathways.

Tier 2 adopts the transient Traceability Framework (TF) to identify underlying causes of differences in the responses of different models to current and future climate forcing. The framework is designed to facilitate model inter-comparisons by tracking a few traceable components across models.

Both Tiers contribute to the aim of isolating the processes responsible for differences between models and their future projections, using a transparent and systematic methodology. The TF represent the flows of carbon in the models and allows for a set of novel experiments. These experiments are based on replacing components and fluxes in the models with common or observed forcing, e.g. forcing the transient TF emulator of the models with NPP or vegetation inputs to soil, to isolate and estimate the relative contribution of processes to carbon storage uncertainties.

Within the project we offer assistance to help implementation of the framework, data harmonization and storage on a common database.

F.2.5 Emergent Constraints and Evaluation Metrics II

P.15 New benchmarks for northern high latitudes

Charles D. Koven[1,†]

[1]Lawrence Berkeley National Laboratory, Berkeley, California, USA
[†]Author to whom correspondence should be addressed; e-mail: cdkoven@lbl.gov

The northern high latitudes, with large stocks of carbon, high anticipated rates of climate change, and importance of abrupt change in ecosystem state with warming due to the importance of freeze/thaw processes, are a crucial

component of the Earth system that global models must represent. The CMIP5 ESMs fared particularly poorly in this region, due to the historical lack of attention paid to high latitude terrestrial processes in global models. I will discuss a variety of benchmarks focused around three areas: soil temperature dynamics and permafrost state, soil carbon stocks and turnover times, and hydrology dynamics. Each of these allow constraints on high latitude dynamics and may help to reduce uncertainty in model projections of the high latitude region.

P.16 Permafrost Benchmark System (PBS)

Kevin M. Schaefer[1,†], Elchin Jafarov[2], Mark Piper[2], Christopher R. Schwalm[3], Kang Wang[2], and Lynn Yarmey[1]

[1]National Snow and Ice Data Center, University of Colorado, Boulder, Colorado, USA
[2]Institute of Arctic and Alpine Research, University of Colorado, Boulder, Colorado, USA
[3]Woods Hole Research Center, Falmouth, Massachusetts, USA
[†]Author to whom correspondence should be addressed; e-mail: kevin.schaefer@nsidc.org

The Permafrost Benchmark System (PBS) will evaluate simulated permafrost dynamics against observed permafrost conditions. The project goals are 1) to develop a set of generic benchmarking tools capable of calculating performance statistics in multiple benchmarking efforts, and 2) develop benchmark datasets of permafrost dynamics based on available observations and 3) apply the PBS by evaluating models that ran the CMIP5 and MsTMIP simulations. We will collaborate with ILAMB to optimize resources and maximize benifit to the modeling community. We will use the core ILAMB infrastructure for benchmark management and model scoring. We will integrate the benchmarks we develop into ILAMB and integrate ILAMB into the Community Surface Dynamics Modeling System (CSDMS) to provide and an online user interface. This will provide an easily accessible, online tool to quickly evaluate model performance and guide model development without having to invest large resources into data preparation and organization. The chosen benchmark datasets include measurements of active layer thickness, permafrost temperature, snow conditions, and frozen soil biogeochemistry. We have formed an informal group of people already developing permafrost benchmarks to coordinate our activities and minimize duplication. The ideal performance target is to match the observations within uncertainty, so the PBS benchmark datasets and evaluation metrics will account for observation uncertainty. The combined IL AMB and PBS infrastructure fills a basic need of modeling teams to evaluate how well their models simulate permafrost dynamics, without a heavy investment in time and resources to organize the observations.

F.2.6 Strategies for Improving Models Through Evaluation

P.17 Theory-guided model evaluation and improvement

Yiqi Luo[1,†] and many others

[1]University of Oklahoma, Norman, Oklahoma, USA
[†]Author to whom correspondence should be addressed; e-mail: yluo@ou.edu

Global land models have become increasingly complicated over the past decades as more and more processes are incorporated into the models to simulate C cycle responses to global change. As a consequence, it becomes very difficult to understand or evaluate complex behaviors of these models. Differences in predictions among models cannot be easily diagnosed and attributed to their sources. In the past few years, we have developed a new theoretical framework to quantify terrestrial carbon storage dynamics. Our theoretical analysis indicates that the ultimate force driving C storage change in an ecosystem is the equilibrium C storage capacity, which is jointly determined by ecosystem C input (e.g., net primary production, NPP) and residence time. Since both C input and residence time vary with time, the equilibrium C storage capacity is time-dependent and acts as a moving target that actual C storage chases. The rate of change in C storage is proportional to the C storage potential, the difference between the current and equilibrium C storage.

The theoretical framework offers a suite of new techniques for evaluating and improving global land carbon cycle models. Those techniques include high-fidelity emulator, three- dimensional (3D) parameter space, traceability analysis, and semi-analytic spin-up (SASU).

A high fidelity emulator is a matrix representation of soil carbon processes. The matrix equation consists of carbon balance equations, each of which carbon input into and output from each of the individual carbon pools. We have developed emulators of CLM3.5, CLM4.5, CABLE, LPJ-GUESS, and regional TECO, which can exactly replicate

simulations of C pools and fluxes with their original models when driven by a limited set of inputs from the full model (GPP, soil temperature, and soil moisture).

The 3D parameter space can place outputs of any carbon cycle models with a common metric to measure differences among models in terms of NPP, carbon residence time, and carbon storage potential.

The traceability analysis is to decompose a complex land model into traceable components based on mutually independent properties of modeled biogeochemical processes. By doing so, we can attribute model-model differences to sources in model structure, parameter, and forcing fields. The traceability analysis also can be used to evaluate effectiveness of newly incorporated modules into existing models, such as adding the N module on simulated C dynamics.

The semi-analytical spin-up (SASU) is the analytic solution to a set of equations that describe carbon transfers within ecosystems over time.

F.2.7 Emergent Constraints and Evaluation Metrics III

P.18 Evaluating the simulations of global nutrient cycles: Available observations and challenges

Ying-Ping Wang[1,†], Benjamin Houlton[2], and Edith Bai[3]

[1]Commonwealth Scientific and Industrial Research Organisation (CSIRO) Oceans and Atmosphere, Aspendale, Victoria 3195, Australia
[2]University of California Davis, Davis, California, USA
[3]Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang, China
[†]Author to whom correspondence should be addressed; e-mail: yingping.wang@csiro.au

Experimental evidence suggests that productivity of most land ecosystems is limited by supplies of major nutrients, particularly nitrogen at high latitudes and phosphorus at low latitudes. However, representation of nutrient limitation in different global land models has rarely been assessed systematically.

Here, I will discuss three types of data for evaluating the performance of global nutrient cycles: spatially explicit data of soil nitrogen and phosphorus pools; nitrogen isotope composition; variations of C:N and N:P ratios of leaf, wood and root tissues by plant functional types or latitude; and field long-term (>10 years) fertilizing experiments or 15N tracer experiments. Examples from the published studies will be presented to show how each type of observations are used to assess global nutrient cycle simulations. Collectively, the combined benchmarking approaches substantially aid in model based projections of global carbon- nutrient interactions.

Nevertheless, three major issue challenges remain. First, estimates of nitrogen fixation from the unmanaged land vary from 58 to over 200 Tg N/year, and the response of the observed of nitrogen fixation to CO_2 can also be highly uncertain. Yet there is currently no globally integrated approach to reduce this uncertainty.

Second, estimates of phosphorus input to land ecosystems through rock weathering and tectonic uplift vary by a factor of two. A recent study also found the phosphorus deposition input is significantly larger than previous estimate. These large uncertainties make the simulations of phosphorus cycles at global scale highly uncertain.

Third, most global nutrient models do not represent nutrient losses from particulate matter (both organic and inorganic). These models need to be coupled to hydraulic models to simulate the nutrient exports, in both organic and inorganic forms, from land to river, which have been measured over all major rivers in the world, and can be used to evaluate global nutrient cycles in the future.

P.19 Empirically derived sensitivity of vegetation to climate as a possible functional constraint for process based land models

Gregory R. Quetin[1,†] and Abigail L. S. Swann[1]

[1]University of Washington, Seattle, Washington, USA
[†]Author to whom correspondence should be addressed; e-mail: gquetin@uw.edu

Vegetated land ecosystems are shaped by climate across the globe to best take advantage of the conditions and resources available. Acclimation to different climatological states changes how each ecosystem functions, with the supply of different resources determining constraints on growth. Here we derive an empirical global map of the

sensitivity of vegetation to climate using the response of satellite-based greenness to interannual variations in surface air temperature and precipitation. We infer constraints on ecosystem function by analyzing how the sensitivity of vegetation to climate varies across climate space. We find four broad climate regions of ecosystem function. There is a cold region below 15°C, which is generally greener during warmer and drier years. There is a transition region between cold climate regions and hotter regions where the sign of vegetation sensitivity changes along a line of 0.017°C/mm/yr, indicative of constraints on productivity driven by a balance between water supply and temperature-dependent atmospheric water demand. A hot dry region above 15°C and below ~1000 mm/year rainfall is browner in warm years and greener in wetter years. Finally, a region beyond 1500 mm/year rainfall greens during warmer years even at the hottest vegetated places on Earth. In this region we propose that increased stress from temperature-dependent atmospheric water demand is offset by increased insolation that increases photosynthesis. These broad empirical patterns of ecosystem function across climate have the potential to provide functional constraints for Earth system models, helping improve our ability to model and predict global vegetation under a changing climate.

P.20 Some suggestions on emergent constraints and metrics on model evaluations over land

Xubin Zeng[1,†], William Lytle[1], Patrick Broxton[1], Nick Dawson[1], and Aihui Wang

[1]University of Arizona, Tucson, Arizona, USA
[2]Institute of Atmospheric Physics, Chinese Academy of Science, Beijing, China
†Author to whom correspondence should be addressed; e-mail: xubin@email.arizona.edu

(1) We have developed global hourly 0.5 degree land surface 2 m temperature (T_{2m}) datasets based on four reanalysis products and the CRUTS3.10 *in situ* dataset for 1948–2009. Our three-step adjustments ensure that our final products have exactly the same monthly-mean maximum (T_x) and minimum (T_n) temperature as the CRU data. One of the uncertainties in our final products can be quantified by their differences (Wang and Zeng 2013).

Based on these results, we make two suggestions for model land surface T_{2m} evaluation metrics:

- To evaluate model monthly mean temperature, which is averaged over all time steps, using the true monthly mean based on hourly values from our datasets, rather than using $T_m = (T_x + T_n)/2$
- To save monthly averaged diurnal cycle from models and compare its range with that based on our datasets, rather than using $DTR = T_x - T_n$.

(2) We have used measurements for several years at five flux tower sites in the U.S. (with a total of 315,576 hours of data) along with *in situ* snow measurements for the coupled evaluation of both below- and above-ground processes from three global reanalysis products and six global land data assimilation products. While errors in T_{2m} are highly correlated with errors in skin temperature for all sites, the correlations between skin and soil temperature errors are weaker, particularly over the sites with seasonal snow (Lytle and Zeng 2016). Therefore, one emergent constraint in model evaluation is the coupled evaluation of daily air, skin, and soil temperatures.

(3) It is well known that snow depth or water equivalent (SWE) varies substantially horizontally and with elevations, but we found that four methods for the spatial interpolation of peak of winter SWE and snow depth based on distance and elevation can result in large errors based on (SNOTEL and COOP) *in situ* data. These errors are reduced substantially by our new method; i.e., the spatial interpolation of these quantities normalized by accumulated snowfall. Our method results in significant improvement in SWE estimates over interpolation techniques that do not consider snowfall, regardless of the number of stations used for the interpolation (Broxton et al. 2016). Therefore, one emergent constraint in model evaluation is the evaluation of daily SWE over the accumulated snowfall.

P.21 Decomposition of CO_2 fertilization effect into contributions by land ecosystem processes: Comparison among CMIP5 Earth system models

Kaoru Tachiiri[1,†], Tomohiro Hajima[1], and Michio Kawamiya[1]

[1]Japan Agency for Marine-Earth Science and Technology, Kanagawa Prefecture, Japan
†Author to whom correspondence should be addressed; e-mail: tachiiri@jamstec.go.jp

Increase in atmospheric CO_2 concentration stimulates plant growth, and promotes carbon uptake by land ecosystems. This process, often called CO_2 fertilization, causes a negative feedback between atmospheric CO_2 concentration and terrestrial carbon uptake. The feed back is considered to have a strong impact on the climate–carbon cycle

system, but that has large inter-model variation in exiting Earth system models (ESMs). In this study, we examined in detail the sensitivity of change in land carbon storage to that in atmospheric CO_2 concentration (ΔCO_2) for the CMIP5 participant ESMs by breaking that down into the ratios of ΔCO_2, changes in gross primary production, leaf area index, net primary production, vegetation carbon, soil carbon, heterotrophic respiration, and land carbon storage. The results showed that increase in atmospheric CO_2 concentration stimulates plant production, litter fall, and heterotrophic respiration with different sensitivities to ΔCO_2 among the models, and major part in sensitivity of land carbon storage to ΔCO_2 could be explained by the sensitivity of plant productivity. The result suggests that to constrain the CO_2 fertilization effect we need to better understand plant primary production, and to do so more observations and experiments are needed. In case the number of ESMs incorporating the nitrogen cycle increases, we may need a new framework to evaluate the carbon and nitrogen cycles with integrated manner to analyze the CO_2 fertilization effect.

F.2.8 Uncertainty Quantification (UQ) Methods

P.22 An uncertainty quantification framework designed for land models

Maoyi Huang[1,†], Zhangshuan Hou[1], Jaideep Ray[2], Laura Swiler[3], and L. Ruby Leung[1]

[1]Pacific Northwest National Laboratory, Richland, Washington, USA
[2]Sandia National Laboratories, Livermore, California, USA
[3]Sandia National Laboratories, Albuquerque, New Mexico, USA
[†]Author to whom correspondence should be addressed; e-mail: maoyi.huang@pnnl.gov

Representing terrestrial processes and their exchanges with the atmosphere, land surface models are important components of Earth system models used to predict climate variations and change. Most land surface models include numerous sub-models, each representing key processes with mathematical equations and model parameters. Optimizing the parameter values may improve model skill in capturing the observed behaviors. In this presentation, we will discuss recent progress in quantifying uncertainty associated with hydrologic parameters in the Community Land Model (CLM) and calibrating those parameters using an uncertainty quantification (UQ) framework that features global sensitivity analysis, parameter screening, classifying the complex system into a few relatively homogeneous regions, and Bayesian inversion using Markov Chain Monte Carlo techniques. The UQ framework has been applied it to flux towers and watersheds under different climate and site conditions in the contiguous United States. Through these studies, they demonstrated that the CLM-simulated latent heat and sensible heat fluxes, and runoff generation are highly sensitive to hydrologic parameters, which could be better constrained using in-situ and remotely-sensed measurements such as the benchmarking datasets available in the International Land Model Benchmarking framework (ILAMB) (e.g., data from AmeriFlux network, streamflow gages, data products from the Moderate Resolution Imaging Spectroradiometer), when integrated with the UQ framework developed by the team. Although only being integrated with CLM, the framework is general and therefore is portable to other land models.

P.23 Use of emulators in uncertainty quantification

George Shu Heng Pau[1,†], Chaopeng Shen[2], and William J. Riley[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, California, USA
[2]Pennsylvania State University, State College, Pennsylvania, USA
[†]Author to whom correspondence should be addressed; e-mail: gpau@lbl.gov

Direct application of robust uncertainty quantification techniques, such as Monte Carlo methods, to high-resolution land models is typically infeasible even with existing high-end computing ecosystems. To reduce the computational burden of applying these techniques, we develop certified reduced order models, or emulators, to efficiently approximate solutions to high-resolution land models at a significant reduced cost. For a watershed-scale land model, we demonstrated that the proper orthogonal decomposition mapping method led to an emulator that had the desired spatial and temporal accuracies. The emulator then allows us to quantify uncertainties at scales relevant to decision support.

P.24 Uncertainty quantification in the ACME land model

Daniel M. Ricciuto[1,†], Khachik Sargsyan[2], and Peter E. Thornton[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[2]Sandia National Laboratories, Livermore, California, USA
[†]Author to whom correspondence should be addressed; e-mail: ricciutodm@ornl.gov

For computationally expensive climate models, Monte-Carlo approaches of exploring the in put parameter space are often prohibitive due to slow convergence with respect to ensemble size. To alleviate this, we build inexpensive surrogates using uncertainty quantification (UQ) methods employing Polynomial Chaos (PC) expansions that approximate the input-output relationships using as few model evaluations as possible. However, when many uncertain input parameters are present, such UQ studies suffer from the curse of dimensionality. In particular, for 50–100 input parameters non-adaptive PC representations have infeasible numbers of basis terms. To this end, we develop and employ Weighted Iterative Bayesian Compressive Sensing to learn the most important input parameter relationships for efficient, sparse PC surrogate construction with posterior uncertainty quantified due to insufficient data. Besides drastic dimensionality reduction, such uncertain surrogate can efficiently replace the model in computationally intensive studies such as forward uncertainty propagation and variance-based sensitivity analysis, as well as design optimization and parameter estimation using observational data.

We apply the surrogate construction and variance-based uncertainty decomposition to Accelerated Climate Model for Energy (ACME) Land Model for several output quantities of interest at model grid cells representing the locations of 100 FLUXNET sites, covering multiple plant functional types and a broad array of climates, varying 65 input parameters over ranges of possible values defined by literature and expert opinion. We find general consistency of the top 10–15 most sensitive parameters across sites and across quantities of interest, with some variation in the relative ranking of these parameters. We find especially strong sensitivity to parameters related to photosynthesis, nitrogen cycling, and allocation. Finally, we assess the quality of the surrogate model and the potential applications of UQ methods for model calibration and benchmarking.

P.25 PEcAn: A community tool to enable synthesis, evaluation & forecasting

Shawn P. Serbin[1,†], Michael C. Dietze[2], and the PEcAn Project team

[1]Brookhaven National Laboratory, Upton, New York, USA
[2]Boston University, Boston, Massachusetts, USA
[†]Author to whom correspondence should be addressed; e-mail: sserbin@bnl.gov

Models are our primary tool for synthesizing our understanding of ecosystems and projecting the impact of global change on ecosystem services associated with carbon, energy and water fluxes and storage. Recently the use of models as a scaffold for data-driven synthesis has expanded and there is increasing interest in formal model–data experimentation (ModEx) frameworks to quantify uncertainties, evaluate models, enable the integration of observations, and guide model developments as well as focus data collection on parameters that drive the greatest uncertainty. However, models remain inaccessible to most ecologists, in large part due to the informatics challenges of managing the flows of information in and out of such models, as well as access to the tools necessary to properly synthesize model results and quantify the uncertainties in projections. Managing the communication between models and data involves three distinct challenges: dealing with the volume of big data; processing unstructured and uncurated long tail data; and the need to capture and propagate uncertainties in model–data comparisons and formal data–model assimilation. Finally, model development has long been an academic cottage industry, with different models lacking compatible formats for inputs, outputs, and settings. This has lead to massive redundancies and minimal reproducibility. As a result, the pace of model improvement has been glacial. PEcAn (http://pecanproject.org/), a tool box for model–data ecoinformatics, tackles many of these challenges. Users interact with models through an intuitive Google-Map based interface, a simple application program interface (API) and standardized file formats. Standardization allows the development of common, reusable tools for processing inputs, visualizing outputs, and automating analyses. PEcAn includes state-of-the-art Hierarchical Bayes tools for model parameterization, data assimilation, uncertainty quantification (UQ) and variance decomposition (VD), as well as the ability to leverage tools for processing uncurated data. In addition to these tools, PEcAn leverages a PostGIS database network to track all inputs, outputs, and model runs, greatly increasing reproducibility and reliability. Within the PEcAn network, the database syncs all results and facilitates file sharing to allow models to talk to each other and enables the community to effectively analyze many models distributed across a global network, thereby increasing the ability to

conduct mulit-model, multi-institutional model comparisons and synthesis activities. In this talk, we will review the capabilities within PEcAn for formal UQ/VD to guide modeling activities but also discuss the many other features and provide an example of the capability for data assimilation and model–data experimentation.

F.3 List of On-site Participants

Gab Abramowitz
Anders Ahlström
Igor Aleinov
Concepcion Arroyo
Dennis Baldocchi
Martin Best
Benjamin Bond-Lamberty
Victor Brovkin
Nuno Carvalhais
Nathan Collier
Stuart J. Davies
Martin De Kauwe
Alberto M. de la Torre
Scott Denning
Ankur Desai
Paul A. Dirmeyer
Richard Ellis
Rosie Fisher
Peter J. Gleckler
Hirofumi Hashimoto
Forrest M. Hoffman
Thomas Holmes
Maoyi Huang
Gustaf Hugelius
Akihiko Ito
Atul K. Jain
Gretchen Keppel-Aleks
Nancy Y. Kiang
Hyungjun Kim
Randal D. Koster
Charles D. Koven
Sujay Kumar
David M. Lawrence
Hongyi Li
Chris Lu
Yiqi Luo
Jiafu Mao
Hank A. Margolis
Nathan McDowell
Umakant Mishra
Paul Moorcroft
Mingquan Mu
George Pau
Gregory Quetin
James T. Randerson
Dan Ricciuto
William J. Riley
Kevin Schaefer
Christopher R. Schwalm
Shawn Serbin
Elena Shevliakova
James Simkins
Andrew Slater
Nicholas G. Smith
Eric J. Stofferahn
Kaoru Tachiiri
Jinyun Tang
Ying Ping Wang
Mathew Williams
Jianyang Xia
Chonggang Xu
Xubin Zeng
Qian Zhang
Sha Zhou
Qing Zhu

Appendix G. References

Abramowitz, G. (2012), Towards a public, standardized, diagnostic benchmarking system for land surface models, *Geosci. Model Dev., 5*(3), 819–827, doi:10.5194/gmd-5-819-2012.

Adachi, M., A. Ito, A. Ishida, W. R. Kadir, P. Ladpala, and Y. Yamagata (2011), Carbon budget of tropical forests in Southeast Asia and the effects of deforestation: An approach using a process-based model and field measurements, *Biogeosci., 8*(9), 2635–2647, doi:10.5194/bg-8-2635-2011.

Adair, E. C., W. J. Parton, S. J. Del Grosso, W. L. Silver, M. E. Harmon, S. A. Hall, I. C. Burke, and S. C. Hart (2008), Simple three-pool model accurately describes patterns of long-term litter decomposition in diverse climates, *Glob. Change Biol., 14*(11), 2636–2660, doi:10.1111/j.1365-2486.2008.01674.x.

AghaKouchak, A., A. Farahmand, F. S. Melton, J. Teixeira, M. C. Anderson, B. D. Wardlow, and C. R. Hain (2015), Remote sensing of drought: Progress, challenges and opportunities, *Rev. Geophys., 53*(2), 452–480, doi:10.1002/2014RG000456.

Ahlström, A., J. Xia, A. Arneth, Y. Luo, and B. Smith (2015), Importance of vegetation dynamics for future terrestrial carbon cycling, *Environ. Res. Lett., 10*(5), 054,019, doi:10.1088/1748-9326/10/8/089501.

Anav, A., P. Friedlingstein, M. Kidston, L. Bopp, P. Ciais, P. Cox, C. Jones, M. Jung, R. Myneni, and Z. Zhu (2013), Evaluating the land and ocean components of the global carbon cycle in the CMIP5 Earth system models, *J. Clim., 26*(18), 6801–6843, doi:10.1175/JCLI-D-12-00417.1.

Asner, G. P., R. E. Martin, C. B. Anderson, and D. E. Knapp (2015), Quantifying forest canopy traits: Imaging spectroscopy versus field survey, *Remote Sens. Environ., 158*, 15–27, doi:10.1016/j.rse.2014.11.011.

Baccini, A., S. J. Goetz, W. S. Walker, N. T. Laporte, M. Sun, D. Sulla-Menashe, J. Hackler, P. S. A. Beck, R. Dubayah, M. A. Friedl, S. Samanta, and R. A. Houghton (2012), Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps, *Nature Clim. Change, 2*(3), 182–185, doi: 10.1038/nclimate1354.

Baldocchi, D. (2008), 'Breathing' of the terrestrial biosphere: Lessons learned from a global network of carbon dioxide flux measurement systems, *Aust. J. Bot., 56*(1), 1–26, doi:10.1071/BT07151.

Baldocchi, D. (2014), Measuring fluxes of trace gases and energy between ecosystems and the atmosphere — The state and future of the eddy covariance method, *Glob. Change Biol., 20*(12), 3600–3609, doi: 10.1111/gcb.12649.

Baldocchi, D., E. Falge, L. Gu, R. Olson, D. Hollinger, S. Running, P. Anthoni, C. Bernhofer, K. Davis, R. Evans, J. Fuentes, A. Goldstein, G. Katul, B. Law, X. Lee, Y. Malhi, T. Meyers, W. Munger, W. Oechel, K. T. Paw, K. Pilegaard, H. P. Schmid, R. Valentini, S. Verma, T. Vesala, K. Wilson, and S. Wofsy (2001), FLUXNET: A new tool to study the temporal and spatial variability of ecosystem–scale carbon dioxide, water vapor, and energy flux densities, *Bull. Am. Meteorol. Soc., 82*(11), 2415–2434, doi:10.1175/15200477(2001)082%3C2415:FANTTS%3E2.3.CO;2.

Baldocchi, D., M. Reichstein, D. Papale, L. Koteen, R. Vargas, D. Agarwal, and R. Cook (2012), The role of trace gas flux networks in the biogeosciences, *Eos Trans. AGU, 93*(23), 217–218, doi: 10.1029/2012EO230001.

Bao, J., H. Ren, Z. Hou, J. Ray, L. Swiler, and M. Huang (2016), Soil moisture estimation using tomographic ground penetrating radar in a MCMC-Bayesian framework, *Math. Geosci.*, in review.

Baret, F., O. Hagolle, B. Geiger, P. Bicheron, B. Miras, M. Huc, B. Berthelot, F. Niño, M. Weiss, O. Samain, J. L. Roujean, and M. Leroy (2007), LAI, fAPAR and fCover CYCLOPES global products derived from VEGETATION: Part 1: Principles of the algorithm, *Remote Sens. Environ., 110*(3), 275–286, doi: 10.1016/j.rse.2007.02.018.

Batjes, N. (2016), Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks, *Geoderma, 269*, 61–68, doi:10.1016/j.geoderma.2016.01.034.

Batjes, N. H., E. Ribeiro, A. van Oostrum, J. Leenaars, T. Hengl, and J. Mendes de Jesus (2017), WoSIS: providing standardised soil profile data for the world, *Earth Syst. Sci. Data, 9*, 1–14, doi:10.5194/essd-9-1-2017.

Beer, C., M. Reichstein, E. Tomelleri, P. Ciais, M. Jung, N. Carvalhais, C. Rödenbeck, M. A. Arain, D. Baldocchi, G. B. Bonan, A. Bondeau, A. Cescatti, G. Lasslop, A. Lindroth, M. Lomas, S. Luyssaert, H. Margolis, K. W. Oleson, O. Roupsard, E. Veenendaal, N. Viovy, C. Williams, F. I. Woodward, and D. Papale (2010), Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate, *Science, 329*(5993), 834–838, doi:10.1126/science.1184984.

Best, M. J., G. Abramowitz, H. R. Johnson, A. J. Pitman, G. Balsamo, A. Boone, M. Cuntz, B. Decharme, P. A. Dirmeyer, J. Dong, M. Ek, Z. Guo, V. Haverd, B. J. J. van den Hurk, G. S. Nearing, B. Pak, C. Peters-Lidard, J. A. Santanello Jr., L. Stevens, and N. Vuichard (2015), The plumbing of land surface models: Benchmarking model performance, *J. Hydrometeor., 16*(3), 1425–1442, doi:10.1175/JHM-D-14-0158.1.

Biederman, J. A., R. L. Scott, M. L. Goulden, R. Vargas, M. E. Litvak, T. E. Kolb, E. A. Yepez, W. C. Oechel, P. D. Blanken, T. W. Bell, J. Garatuza-Payan, G. E. Maurer, S. Dore, and S. P. Burns (2016), Terrestrial carbon balance in a drier world: The effects of water availability in southwestern North America, *Glob. Change Biol., 22*(5), 1867–1879, doi:10.1111/gcb.13222.

Bilionis, I., B. A. Drewniak, and E. M. Constantinescu (2015), Crop physiology calibration in the CLM, *Geosci. Model Dev., 8*(4), 1071–1083, doi:10.5194/gmd-8-1071-2015.

Bloom, A. A., J.-F. Exbrayat, I. R. van der Velde, L. Feng, and M. Williams (2016), The decadal state of the terrestrial carbon cycle: Global retrievals of terrestrial carbon allocation, pools, and residence times, *Proc. Nat. Acad. Sci., 113*(5), 1285–1290, doi:10.1073/pnas.1515160113.

Boisier, J. P., N. de Noblet-Ducoudré, A. J. Pitman, F. T. Cruz, C. Delire, B. J. J. M. van den Hurk, M. K. van der Molen, C. Müller, and A. Voldoire (2012), Attributing the impacts of land-cover changes in temperate regions on surface temperature and heat fluxes to specific causes: Results from the first LUCID set of simulations, *J. Geophys. Res. Atmos., 117*(D12), doi:10.1029/2011JD017106.

Boisier, J. P., N. de Noblet-Ducoudré, and P. Ciais (2013), Inferring past land use-induced changes in surface albedo from satellite observations: A useful tool to evaluate model simulations, *Biogeosci., 10*(3), 1501– 1516, doi:10.5194/bg-10-1501-2013.

Boisier, J. P., N. de Noblet-Ducoudré, and P. Ciais (2014), Historical land-use-induced evapotranspiration changes estimated from present-day observations and reconstructed land-cover maps, *Hydrol. Earth Syst. Sci., 18*(9), 3571–3590, doi:10.5194/hess-18-3571-2014.

Bonan, G. B. (2014), Connecting mathematical ecosystems, real-world ecosystems, and climate science, *New Phytol., 202*(3), 731–733, doi:10.1111/nph.12802.

Bond-Lamberty, B., A. P. Smith, and V. Bailey (2016a), Temperature and moisture effects on greenhouse gas emissions from deep active-layer boreal soils, *Biogeosci. Discuss.*, 2016, 1–36, doi:10.5194/bg-2016-234.

Bond-Lamberty, B., D. Epron, J. Harden, M. E. Harmon, F. M. Hoffman, J. Kumar, A. D. McGuire, and R. Vargas (2016b), Estimating heterotrophic respiration at large scales: Challenges, approaches, and next steps, *Ecosphere, 7*(6), doi:10.1002/ecs2.1380.

Bouskill, N. J., W. J. Riley, and J. Tang (2014), Meta-analysis of high-latitude nitrogen-addition and warming studies implies ecological mechanisms overlooked by land models, *Biogeosci., 11*(23), 6969–6983, doi:10.5194/bg-11-6969-2014.

Braakhekke, M. C., C. Beer, M. Schrumpf, A. Ekici, B. Ahrens, M. R. Hoosbeek, B. Kruijt, P. Kabat, and M. Reichstein (2014), The use of radiocarbon to constrain current and future soil organic matter turnover and transport in a temperate forest, *J. Geophys. Res. Biogeosci., 119*(3), 372–391, doi: 10.1002/2013JG002420.

Brown, R. D., and B. Brasnett (2010), Canadian Meteorological Centre (CMC) daily snow depth analysis data, Environment Canada, National Snow and Ice Data Center, Boulder, Colorado, USA, Digital media.

Brown, T. B., K. R. Hultine, H. Steltzer, E. G. Denny, M. W. Denslow, J. Granados, S. Henderson, D. Moore, S. Nagai, M. SanClements, A. Sánchez-Azofeifa, O. Sonnentag, D. Tazik, and A. D. Richardson (2016), Using phenocams to monitor our changing Earth: Toward a global phenocam network, *Front. Ecol. Environ., 14*(2), 84–93, doi:10.1002/fee.1222.

Brutel-Vuilmet, C., M. Ménégoz, and G. Krinner (2013), An analysis of present and future seasonal Northern Hemisphere land snow cover simulated by CMIP5 coupled climate models, *The Cryosphere, 7*(1), 67–80, doi:10.5194/tc-7-67-2013.

Burakowski, E. A., S. V. Ollinger, G. B. Bonan, C. P. Wake, J. E. Dibb, and D. Y. Hollinger (2016), Evaluating the climate effects of reforestation in New England using a Weather Research and Forecasting (WRF) model multiphysics ensemble, *J. Clim., 29*(14), 5141–5156, doi:10.1175/JCLI-D-15-0286.1.

Burke, I. C., C. M. Yonker, W. J. Parton, C. V. Cole, D. S. Schimel, and K. Flach (1989), Texture, climate, and cultivation effects on soil organic matter content in U.S. grassland soils, *Soil Sci. Soc. Am. J., 53*(3), 800–805, doi:10.2136/sssaj1989.03615995005300030029x.

Caldararu, S., D. W. Purves, and P. I. Palmer (2014), Phenology as a strategy for carbon optimality: A global model, *Biogeosci., 11*(3), 763–778, doi:10.5194/bg-11-763-2014.

Carvalhais, N., M. Forkel, M. Khomik, J. Bellarby, M. Jung, M. Migliavacca, M. Mu, S. Saatchi, M. Santoro, M. Thurner, U. Weber, B. Ahrens, C. Beer, A. Cescatti, J. T. Randerson, and M. Reichstein (2014), Global covariation of carbon turnover times with climate in terrestrial ecosystems, *Nature, 514*(7521), 213–217.

Cavaleri, M. A., S. C. Reed, W. K. Smith, and T. E. Wood (2015), Urgent need for warming experiments in tropical forests, *Glob. Change Biol., 21*(6), 2111–2121, doi:10.1111/gcb.12860.

Cervarich, M., S. Shu, A. K. Jain, A. Arneth, J. Canadell, P. Friedlingstein, R. Houghton, E. Kato, C. Koven, P. Patra, B. Poulter, S. Sitch, B. Stocker, N. Viovy, A. Wiltshire, and N. Zeng (2016), The terrestrial carbon budget of South and Southeast Asia, *Environ. Res. Lett.*, in revision.

Chen, J., Y. Luo, J. Xia, Z. Shi, L. Jiang, S. Niu, X. Zhou, and J. Cao (2016), Differential responses of ecosystem respiration components to experimental warming in a meadow grassland on the Tibetan Plateau, *Agr. Forest Meteorol., 220*, 21–29, doi:10.1016/j.agrformet.2016.01.010.

Chen, L., and P. A. Dirmeyer (2016), Adapting observationally based metrics of biogeophysical feedbacks from land cover/land use change to climate modeling, *Environ. Res. Lett., 11*(3), 034,002, doi: 10.1088/1748-9326/11/3/034002.

Chorin, A. J., and X. Tu (2009), Implicit sampling for particle filters, *Proc. Nat. Acad. Sci., 106*(41), 17,249– 17,254, doi:10.1073/pnas.0909196106.

Ciais, P., C. Sabine, G. Bala, L. Bopp, V. Brovkin, J. Canadell, A. Chhabra, R. DeFries, J. Galloway, M. Heimann, C. Jones, C. Le Quéré, R. B. Myneni, S. Piao, and P. Thornton (2013), 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*, edited by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, pp. 465–570, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Cohen, W. B., T. K. Maiersperger, Z. Yang, S. T. Gower, D. P. Turner, W. D. Ritts, M. Berterretche, and S. W. Running (2003), Comparisons of land cover and LAI estimates derived from ETM+ and MODIS for four sites in North America: A quality assessment of 2000/2001 provisional MODIS products, *Remote Sens. Environ., 88*(3), 233–255, doi:10.1016/j.rse.2003.06.006.

Collier, N., F. M. Hoffman, M. Mu, J. T. Randerson, W. J. Riley, C. D. Koven, G. Keppel-Aleks, and D. M. Lawrence (2016), International Land model Benchmarking (ILAMB) package v002.00, Software package, doi:10.18139/ILAMB.v002.00/1251621.

Covey, C., P. J. Gleckler, C. Doutriaux, D. N. Williams, A. Dai, J. Fasullo, K. Trenberth, and A. Berg (2016), Metrics for the diurnal cycle of precipitation: Toward routine benchmarks for climate models, *J. Clim., 29*(12), 4461–4471, doi:10.1175/JCLI-D-15-0664.1.

Cox, P. M., D. Pearson, B. B. Booth, P. Friedlingstein, C. Huntingford, C. D. Jones, and C. M. Luke (2013), Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability, *Nature, 494*(7437), 341–344, doi:10.1038/nature11882.

Csilléry, K., M. G. Blum, O. E. Gaggiotti, and O. François (2010), Approximate Bayesian Computation (ABC) in practice, *Trends Ecol. Evol., 25*(7), 410–417, doi:10.1016/j.tree.2010.04.001.

Csilléry, K., O. François, and M. G. B. Blum (2012), abc: An R package for Approximate Bayesian Computation (ABC), *Methods Ecol. Evol., 3*(3), 475–479, doi:10.1111/j.2041-210X.2011.00179.x.

Cuesta-Valero, F. J., A. García-García, H. Beltrami, and J. E. Smerdon (2016), First assessment of continental energy storage in CMIP5 simulations, *Geophys. Res. Lett., 43*(10), 5326–5335, doi: 10.1002/2016GL068496.

Davidson, E. A., and I. A. Janssens (2006), Temperature sensitivity of soil carbon decomposition and feedbacks to climate change, *Nature, 440*(7081), 165–173, doi:10.1038/nature04514.

Davidson, E. A., S. Samanta, S. S. Caramori, and K. Savage (2012), The Dual Arrhenius and Michaelis– Menten kinetics model for decomposition of soil organic matter at hourly to seasonal time scales, *Glob. Change Biol., 18*(1), 371–384, doi:10.1111/j.1365-2486.2011.02546.x.

De Kauwe, M. G., B. E. Medlyn, S. Zaehle, A. P. Walker, M. C. Dietze, T. Hickler, A. K. Jain, Y. Luo, W. J. Parton, I. C. Prentice, B. Smith, P. E. Thornton, S. Wang, Y.-P. Wang, D. Wårlind, E. Weng, K. Y. Crous, D. S. Ellsworth, P. J. Hanson, H.-S. Kim, J. M. Warren, R. Oren, and R. J. Norby (2013), Forest water use and water use efficiency at elevated CO_2: A model-data intercomparison at two contrasting temperate forest FACE sites, *Glob. Change Biol., 19*(6), 1759–1779, doi:10.1111/gcb.12164.

De Kauwe, M. G., B. E. Medlyn, S. Zaehle, A. P. Walker, M. C. Dietze, Y.-P. Wang, Y. Luo, A. K. Jain, B. El-Masri, T. Hickler, D. Wårlind, E. Weng, W. J. Parton, P. E. Thornton, S. Wang, I. C. Prentice, S. Asao, B. Smith, H. R. McCarthy, C. M. Iversen, P. J. Hanson, J. M. Warren, R. Oren, and R. J. Norby (2014), Where does the carbon go? A model–data intercomparison of vegetation carbon allocation and turnover processes at two temperate forest free-air CO_2 enrichment sites, *New Phytol., 203*(3), 883–899, doi:10.1111/nph.12847.

de Noblet-Ducoudré, N., J.-P. Boisier, A. Pitman, G. B. Bonan, V. Brovkin, F. Cruz, C. Delire, V. Gayler, B. J. J. M. van den Hurk, P. J. Lawrence, M. K. van der Molen, C. Müller, C. H. Reick, B. J. Strengers, and A. Voldoire (2012), Determining robust impacts of land-use-induced land cover changes on surface climate over North America and Eurasia: Results from the first set of LUCID experiments, *J. Clim., 25*(9), 3261–3281, doi:10.1175/JCLI-D-11-00338.1.

Dieleman, W. I. J., S. Vicca, F. A. Dijkstra, F. Hagedorn, M. J. Hovenden, K. S. Larsen, J. A. Morgan, A. Volder, C. Beier, J. S. Dukes, J. King, S. Leuzinger, S. Linder, Y. Luo, R. Oren, P. De Angelis, D. Tingey, M. R. Hoosbeek, and I. A. Janssens (2012), Simple additive effects are rare: A quantitative review of plant biomass and soil process responses to combined manipulations of CO_2 and temperature, *Glob. Change Biol., 18*(9), 2681–2693, doi:10.1111/j.1365-2486.2012.02745.x.

Dietze, M. C., D. S. LeBauer, and R. Kooper (2013), On improving the communication between models and data, *Plant Cell Environ., 36*(9), 1575–1585, doi:10.1111/pce.12043.

Dietze, M. C., S. P. Serbin, C. Davidson, A. R. Desai, X. Feng, R. Kelly, R. Kooper, D. LeBauer, J. Mantooth, K. McHenry, and D. Wang (2014), A quantitative assessment of a terrestrial biosphere model's data needs across North American biomes, *J. Geophys. Res. Biogeosci., 119*(3), 286–300, doi: 10.1002/2013JG002392.

Dijkstra, P., J. C. Blankinship, P. C. Selmants, S. C. Hart, G. W. Koch, E. Schwartz, and B. A. Hungate (2011), Probing carbon flux patterns through soil microbial metabolic networks using parallel position-specific tracer labeling, *Soil Biology and Biochemistry, 43*(1), 126–132, doi: 10.1016/j.soilbio.2010.09.022.

Doetterl, S., A. Stevens, J. Six, R. Merckx, K. Van Oost, M. Casanova Pinto, A. Casanova-Katny, C. Munoz, M. Boudin, E. Zagal Venegas, and P. Boeckx (2015), Soil carbon storage controlled by interactions between geochemistry and climate, *Nat. Geosci., 8*(10), 780–783, doi:10.1038/ngeo2516.

Doney, S. C., I. Lima, J. K. Moore, K. Lindsay, M. J. Behrenfeld, T. K. Westberry, N. Mahowald, D. M. Glover, and T. Takahashi (2009), Skill metrics for confronting global upper ocean ecosystembiogeochemistry models against field and remote sensing data, *J. Mar. Syst., 76*(1–2), 95–112, doi: 10.1016/j.jmarsys.2008.05.015.

Drignei, D., C. E. Forest, and D. Nychka (2008), Parameter estimation for computationally intensive nonlinear regression with an application to climate modeling, *Ann. Appl. Stat., 2*(4), 1217–1230, doi:10.1214/08AOAS210.

Duan, Q., J. Schaake, V. Andrassian, S. Franks, G. Goteti, H. Gupta, Y. Gusev, F. Habets, A. Hall, L. Hay, T. Hogue, M. Huang, G. Leavesley, X. Liang, O. Nasonova, J. Noilhan, L. Oudin, S. Sorooshian, T. Wagener, and E. Wood (2006), Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *J. Hydrol., 320*(1–2), 3–17, doi: 10.1016/j.jhydrol.2005.07.031.

Dukes, J. S., N. R. Chiariello, E. E. Cleland, L. A. Moore, M. R. Shaw, S. Thayer, T. Tobeck, H. A. Mooney, and C. B. Field (2005), Responses of grassland production to single and multiple global environmental changes, *PLoS Biol., 3*(10), e319, doi:10.1371/journal.pbio.0030319.

Eaton, B., J. Gregory, B. Drach, K. Taylor, S. Hankin, J. Caron, R. Signell, P. Bentley, G. Rappa, H. Höck, A. Pamment, and M. Juckes (2011), NetCDF Climate and Forecast (CF) metadata conventions, *Tech. Rep. Version 1.6*.

Edwards, N. R., D. Cameron, and J. Rougier (2011), Precalibrating an intermediate complexity climate model, *Clim. Dyn., 37*(7), 1469–1482, doi:10.1007/s00382-010-0921-0.

Euskirchen, E. S., A. D. McGuire, F. S. Chapin, S. Yi, and C. C. Thompson (2009), Changes in vegetation in northern Alaska under scenarios of climate change, 2003–2100: Implications for climate feedbacks, *Ecol. Appl., 19*(4), 1022–1043, doi:10.1890/08-0806.1.

Eyring, V., M. Righi, A. Lauer, M. Evaldsson, S. Wenzel, C. Jones, A. Anav, O. Andrews, I. Cionni, E. L. Davin, C. Deser, C. Ehbrecht, P. Friedlingstein, P. Gleckler, K.-D. Gottschaldt, S. Hagemann, M. Juckes, S. Kindermann, J. Krasting, D. Kunert, R. Levine, A. Loew, J. Mäkelä, G. Martin, E. Mason, A. S. Phillips, S. Read, C. Rio, R. Roehrig, D. Senftleben, A. Sterl, L. H. van Ulft, J. Walton, S. Wang, and K. D. Williams (2016a), ESMValTool (v1.0) – A community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP, *Geosci. Model Dev., 9*(5), 1747–1802, doi:10.5194/gmd-91747-2016.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor (2016b), Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, *Geosci. Model Dev., 9*(5), 1937–1958, doi:10.5194/gmd-9-1937-2016.

Fekete, B. M., C. J. Vörösmarty, and W. Grabs (2002), High-resolution fields of global runoff combining observed river discharge and simulated water balances, *Global Biogeochem. Cycles, 16*(3), 15–1–15–10, doi:10.1029/1999GB001254.

Ferraro, R., D. E. Waliser, P. Gleckler, K. E. Taylor, and V. Eyring (2015), Evolving Obs4MIPs to support phase 6 of the Coupled Model Intercomparison Project (CMIP6), *Bull. Am. Meteorol. Soc., 96*(8), ES131– ES133, doi:10.1175/BAMS-D-14-00216.1.

Fischer, E. M., K. W. Oleson, and D. M. Lawrence (2012), Contrasting urban and rural heat stress responses to climate change, *Geophys. Res. Lett., 39*(3), doi:10.1029/2011GL050576.

Fisher, R. A., M. Williams, A. L. da Costa, Y. Malhi, R. F. da Costa, S. Almeida, and P. Meir (2007), The response of an Eastern Amazonian rain forest to drought stress: Results and modelling analyses from a throughfall exclusion experiment, *Glob. Change Biol., 13*(11), 2361–2378, doi:10.1111/j.13652486.2007.01417.x.

Fisher, R. A., S. Muszala, M. Verteinstein, P. Lawrence, C. Xu, N. G. McDowell, R. G. Knox, C. Koven, J. Holm, B. M. Rogers, A. Spessa, D. Lawrence, and G. Bonan (2015), Taking off the training wheels: The properties of a dynamic vegetation model without climate envelopes, CLM4.5(ED), *Geosci. Model Dev., 8*(11), 3593–3619, doi:10.5194/gmd-8-3593-2015.

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen (2013), Evaluation of climate models, 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 T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, pp. 741–866, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Frank, D., M. Reichstein, M. Bahn, K. Thonicke, D. Frank, M. D. Mahecha, P. Smith, M. van der Velde, S. Vicca, F. Babst, C. Beer, N. Buchmann, J. G. Canadell, P. Ciais, W. Cramer, A. Ibrom, F. Miglietta, B. Poulter, A. Rammig, S. I. Seneviratne, A. Walz, M. Wattenbach, M. A. Zavala, and J. Zscheischler (2015), Effects of climate extremes on the terrestrial carbon cycle: Concepts, processes and potential future impacts, *Glob. Change Biol., 21*(8), 2861–2880, doi:10.1111/gcb.12916.

Friedlingstein, P., P. M. Cox, R. A. Betts, L. Bopp, W. von Bloh, V. Brovkin, S. C. Doney, M. Eby, I. Fung, B. Govindasamy, J. John, C. D. Jones, F. Joos, T. Kato, M. Kawamiya, W. Knorr, K. Lindsay, H. D. Matthews, T. Raddatz, P. Rayner, C. Reick, E. Roeckner, K.-G. Schnitzler, R. Schnur, K. Strassmann, S. Thompson, A. J. Weaver, C. Yoshikawa, and N. Zeng (2006), Climate–carbon cycle feedback analysis: Results from the C^4MIP model intercomparison, *J. Clim., 19*(14), 3373–3353, doi:10.1175/JCLI3800.1.

Friedlingstein, P., M. Meinshausen, V. K. Arora, C. D. Jones, A. Anav, S. K. Liddicoat, and R. Knutti (2014a), Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks, *J. Clim., 27*(2), 511– 526, doi:10.1175/JCLI-D-12-00579.1.

Friedlingstein, P., R. M. Andrew, J. Rogelj, G. P. Peters, J. G. Canadell, R. Knutti, G. Luderer, M. R. Raupach, M. Schaeffer, D. P. van Vuuren, and C. Le Quéré (2014b), Persistent growth of CO_2 emissions and implications for reaching climate targets, *Nat. Geosci., 7*(10), 709–715, doi:10.1038/ngeo2248.

Fuccillo, K. K., T. M. Crimmins, C. E. de Rivera, and T. S. Elder (2015), Assessing accuracy in citizen science-based plant phenology monitoring, *Int. J. Biometeorol., 59*(7), 917–926, doi:10.1007/s00484014-0892-7.

Furtado, J. C., J. L. Cohen, A. H. Butler, E. E. Riddle, and A. Kumar (2015), Eurasian snow cover variability and links to winter climate in the CMIP5 models, *Clim. Dyn., 45*(9), 2591–2605, doi:10.1007/s00382015-2494-4.

Ghanem, R. G., and P. D. Spanos (1991), *Stochastic Finite Elements: A Spectral Approach*, Sprinter Verlag, New York.

Ghimire, B., W. J. Riley, C. D. Koven, M. Mu, and J. T. Randerson (2016), Representing leaf and root physiological traits in CLM improves global carbon and nitrogen cycling predictions, *J. Adv. Model. Earth Syst., 8*(2), 598–613, doi:10.1002/2015MS000538.

Giglio, L., J. T. Randerson, and G. R. van der Werf (2013), Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4), *J. Geophys. Res. Biogeosci., 118*(1), 317–328, doi:10.1002/jgrg.20042.

Giorgi, F., E.-S. Im, E. Coppola, N. S. Diffenbaugh, X. J. Gao, L. Mariotti, and Y. Shi (2011), Higher hydroclimatic intensity with global warming, *J. Clim., 24*(20), 5309–5324, doi:10.1175/2011JCLI3979.1.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2008), Performance metrics for climate models, *J. Geophys. Res. Atmos., 113*(D6), doi:10.1029/2007JD008972.

Gleckler, P. J., C. Doutriaux, P. J. Durack, K. E. Taylor, Y. Zhang, D. N. Williams, E. Mason, and J. Servonnat (2016), A more powerful reality test for climate models, *Eos Trans. AGU, 97*, doi: 10.1029/2016EO051663.

Gong, W., Q. Duan, J. Li, C. Wang, Z. Di, Y. Dai, A. Ye, and C. Miao (2015), Multi-objective parameter optimization of common land model using adaptive surrogate modeling, *Hydrol. Earth Syst. Sci., 19*(5), 2409–2425, doi:10.5194/hess-19-2409-2015.

Goodman, J., and J. Weare (2010), Ensemble samplers with affine invariance, *Comm. App. Math Comp. Sci., 5*(1), 65–80, doi:10.2140/camcos.2010.5.65.

Goodwin, N. (2015), Bridging the gap between deterministic and probabilistic uncertainty quantification using advanced proxy based methods, in *Proceedings of the SPE Reservoir Simulation Symposium (23–25 February 2015)*, Society of Petroleum Engineers (SPE), Houston, Texas, USA, doi:10.2118/173301-MS.

Gotangco-Castillo, C. K., S. Levis, and P. Thornton (2012), Evaluation of the new CNDV option of the Community Land Model: Effects of dynamic vegetation and interactive nitrogen on CLM4 means and variability, *J. Clim., 25*(11), 3702–3714, doi:10.1175/JCLI-D-11-00372.1.

Grant, R. F., D. D. Baldocchi, and S. Ma (2012a), Ecological controls on net ecosystem productivity of a seasonally dry annual grassland under current and future climates: Modelling with ecosys, *Agr. Forest Meteorol., 152*, 189–200, doi:10.1016/j.agrformet.2011.09.012.

Grant, R. F., A. R. Desai, and B. N. Sulman (2012b), Modelling contrasting responses of wetland productivity to changes in water table depth, *Biogeosci., 9*(11), 4215–4231, doi:10.5194/bg-9-4215-2012.

Graven, H. D., R. F. Keeling, S. C. Piper, P. K. Patra, B. B. Stephens, S. C. Wofsy, L. R. Welp, C. Sweeney, P. P. Tans, J. J. Kelley, B. C. Daube, E. A. Kort, G. W. Santoni, and J. D. Bent (2013), Enhanced seasonal exchange of CO_2 by northern ecosystems since 1960, *Science, 341*(6150), 1085–1089, doi: 10.1126/science.1239207.

Grossman, D. (2016), Amazon rainforest to get a growth check, *Science, 352*(6286), 635–636, doi: 10.1126/science.352.6286.635.

Guanter, L., Y. Zhang, M. Jung, J. Joiner, M. Voigt, J. A. Berry, C. Frankenberg, A. R. Huete, P. Zarco-Tejada, J.-E. Lee, M. S. Moran, G. Ponce-Campos, C. Beer, G. Camps-Valls, N. Buchmann, D. Gianelle, K. Klumpp, A. Cescatti, J. M. Baker, and T. J. Griffis (2014), Global and time-resolved monitoring of crop photosynthesis with chlorophyll fluorescence, *Proc. Nat. Acad. Sci., 111*(14), E1327–E1333, doi: 10.1073/pnas.1320008111.

Hall, D. K., G. A. Riggs, and V. V. Salomonson (2006), MODIS/Terra snow cover daily L3 global 500m grid V005, National Snow and Ice Data Center, Boulder, Colorado, USA, Digital media.

Hall, D. K., G. A. Riggs, J. L. Foster, and S. V. Kumar (2010), Development and evaluation of a cloud-gap-filled MODIS daily snow-cover product, *Remote Sens. Environ., 114*(3), 496–503, doi: 10.1016/j.rse.2009.10.007.

Hammond, G. E., P. C. Lichtner, and R. T. Mills (2014), Evaluating the performance of parallel subsurface simulators: An illustrative example with PFLOTRAN, *Water Resour. Res., 50*(1), 208–228, doi: 10.1002/2012WR013483.

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend (2013), High-resolution global maps of 21st-century forest cover change, *Science, 342*(6160), 850–853, doi:10.1126/science.1244693.

Hantson, S., A. Arneth, S. P. Harrison, D. I. Kelley, I. C. Prentice, S. S. Rabin, S. Archibald, F. Mouillot, S. R. Arnold, P. Artaxo, D. Bachelet, P. Ciais, M. Forrest, P. Friedlingstein, T. Hickler, J. O. Kaplan, S. Kloster, W. Knorr, G. Lasslop, F. Li, S. Mangeon, J. R. Melton, A. Meyn, S. Sitch, A. Spessa, G. R. van der Werf, A. Voulgarakis, and C. Yue (2016), The status and challenge of global fire modelling, *Biogeosci., 13*(11), 3359–3375, doi:10.5194/bg-13-3359-2016.

Harden, J. W., C. D. Koven, C.-L. Ping, G. Hugelius, A. D. McGuire, P. Camill, T. Jorgenson, P. Kuhry, G. J. Michaelson, J. A. O'Donnell, E. A. G. Schuur, C. Tarnocai, K. Johnson, and G. Grosse (2012), Field information links permafrost carbon to physical vulnerabilities of thawing, *Geophys. Res. Lett., 39*(15), doi:10.1029/2012GL051958.

Haughton, N., G. Abramowitz, A. J. Pitman, D. Or, M. J. Best, H. R. Johnson, G. Balsamo, A. Boone, M. Cuntz, B. Decharme, P. A. Dirmeyer, J. Dong, M. Ek, Z. Guo, V. Haverd, B. J. J. van den Hurk, G. S. Nearing, B. Pak, J. A. Santanello Jr., L. E. Stevens, and N. Vuichard (2016), The plumbing of land surface models: Is poor performance a result of methodology or data quality?, *J. Hydrometeor., 17*(6), 1705–1723, doi:10.1175/JHM-D-15-0171.1.

He, Y., S. E. Trumbore, M. S. Torn, J. W. Harden, L. J. S. Vaughn, S. D. Allison, and J. T. Randerson (2016), Radiocarbon constraints imply reduced carbon uptake by soils during the 21st century, *Science, 353*(6306), 1419–1424, doi:10.1126/science.aad4273.

Hengl, T., J. M. de Jesus, R. A. MacMillan, N. H. Batjes, G. B. M. Heuvelink, E. Ribeiro, A. Samuel-Rosa, B. Kempen, J. G. B. Leenaars, M. G. Walsh, and M. R. Gonzalez (2014), SoilGrids1km — Global soil information based on automated mapping, *PLoS ONE, 9*(8), 1–17, doi:10.1371/journal.pone.0105992.

Hengl, T, J. Mendes de Jesus, G. B. M. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, M. A. Guevara, R. Vargas, R. A. MacMillan, N. H. Batjes, J. G. B. Leenaars, E. Ribeiro, I. Wheeler, S. Mantel, and B. Kempen (2017), SoilGrids250m: Global gridded soil information based on machine learning, *PLoS ONE*, 12(2), e0169748, doi:10.1371/journal.pone.0169748.

Hey, T., S. Tansley, and K. Tolle (Eds.) (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, 284 pp., Microsoft Corporation, Redmond, Washington, USA.

Hickler, T., B. Smith, I. C. Prentice, K. Mjöfors, P. Miller, A. Arneth, and M. T. Sykes (2008), CO_2 fertilization in temperate FACE experiments not representative of boreal and tropical forests, *Glob. Change Biol., 14*(7), 1531–1542, doi:10.1111/j.1365-2486.2008.01598.x.

Higdon, D., J. Gattiker, B. Williams, and M. Rightley (2008), Computer model calibration using high-dimensional output, *J. Am. Stat. Assoc., 103*(482), 570–583, doi:10.1198/016214507000000888.

Hoffman, F. M., J. W. Larson, R. T. Mills, B.-G. J. Brooks, A. R. Ganguly, W. W. Hargrove, J. Huang, J. Kumar, and R. R. Vatsavai (2011), Data Mining in Earth System Science (DMESS 2011), in *Proceedings of the International Conference on Computational Science (ICCS 2011), Procedia Comput. Sci.*, vol. 4, edited by M. Sato, S. Matsuoka, P. M. Sloot, G. D. van Albada, and J. Dongarra, pp. 1450–1455, Elsevier, Amsterdam, doi:10.1016/j.procs.2011.04.157.

Hoffman, F. M., J. Kumar, R. T. Mills, and W. W. Hargrove (2013), Representativeness-based sampling network design for the State of Alaska, *Landscape Ecol., 28*(8), 1567–1586, doi:10.1007/s10980-0139902-0.

Hoffman, F. M., J. T. Randerson, V. K. Arora, Q. Bao, P. Cadule, D. Ji, C. D. Jones, M. Kawamiya, S. Khatiwala, K. Lindsay, A. Obata, E. Shevliakova, K. D. Six, J. F. Tjiputra, E. M. Volodin, and T. Wu (2014), Causes and implications of persistent atmospheric carbon dioxide biases in Earth System Models, *J. Geophys. Res. Biogeosci., 119*(2), 141–162, doi:10.1002/2013JG002381.

Holden, P. B., N. R. Edwards, K. I. C. Oliver, T. M. Lenton, and R. D. Wilkinson (2010), A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1, *Clim. Dyn., 35*(5), 785–806, doi:10.1007/s00382-009-0630-8.

Hou, Z., M. Huang, L. R. Leung, G. Lin, and D. M. Ricciuto (2012), Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the Community Land Model, *J. Geophys. Res. Atmos., 117*(D15), doi:10.1029/2012JD017521.

Houghton, R. A. (2013), Keeping management effects separate from environmental effects in terrestrial carbon accounting, *Glob. Change Biol., 19*(9), 2609–2612, doi:10.1111/gcb.12233.

Houghton, R. A., J. I. House, J. Pongratz, G. R. van der Werf, R. S. DeFries, M. C. Hansen, C. Le Quéré and N. Ramankutt (2012), Carbon emissions from land use and land-cover change, *Biogeosci., 9*(12), 5125–5142, doi:10.5194/bg-9-5125-2012.

Huang, M., Z. Hou, L. R. Leung, Y. Ke, Y. Liu, Z. Fang, and Y. Sun (2013), Uncertainty analysis of runoff simulations and parameter identifiability in the Community Land Model: Evidence from MOPEX basins, *J. Hydrometeor., 14*(6), 1754–1772, doi:10.1175/JHM-D-12-0138.1.

Huang, M., J. Ray, Z. Hou, H. Ren, Y. Liu, and L. Swiler (2016), On the applicability of surrogate-based Markov chain Monte Carlo-Bayesian inversion to the Community Land Model: Case studies at flux tower sites, *J. Geophys. Res. Atmos., 12*(13), 7548–7563, doi:10.1002/2015JD024339.

Hugelius, G., J. Strauss, S. Zubrzycki, J. W. Harden, E. A. G. Schuur, C.-L. Ping, L. Schirrmeister, G. Grosse, G. J. Michaelson, C. D. Koven, J. A. O'Donnell, B. Elberling, U. Mishra, P. Camill, Z. Yu, J. Palmtag, and P. Kuhry (2014), Estimated stocks of circumpolar permafrost carbon with quantified uncertainty ranges and identified data gaps, *Biogeosci., 11*(23), 6573–6593, doi:10.5194/bg-11-6573-2014.

Huntzinger, D. N., C. Schwalm, A. M. Michalak, K. Schaefer, A. W. King, Y. Wei, A. Jacobson, S. Liu, R. B. Cook, W. M. Post, G. Berthier, D. Hayes, M. Huang, A. Ito, H. Lei, C. Lu, J. Mao, C. H. Peng, S. Peng, B. Poulter, D. Riccuito, X. Shi, H. Tian, W. Wang, N. Zeng, F. Zhao, and Q. Zhu (2013), The North American Carbon Program Multi-Scale Synthesis and Terrestrial Model Intercomparison Project — Part 1: Overview and experimental design, *Geosci. Model Dev., 6*(6), 2121–2133, doi:10.5194/gmd6-2121-2013.

Huntzinger, D. N., C. R. Schwalm, A. M. Michalak, K. Schaefer, Y. Wei, R. B. Cook, and A. R. Jacobson (2014), NACP MsTMIP: Summary of model structure and characteristics, Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA, doi:10.3334/ORNLDAAC/1228.

Huntzinger, D. N., et al. (2016), NACP MsTMIP: Global 0.5-deg terrestrial biosphere model outputs (version 1) in standard format, Data set, available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA, doi:10.3334/ORNLDAAC/1225.

Ivanova, D. P., P. J. Gleckler, K. E. Taylor, P. J. Durack, and K. D. Marvel (2016), Moving beyond the total sea ice extent in gauging model biases, *J. Clim.*, accepted.

Järvinen, H., P. Räisänen, M. Laine, J. Tamminen, A. Ilin, E. Oja, A. Solonen, and H. Haario (2010), Estimation of ECHAM5 climate model closure parameters with adaptive MCMC, *Atmos. Chem. Phys., 10*(20), 9993–10,002, doi:10.5194/acp-10-9993-2010.

Jenkinson, D. S., P. B. S. Hart, J. H. Rayner, and L. C. Parry (1987), Modelling the turnover of organic matter in long-term experiments at Rothamsted, *Tech. rep.*, Food and Agriculture Organization of the United Nations.

Jones, C. D., V. Arora, P. Friedlingstein, L. Bopp, V. Brovkin, J. Dunne, H. Graven, F. Hoffman, T. Ilyina, J. G. John, M. Jung, M. Kawamiya, C. Koven, J. Pongratz, T. Raddatz, J. Randerson, and S. Zaehle (2016), C^4MIP – The Coupled Climate–Carbon Cycle Model Intercomparison Project: Experimental protocol for CMIP6, *Geosci. Model Dev., 9*(8), 2853–2880, doi:10.5194/gmd-9-2853-2016.

Jung, M., M. Reichstein, H. A. Margolis, A. Cescatti, A. D. Richardson, M. A. Arain, A. Arneth, C. Bernhofer, D. Bonal, J. Chen, D. Gianelle, N. Gobron, G. Kiely, W. Kutsch, G. Lasslop, B. E. Law, A. Lindroth, L. Merbold, L. Montagnani, E. J. Moors, D. Papale, M. Sottocornola, F. Vaccari, and C. Williams (2011), Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, *J. Geophys. Res. Biogeosci., 116*(G3), doi:10.1029/2010JG001566.

Kaiser, J. W., A. Heil, M. O. Andreae, A. Benedetti, N. Chubarova, L. Jones, J.-J. Morcrette, M. Razinger, M. G. Schultz, M. Suttie, and G. R. van der Werf (2012), Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power, *Biogeosci., 9*(1), 527–554, doi: 10.5194/bg-9-527-2012.

Kattge, J., S. Díaz, S. Lavorel, I. C. Prentice, P. Leadley, G. Bönisch, E. Garnier, M. Westoby, P. B. Reich, I. J. Wright, J. H. C. Cornelissen, C. Violle, S. P. Harrison, P. M. van Bodegom, M. Reichstein, B. J. Enquist, N. A. Soudzilovskaia, D. D. Ackerly, M. Anand, O. Atkin, M. Bahn, T. R. Baker, D. Baldocchi, R. Bekker, C. C. Blanco, B. Blonder, W. J. Bond, R. Bradstock, D. E. Bunker, F. Casanoves, J. Cavender-Bares, J. Q. Chambers, F. S. Chapin III, J. Chave, D. Coomes, W. K. Cornwell, J. M. Craine, B. H. Dobrin, L. Duarte, W. Durka, J. Elser, G. Esser, M. Estiarte, W. F. Fagan, J. Fang, F. Fernández-Méndez, A. Fidelis, B. Finegan, O. Flores, H. Ford, D. Frank, G. T. Freschet, N. M. Fyllas, R. V. Gallagher, W. A. Green, A. G. Gutierrez, T. Hickler, S. I. Higgins, J. G. Hodgson, A. Jalili, S. Jansen, C. A. Joly, A. J. Kerkhoff, D. Kirkup, K. Kitajima, M. Kleyer, S. Klotz, J. M. H. Knops, K. Kramer, I. Kühn, H. Kurokawa, D. Laughlin, T. D. Lee, M. Leishman, F. Lens, T. Lenz, S. L. Lewis, J. Lloyd, J. Llusià, F. Louault, S. Ma, M. D. Mahecha, P. Manning, T. Massad, B. E. Medlyn, J. Messier, A. T. Moles, S. C. Müller, K. Nadrowski, S. Naeem, Ü. Niinemets, S. Nöllert, A. Nüske, R. Ogaya, J. Oleksyn, V. G. Onipchenko, Y. Onoda, J. Ordoñez, G. Overbeck, W. A. Ozinga, S. Patiño, S. Paula, J. G. Pausas, J. Peñuelas, O. L. Phillips, V. Pillar, H. Poorter, L. Poorter, P. Poschlod, A. Prinzing, R. Proulx, A. Rammig, S. Reinsch, B. Reu, L. Sack, B. Salgado-Negret, J. Sardans, S. Shiodera, B. Shipley, A. Siefert, E. Sosinski, J.F. Soussana, E. Swaine, N. Swenson, K. Thompson, P. Thornton, M. Waldram, E. Weiher, M. White, S. White, S. J. Wright, B. Yguel, S. Zaehle, A. E. Zanne, and C. Wirth (2011), TRY — A global database of plant traits, *Glob. Change Biol., 17*(9), 2905–2935, doi:10.1111/j.1365-2486.2011.02451.x.

Keenan, T. F., E. Davidson, A. M. Moffat, W. Munger, and A. D. Richardson (2012), Using model-data fusion to interpret past trends, and quantify uncertainties in future projections, of terrestrial ecosystem carbon cycling, *Glob. Change Biol., 18*(8), 2255–2569, doi:10.1111/j.1365-2486.2012.02684.x.

Keenan, T. F., D. Y. Hollinger, G. Bohrer, D. Dragoni, J. W. Munger, H. P. Schmid, and A. D. Richardson (2013), Increase in forest water-use efficiency as atmospheric carbon dioxide concentrations rise, *Nature, 499*(7458), 324–327, doi:10.1038/nature12291.

Keiluweit, M., J. J. Bougoure, P. S. Nico, J. Pett-Ridge, P. K. Weber, and M. Kleber (2015), Mineral protection of soil carbon counteracted by root exudates, *Nature Clim. Change, 5*(6), 588–595, doi: 10.1038/nclimate2580.

Keppel-Aleks, G., J. T. Randerson, K. Lindsay, B. B. Stephens, J. K. Moore, S. C. Doney, P. E. Thornton, N. M. Mahowald, F. M. Hoffman, C. Sweeney, P. P. Tans, P. O. Wennberg, and S. C. Wofsy (2013), Atmospheric carbon dioxide variability in the Community Earth System Model: Evaluation and transient dynamics during the twentieth and twenty-first centuries, *J. Clim., 26*(13), 4447–4475, doi:10.1175/JCLID-12-00589.1.

Keppel-Aleks, G., A. S. Wolf, M. Mu, S. C. Doney, D. C. Morton, P. S. Kasibhatla, J. B. Miller, E. J. Dlugokencky, and J. T. Randerson (2014), Separating the influence of temperature, drought, and fire on interannual variability in atmospheric CO_2, *Global Biogeochem. Cycles, 28*(11), 1295–1310, doi: 10.1002/2014GB004890.

Kim, H. (2010), Role of rivers in the spatiotemporal variations of terrestrial hydrological circulations, Ph.D. thesis, University of Tokyo.

Kim, H., P. J.-F. Yeh, T. Oki, and S. Kanae (2009), Role of rivers in the seasonal variations of terrestrial water storage over global basins, *Geophys. Res. Lett., 36*(17), doi:10.1029/2009GL039006.

Kim, H., et al. (2016), A century-long global surface meteorology for offline terrestrial simulations, in preparation.

Kloster, S., N. M. Mahowald, J. T. Randerson, P. E. Thornton, F. M. Hoffman, S. Levis, P. J. Lawrence, J. J. Feddema, K. W. Oleson, and D. M. Lawrence (2010), Fire dynamics during the 20th century simulated by the Community Land Model, *Biogeosci., 7*(6), 1877–1902, doi:10.5194/bg-7-1877-2010.

Knutti, R., and J. Sedlacek (2013), Robustness and uncertainties in the new CMIP5 climate model projections, *Nature Clim. Change, 3*(4), 369–373, doi:10.1038/nclimate1716.

Koster, R. D., M. J. Suarez, and M. Heiser (2000), Variance and predictability of precipitation at seasonal-to-interannual timescales, *J. Hydrometeor., 1*(1), 26–46, doi:10.1175/15257541(2000)001%3C0026:VAPOPA%3E2.0.CO;2.

Koster, R. D., P. A. Dirmeyer, Z. C. Guo, G. Bonan, E. Chan, P. Cox, C. T. Gordon, S. Kanae, E. Kowalczyk, D. Lawrence, P. Liu, C. H. Lu, S. Malyshev, B. McAvaney, K. Mitchell, D. Mocko, T. Oki, K. Ole-son, A. Pitman, Y. C. Sud, C. M. Taylor, D. Verseghy, R. Vasic, Y. K. Xue, T. Yamada, et al. (2004), Regions of strong coupling between soil moisture and precipitation, *Science, 305*(5687), 1138–1140, doi:10.1126/science.1100217.

Koven, C. D., B. Ringeval, P. Friedlingstein, P. Ciais, P. Cadule, D. Khvorostyanov, G. Krinner, and C. Tarnocai (2011), Permafrost carbon-climate feedbacks accelerate global warming, *Proc. Nat. Acad. Sci., 108*(36), 14,769–14,774, doi:10.1073/pnas.1103910108.

Koven, C. D., W. J. Riley, and A. Stern (2013), Analysis of permafrost thermal dynamics and response to climate change in the CMIP5 Earth system models, *J. Clim., 26*(6), 1877–1900, doi:10.1175/JCLI-D-1200228.1.

Koven, C. D., J. Q. Chambers, K. Georgiou, R. Knox, R. Negron-Juarez, W. J. Riley, V. K. Arora, V. Brovkin, P. Friedlingstein, and C. D. Jones (2015), Controls on terrestrial carbon feedbacks by productivity vs. turnover in the CMIP5 Earth System Models, *Biogeosci., 12*(17), 5211–5228, doi:10.5194/bg-12-52112015.

Kuczera, G., D. Kavetski, B. Renard, and M. Thyer (2010), A limited-memory acceleration strategy for MCMC sampling in hierarchical Bayesian calibration of hydrological models, *Water Resour. Res., 46*(7), doi:10.1029/2009WR008985.

Kumar, J., J. Weiner, W. W. Hargrove, S. P. Norman, F. M. Hoffman, and D. Newcomb (2015), Characterization and classification of vegetation canopy structure and distribution within the Great Smoky Mountains National Park using LiDAR, in *Proceedings of the 15th IEEE International Conference on Data Mining Workshops (ICDMW 2015)*, edited by P. Cui, J. Dy, C. Aggarwal, Z.-H. Zhou, A. Tuzhilin, H. Xiong, and X. Wu, pp. 1478–1485, Institute of Electrical and Electronics Engineers (IEEE), Conference Publishing Services (CPS), doi:10.1109/ICDMW.2015.178.

Kumar, J., F. M. Hoffman, W. W. Hargrove, and N. Collier (2016), Understanding the representativeness of FLUXNET for upscaling carbon flux from eddy covariance measurements, *Earth Syst. Sci. Data Discuss.*, 2016, 1–25, doi:10.5194/essd-2016-36.

Kumar, S. V., C. D. Peters-Lidard, Y. Tian, P. R. Houser, J. Geiger, S. Olden, L. Lighty, J. L. Eastman, B. Doty, P. Dirmeyer, J. Adams, K. Mitchell, E. F. Wood, and J. Sheffield (2006), Land information system: An interoperable framework for high resolution land surface modeling, *Environ. Modell. Softw., 21*(10), 1402–1415, doi:10.1016/j.envsoft.2005.07.004.

Kumar, S. V., C. D. Peters-Lidard, J. Santanello, K. Harrison, Y. Liu, and M. Shaw (2012), Land surface Verification Toolkit (LVT) — A generalized framework for land surface model evaluation, *Geosci. Model Dev., 5*(3), 869–886, doi:10.5194/gmd-5-869-2012.

Laloy, E., and J. A. Vrugt (2012), High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS) and high-performance computing, *Water Resour. Res., 48*(1), doi: 10.1029/2011WR010608.

Landerer, F. W., and S. C. Swenson (2012), Accuracy of scaled grace terrestrial water storage estimates, *Water Resour. Res., 48*(4), doi:10.1029/2011WR011453.

Langford, Z., J. Kumar, F. M. Hoffman, R. J. Norby, S. D. Wullschleger, V. L. Sloan, and C. M. Iversen (2016), Mapping Arctic plant functional type distributions in the Barrow Environmental Observatory using WorldView-2 and LiDAR datasets, *Remote Sens., 8*(9), 733, doi:10.3390/rs8090733.

Lasslop, G., M. Reichstein, D. Papale, A. D. Richardson, A. Arneth, A. Barr, P. Stoy, and G. Wohlfahrt (2010), Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: Critical issues and global evaluation, *Glob. Change Biol., 16*(1), 187–208, doi: 10.1111/j.1365-2486.2009.02041.x.

Lawrence, D. M., G. C. Hurtt, A. Arneth, V. Brovkin, K. V. Calvin, A. D. Jones, C. D. Jones, P. J. Lawrence, N. de Noblet-Ducoudré, J. Pongratz, S. I. Seneviratne, and E. Shevliakova (2016), The Land Use Model Intercomparison Project (LUMIP): Rationale and experimental design, *Geosci. Model Dev., 9*(9), 2973– 2998, doi:10.5194/gmd-9-2973-2016.

Lawrence, P. J., J. J. Feddema, G. B. Bonan, G. A. Meehl, B. C. O'Neill, K. W. Oleson, S. Levis, D. M. Lawrence, E. Kluzek, K. Lindsay, and P. E. Thornton (2012), Simulating the biogeochemical and biogeophysical impacts of transient land cover change and wood harvest in the Community Climate System Model (CCSM4) from 1850 to 2100, *J. Clim., 25*(9), 3071–3095, doi:10.1175/JCLI-D-11-00256.1.

Le Maitre, O., and O. M. Knio (2010), *Sepctral Methods for Uncertainty Quantification*, Springer, New York, doi:10.1007/978-90-481-3520-2.

Le Quéré, C., R. Moriarty, R. M. Andrew, J. G. Canadell, S. Sitch, J. I. Korsbakken, P. Friedlingstein, G. P. Peters, R. J. Andres, T. A. Boden, R. A. Houghton, J. I. House, R. F. Keeling, P. Tans, A. Arneth, D. C. E. Bakker, L. Barbero, L. Bopp, J. Chang, F. Chevallier, L. P. Chini, P. Ciais, M. Fader, R. A. Feely, T. Gkritzalis, I. Harris, J. Hauck, T. Ilyina, A. K. Jain, E. Kato, V. Kitidis, K. Klein Goldewijk, C. Koven, P. Landschützer, S. K. Lauvset, N. Lefèvre, A. Lenton, I. D. Lima, N. Metzl, F. Millero, D. R. Munro, A. Murata, J. E. M. S. Nabel, S. Nakaoka, Y. Nojiri, K. O'Brien, A. Olsen, T. Ono, F. F. Pérez, B. Pfeil, D. Pierrot, B. Poulter, G. Rehder, C. R¨odenbeck, S. Saito, U. Schuster, J. Schwinger, R. Séférian, T. Steinhoff, B. D. Stocker, A. J. Sutton, T. Takahashi, B. Tilbrook, I. T. van der Laan-Luijkx, G. R. van der Werf, S. van Heuven, D. Vandemark, N. Viovy, A. Wiltshire, S. Zaehle, and N. Zeng (2015), Global carbon budget 2015, *Earth Syst. Sci. Data, 7*(2), 349–396, doi:10.5194/essd-7-349-2015.

LeBauer, D. S., D. Wang, K. T. Richter, C. C. Davidson, and M. C. Dietze (2013), Facilitating feedbacks between field measurements and ecosystem models, *Ecol. Monogr., 83*(2), 133–154, doi:10.1890/120137.1.

Lee, X., M. L. Goulden, D. Y. Hollinger, A. Barr, T. A. Black, G. Bohrer, R. Bracho, B. Drake, A. Goldstein, L. Gu, G. Katul, T. Kolb, B. E. Law, H. Margolis, T. Meyers, R. Monson, W. Munger, R. Oren, K. T. Paw U, A. D. Richardson, H. P. Schmid, R. Staebler, S. Wofsy, and L. Zhao (2011), Observed increase in local cooling effect of deforestation at higher latitudes, *Nature, 479*(7373), 384–387, doi:10.1038/nature10588.

Lehmann, J., and M. Kleber (2015), The contentious nature of soil organic matter, *Nature, 528*(7580), 60– 68, doi:10.1038/nature16069.

Lejeune, Q., S. I. Seneviratne, and E. L. Davin (2016), Comparative assessment of mid-latitude land-cover change effects on temperature in historical LUCID and CMIP5 simulations, *J. Clim.*, submitted.

Leuning, R. (1995), A critical appraisal of a combined stomatal-photosynthesis model for C_3 plants, *Plant Cell Environ., 18*(4), 339–355, doi:10.1111/j.1365-3040.1995.tb00370.x.

Li, F., S. Levis, and D. S. Ward (2013), Quantifying the role of fire in the Earth system – Part 1: Improved global fire modeling in the Community Earth System Model (CESM1), *Biogeosci., 10*(4), 2293–2314, doi:10.5194/bg-10-2293-2013.

Li, J., G. Wang, S. D. Allison, M. A. Mayes, and Y. Luo (2014), Soil carbon sensitivity to temperature and carbon use efficiency compared across microbial-ecosystem models of varying complexity, *Biogeochemistry, 119*(1), 67–84, doi:10.1007/s10533-013-9948-8.

Li, Y., M. Zhao, S. Motesharrei, Q. Mu, E. Kalnay, and S. Li (2015), Local cooling and warming effects of forests based on satellite observations, *Nat. Commun., 6*, 6603, doi:10.1038/ncomms7603.

Lindsay, K., G. B. Bonan, S. C. Doney, F. M. Hoffman, D. M. Lawrence, M. C. Long, N. M. Mahowald, J. K. Moore, J. T. Randerson, and P. E. Thornton (2014), Preindustrial-control and twentieth-century carbon cycle experiments with the Earth system model CESM1(BGC), *J. Clim., 27*(24), 8981–9005, doi: 10.1175/JCLI-D-12-00565.1.

Liu, Y., G. Bisht, Z. M. Subin, W. J. Riley, and G. S. H. Pau (2016a), A hybrid reduced-order model of fine-resolution hydrologic simulations at a polygonal tundra site, *Vadose Zone J., 15*(2), doi: 10.2136/vzj2015.05.0068.

Liu, Y., G. S. H. Pau, and S. Finsterle (2016b), Implicit sampling combined with reduced order modeling for the inversion of vadose zone hydrological data, submitted.

Liu, Y. Y., A. I. J. M. van Dijk, R. A. M. de Jeu, J. G. Canadell, M. F. McCabe, J. P. Evans, and G. Wang (2015), Recent reversal in loss of global terrestrial biomass, *Nature Clim. Change, 5*(5), 470–474, doi: 10.1038/nclimate2581.

Lo, M.-H., J. S. Famiglietti, P. J.-F. Yeh, and T. H. Syed (2010), Improving parameter estimation and water table depth simulation in a land surface model using GRACE water storage and estimated base flow data, *Water Resour. Res., 46*(5), doi:10.1029/2009WR007855.

Lucas, R., J. Ang, K. Bergman, S. Borkar, W. Carlson, L. Carrington, G. Chiu, R. Colwell, W. Dally, J. Dongarra, A. Geist, R. Haring, J. Hittinger, A. Hoisie, D. Klein, P. Kogge, R. Lethin, V. Sarkar, R. Schreiber, J. Shalf, T. Sterling, R. Stevens, J. Bashor, R. Brightwell, P. Coteus, E. Debenedictus, J. Hiller, K. H. Kim, H. Langston, R. Murphy, C. Webster, S. Wild, G. Grider, R. Ross, S. Leyffer, and J. Laros III (2014), DOE Advanced Scientific Computing Advisory Subcommittee (ASCAC) report: Top ten exascale research challenges, *Technical report*, U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Germantown, Maryland, USA, doi:10.2172/1222713.

Luo, Y. (2001), Transient ecosystem responses to Free-Air CO_2 Enrichment (FACE): Experimental evidence and methods of analysis, *New Phytol., 152*(1), 3–8, doi:10.1046/j.0028-646X.2001.00247.x.

Luo, Y., and E. Weng (2011), Dynamic disequilibrium of the terrestrial carbon cycle under global change, *Trends Ecol. Evol., 26*(2), 96–104, doi:10.1016/j.tree.2010.11.003.

Luo, Y., L. W. White, J. G. Canadell, E. H. DeLucia, D. S. Ellsworth, A. Finzi, J. Lichter, and W. H. Schlesinger (2003), Sustainability of terrestrial carbon sequestration: A case study in Duke Forest with inversion approach, *Global Biogeochem. Cycles, 17*(1), doi:10.1029/2002GB001923.

Luo, Y., K. Ogle, C. Tucker, S. Fei, C. Gao, S. LaDeau, J. S. Clark, and D. S. Schimel (2011), Ecological forecasting and data assimilation in a data-rich era, *Ecol. Appl., 21*(5), 1429–1442, doi:10.1890/091275.1.

Luo, Y., T. F. Keenan, and M. Smith (2015), Predictability of the terrestrial carbon cycle, *Glob. Change Biol., 21*(5), 1737–1751, doi:10.1111/gcb.12766.

Luo, Y., A. Ahlström, S. D. Allison, N. H. Batjes, V. Brovkin, N. Carvalhais, A. Chappell, P. Ciais, E. A. Davidson, A. Finzi, K. Georgiou, B. Guenet, O. Hararuk, J. W. Harden, Y. He, F. Hopkins, L. Jiang, C. D. Koven, R. B. Jackson, C. D. Jones, M. J. Lara, J. Liang, A. D. McGuire, W. Parton, C. Peng, J. T. Randerson, A. Salazar, C. A. Sierra, M. J. Smith, H. Tian, K. E. O. Todd-Brown, M. Torn, K. J. van Groenigen, Y. P. Wang, T. O. West, Y. Wei, W. R. Wieder, J. Xia, X. Xu, X. Xu, and T. Zhou (2016), Toward more realistic projections of soil carbon dynamics by Earth system models, *Global Biogeochem. Cycles, 30*(1), 40–56, doi:10.1002/2015GB005239.

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.

Luyssaert, S., I. Inglima, M. Jung, A. D. Richardson, M. Reichstein, D. Papale, S. L. Piao, E. D. Schulze, L. Wingate, G. Matteucci, L. Aragao, M. Aubinet, C. Beer, C. Bernhofer, K. G. Black, D. Bonal, J. M. Bonnefond, J. Chambers, P. Ciais, B. Cook, K. J. Davis, A. J. Dolman, B. Gielen, M. Goulden, J. Grace, A. Granier, A. Grelle, T. Griffis, T. Gr¨unwald, G. Guidolotti, P. J. Hanson, R. Harding, D. Y. Hollinger, L. R. Hutyra, P. Kolari, B. Kruijt, W. Kutsch, F. Lagergren, T. Laurila, B. E. Law, G. Le Maire, A. Lindroth, D. Loustau, Y. Malhi, J. Mateus, M. Migliavacca, L. Misson, L. Montagnani, J. Moncrieff, E. Moors, J. W. Munger, E. Nikinmaa, S. V. Ollinger, G. Pita, C. Rebmann, O. Roupsard, N. Saigusa, M. J. Sanz, G. Seufert, C. Sierra, M. L. Smith, J. Tang, R. Valentini, T. Vesala, and I. A. Janssens (2007), CO_2 balance of boreal, temperate, and tropical forests derived from a global database, *Glob. Change Biol., 13*(12), 2509–2537, doi:10.1111/j.1365-2486.2007.01439.x.

Ma, X., M. Al-Harbi, A. Datta-Gupta, and Y. Efendiev (2008), An efficient two-stage sampling method for uncertainty quantification in history matching geological models, *Soc. Petrol. Eng. J., 13*(1), 77–87, doi:10.2118/102476-PA.

Maeda, E. E., H. Kim, L. E. O. C. Arag´ao, J. S. Famiglietti, and T. Oki (2015), Disruption of hydroecological equilibrium in southwest Amazon mediated by drought, *Geophys. Res. Lett., 42*(18), 7546–7553, doi: 10.1002/2015GL065252.

Malyshev, S., E. Shevliakova, R. J. Stouffer, and S. W. Pacala (2015), Contrasting local versus regional effects of land-use-change-induced heterogeneity on historical climate: Analysis with the GFDL Earth system model, *J. Clim., 28*(13), 5448–5469, doi:10.1175/JCLI-D-14-00586.1.

Manzoni, S., S. M. Schaeffer, G. Katul, A. Porporato, and J. P. Schimel (2014), A theoretical analysis of microbial eco-physiological and diffusion limitations to carbon cycling in drying soils, *Soil Biology and Biochemistry, 73*, 69–83, doi:10.1016/j.soilbio.2014.02.008.

Manzoni, S., F. Moyano, T. K¨atterer, and J. Schimel (2016), Modeling coupled enzymatic and solute transport controls on decomposition in drying soils, *Soil Biology and Biochemistry, 95*, 275–287, doi: 10.1016/j.soilbio.2016.01.006.

Mao, J., W. Fu, X. Shi, D. M. Ricciuto, J. B. Fisher, R. E. Dickinson, Y. Wei, W. Shem, S. Piao, K. Wang, C. R. Schwalm, H. Tian, M. Mu, A. Arain, P. Ciais, R. Cook, Y. Dai, D. Hayes, F. M. Hoffman, M. Huang, S. Huang, D. N. Huntzinger, A. Ito, A. Jain, A. W. King, H. Lei, C. Lu, A. M. Michalak, N. Parazoo, C. Peng, S. Peng, B. Poulter, K. Schaefer, E. Jafarov, P. E. Thornton, W. Wang, N. Zeng, Z. Zeng, F. Zhao, Q. Zhu, and Z. Zhu (2015), Disentangling climatic and anthropogenic controls on global terrestrial evapotranspiration trends, *Environ. Res. Lett., 10*(9), 094,008, doi:10.1088/1748-9326/10/9/094008.

Mao, J., D. M. Ricciuto, P. E. Thornton, J. M. Warren, A. W. King, X. Shi, C. M. Iversen, and R. J. Norby (2016), Evaluating the Community Land Model in a pine stand with shading manipulations and $^{13}CO_2$ labeling, *Biogeosci., 13*(3), 641–657, doi:10.5194/bg-13-641-2016.

Martin, J., L. C. Wilcox, C. Burstedde, and O. Ghattas (2012), A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion, *SIAM J. Sci. Comput., 34*(3), A1460–A1487, doi:10.1137/110845598.

Matthews, E. (1997), Global litter production, pools, and turnover times: Estimates from measurement data and regression models, *J. Geophys. Res. Atmos., 102*(D15), 18,771–18,800, doi:10.1029/97JD02956.

Mayes, M. A., K. R. Heal, C. C. Brandt, J. R. Phillips, and P. M. Jardine (2012), Relation between soil order and sorption of dissolved organic carbon in temperate subsoils, *Soil Sci. Soc. Am. J., 76*(3), 1027–1037, doi:10.2136/sssaj2011.0340.

McGuire, A. D., S. Sitch, J. S. Clein, R. Dargaville, G. Esser, J. Foley, M. Heimann, F. Joos, J. Kaplan, D. W. Kicklighter, R. A. Meier, J. M. Melillo, B. Moore, I. C. Prentice, N. Ramankutty, T. Reichenau, A. Schloss, H. Tian, L. J. Williams, and U. Wittenberg (2001), Carbon balance of the terrestrial biosphere in the Twentieth Century: Analyses of CO_2, climate and land use effects with four process-based ecosystem models, *Global Biogeochem. Cycles, 15*(1), 183–206.

McGuire, A. D., C. Koven, D. M. Lawrence, J. S. Clein, J. Xia, C. Beer, E. Burke, G. Chen, X. Chen, C. Delire, E. Jafarov, A. H. MacDougall, S. Marchenko, D. Nicolsky, S. Peng, A. Rinke, K. Saito, W. Zhang, R. Alkama, T. J. Bohn, P. Ciais, B. Decharme, A. Ekici, I. Gouttevin, T. Hajima, D. J. Hayes, D. Ji, G. Krinner, D. P. Lettenmaier, Y. Luo, P. A. Miller, J. C. Moore, V. Romanovsky, C. Schädel, K. Schaefer, E. A. Schuur, B. Smith, T. Sueyoshi, and Q. Zhuang (2016), Variability in the sensitivity among model simulations of permafrost and carbon dynamics in the permafrost region between 1960 and 2009, *Global Biogeochem. Cycles, 30*(7), 1015–1037, doi:10.1002/2016GB005405.

Medlyn, B. E., S. Zaehle, M. G. De Kauwe, A. P. Walker, M. C. Dietze, P. J. Hanson, T. Hickler, A. K. Jain, Y. Luo, W. Parton, I. C. Prentice, P. E. Thornton, S. Wang, Y.-P. Wang, E. Weng, C. M. Iversen, H. R. McCarthy, J. M. Warren, R. Oren, and R. J. Norby (2015), Using ecosystem experiments to improve vegetation models, *Nature Clim. Change, 5*(6), 528–534, doi:10.1038/nclimate2621.

Medlyn, B. E., M. G. De Kauwe, S. Zaehle, A. P. Walker, R. A. Duursma, K. Luus, M. Mishurov, B. Pak, B. Smith, Y.-P. Wang, X. Yang, K. Y. Crous, J. E. Drake, T. E. Gimeno, C. A. Macdonald, R. J. Norby, S. A. Power, M. G. Tjoelker, and D. S. Ellsworth (2016), Using models to guide field experiments: *a priori* predictions for the CO_2 response of a nutrient-and water-limited native Eucalypt woodland, *Glob. Change Biol., 22*(8), 2834–2851, doi:10.1111/gcb.13268.

Medvigy, D., S. C. Wofsy, J. W. Munger, D. Y. Hollinger, and P. R. Moorcroft (2009), Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem Demography model version 2, *J. Geophys. Res. Biogeosci., 114*(G1), doi:10.1029/2008JG000812.

Meehl, G. A., R. Moss, K. E. Taylor, V. Eyring, R. J. Stouffer, S. Bony, and B. Stevens (2014), Climate model intercomparisons: Preparing for the next phase, *Eos Trans. AGU, 95*(9), 77–78, doi: 10.1002/2014EO090001.

Milly, P. C. D., S. L. Malyshev, E. Shevliakova, K. A. Dunne, K. L. Findell, T. Gleeson, Z. Liang, P. Phillipps, R. J. Stouffer, and S. Swenson (2014), An enhanced model of land water and energy for global hydrologic and Earth-system studies, *J. Hydrometeor., 15*(5), 1739–1761, doi:10.1175/JHM-D-130162.1.

Mishra, U., and W. J. Riley (2015), Scaling impacts on environmental controls and spatial heterogeneity of soil organic carbon stocks, *Biogeosci., 12*(13), 3993–4004, doi:10.5194/bg-12-3993-2015.

Mishra, U., J. D. Jastrow, R. Matamala, G. Hugelius, C. D. Koven, J. W. Harden, C. L. Ping, G. J. Michaelson, Z. Fan, R. M. Miller, A. D. McGuire, C. Tarnocai, P. Kuhry, W. J. Riley, K. Schaefer, E. A. G. Schuur, M. T. Jorgenson, and L. D. Hinzman (2013), Empirical estimates to reduce modeling uncertainties of soil organic carbon in permafrost regions: A review of recent progress and remaining challenges, *Environ. Res. Lett., 8*(3), 035,020, doi:10.1088/1748-9326/8/3/035020.

Mishra, U., B. Drewniak, J. D. Jastrow, R. M. Matamala, and U. W. A. Vitharana (2016), Spatial representation of organic carbon and active-layer thickness of high latitude soils in CMIP5 earth system models, *Geoderma*, doi:10.1016/j.geoderma.2016.04.017.

Moorcroft, P. R., G. C. Hurtt, and S. W. Pacala (2001), A method for scaling vegetation dynamics: The Ecosystem Demography model (ED), *Ecol. Monogr., 71*(4), 557–586, doi:10.1890/00129615(2001)071[0557:AMFSVD]2.0.CO;2.

Mu, M., J. T. Randerson, W. J. Riley, C. D. Koven, G. Keppel-Aleks, D. M. Lawrence, and F. M. Hoffman (2016a), International Land model Benchmarking (ILAMB) package v001.00, Software package, doi: 10.18139/ILAMB.v001.00/1251597.

Mu, M., et al. (2016b), Development of version 1 of the International Land Model Benchmarking (ILAMB) system and its application to CMIP5 Earth system models, *J. Adv. Model. Earth Syst.*, in preparation.

Mu, Q., M. Zhao, and S. W. Running (2011), Improvements to a MODIS global terrestrial evapotranspiration algorithm, *Remote Sens. Environ., 115*(8), 1781–1800, doi:10.1016/j.rse.2011.02.019.

Müller, J., R. Paudel, C. A. Shoemaker, J. Woodbury, Y. Wang, and N. Mahowald (2015), CH_4 parameter estimation in CLM4.5bgc using surrogate global optimization, *Geosci. Model Dev., 8*(10), 3285–3310, doi:10.5194/gmd-8-3285-2015.

Myneni, R. B., S. Hoffman, Y. Knyazikhin, J. L. Privette, J. Glassy, Y. Tian, Y. Wang, X. Song, Y. Zhang, G. R. Smith, A. Lotsch, M. Friedl, J. T. Morisette, P. Votava, R. R. Nemani, and S. W. Running (2002), Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data, *Remote Sens. Environ., 83*(1–2), 214–231, doi:10.1016/S0034-4257(02)00074-3.

Nearing, G. S., D. M. Mocko, C. D. Peters-Lidard, S. V. Kumar, and Y. Xia (2016), Benchmarking NLDAS-2 soil moisture and evapotranspiration to separate uncertainty contributions, *J. Hydrometeor., 17*(3), 745– 759, doi:10.1175/JHM-D-15-0063.1.

Negrón-Juárez, R. I., C. D. Koven, W. J. Riley, R. G. Knox, and J. Q. Chambers (2015), Observed allocations of productivity and biomass, and turnover times in tropical forests are not accurately represented in CMIP5 Earth system models, *Environ. Res. Lett., 10*(6), 064,017, doi:10.1088/1748-9326/10/6/064017.

Nicolás, C., J. N. Kennedy, T. Hernández, C. García, and J. Six (2014), Soil aggregation in a semiarid soil amended with composted and non-composted sewage sludge — A field experiment, *Geoderma, 219–220*, 24–31, doi:10.1016/j.geoderma.2013.12.017.

Noormets, A., J. Chen, and T. R. Crow (2007), Age-dependent changes in ecosystem carbon fluxes in managed forests in Northern Wisconsin, USA, *Ecosystems, 10*(2), 187–203, doi:10.1007/s10021-007-9018-y.

Norby, R. J., M. G. De Kauwe, T. F. Domingues, R. A. Duursma, D. S. Ellsworth, D. S. Goll, D. M. Lapola, K. A. Luus, A. R. MacKenzie, B. E. Medlyn, R. Pavlick, A. Rammig, B. Smith, R. Thomas, K. Thonicke, A. P. Walker, X. Yang, and S. Zaehle (2016), Model–data synthesis for the next generation of forest free-air CO_2 enrichment (FACE) experiments, *New Phytol., 209*(1), 17–28, doi:10.1111/nph.13593.

Oleson, K. W., D. M. Lawrence, G. B, M. G. Flanner, E. Kluzek, P. J, S. Levis, S. C. Swenson, E. Thornton, J. Feddema, C. L. Heald, J.-F. Lamarque, G.-Y. Niu, T. Qian, S. Running, K. Sakaguchi, L. Yang, X. Zeng, and X. Zeng (2010), Technical description of version 4.0 of the Community Land Model (CLM), *Technical Note NCAR/TN-503+STR*, National Center for Atmospheric Research, Boulder, Colorado, USA.

Ollinger, S. V., M. L. Smith, M. E. Martin, R. A. Hallett, C. L. Goodale, and J. D. Aber (2002), Regional variation in foliar chemistry and N cycling among forests of diverse history and composition, *Ecology, 83*(2), 339–355, doi:10.1890/0012-9658(2002)083[0339:RVIFCA]2.0.CO;2.

Olson, R., R. Sriver, M. Goes, N. M. Urban, H. D. Matthews, M. Haran, and K. Keller (2012), A climate sensitivity estimate using Bayesian fusion of instrumental observations and an Earth system model, *J. Geophys. Res. Atmos., 117*(D4), doi:10.1029/2011JD016620.

Overpeck, J. T., G. A. Meehl, S. Bony, and D. R. Easterling (2011), Climate data challenges in the 21st century, *Science, 331*(6018), 700–702, doi:10.1126/science.1197869.

Pan, Y., R. A. Birdsey, J. Fang, R. Houghton, P. E. Kauppi, W. A. Kurz, O. L. Phillips, A. Shvidenko, S. L. Lewis, J. G. Canadell, P. Ciais, R. B. Jackson, S. Pacala, A. D. McGuire, S. Piao, A. Rautiainen, S. Sitch, and D. Hayes (2011), A large and persistent carbon sink in the world's forests, *Science*, doi: 10.1126/science.1201609.

Parton, W. J., D. S. Schimel, C. V. Cole, and D. S. Ojima (1987), Analysis of factors controlling soil organic matter levels in Great Plains grasslands, *Soil Sci. Soc. Am. J., 51*(5), 1173–1179, doi: 10.2136/sssaj1987.03615995005100050015x.

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, doi:10.1007/BF02180320.

Parton, W. J., J. M. O. Scurlock, D. S. Ojima, T. G. Gilmanov, R. J. Scholes, D. S. Schimel, T. Kirchner, J.-C. Menaut, T. Seastedt, E. Garcia Moya, A. Kamnalrut, and J. I. Kinyamario (1993), Observations and modeling of biomass and soil organic matter dynamics for the grassland biome worldwide, *Global Biogeochem. Cycles, 7*(4), 785–809, doi:10.1029/93GB02042.

Parton, W. J., J. A. Morgan, G. Wang, and S. Del Grosso (2007), Projected ecosystem impact of the Prairie Heating and CO_2 Enrichment experiment, *New Phytol., 174*(4), 823–834, doi:10.1111/j.14698137.2007.02052.x.

Pau, G. S. H., G. Bisht, and W. J. Riley (2014), A reduced-order modeling approach to represent subgridscale hydrological dynamics for land-surface simulations: Application in a polygonal tundra landscape, *Geosci. Model Dev., 7*(5), 2091–2105, doi:10.5194/gmd-7-2091-2014.

Pau, G. S. H., C. Shen, W. J. Riley, and Y. Liu (2016), Accurate and efficient prediction of fine-resolution hydrologic and carbon dynamic simulations from coarse-resolution models, *Water Resour. Res., 52*(2), 791–812, doi:10.1002/2015WR017782.

Pelletier, J. D., P. D. Broxton, P. Hazenberg, X. Zeng, P. A. Troch, G.-Y. Niu, Z. Williams, M. A. Brunke, and D. Gochis (2016), A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling, *J. Adv. Model. Earth Syst., 8*(1), 41–65, doi: 10.1002/2015MS000526.

Pendall, E., J. L. Heisler-White, D. G. Williams, F. A. Dijkstra, Y. Carrillo, J. A. Morgan, and D. R. LeCain (2013), Warming reduces carbon losses from grassland exposed to elevated atmospheric carbon dioxide, *PLoS ONE, 8*(8), e71,921, doi:10.1371/journal.pone.0071921.

Petropoulos, G. P., G. Ireland, and B. Barrett (2015), Surface soil moisture retrievals from remote sensing: Current status, products & future trends, *Phys. Chem. Earth, 83–84*, 36–56, doi: 10.1016/j.pce.2015.02.009.

Piao, S., S. Sitch, P. Ciais, P. Friedlingstein, P. Peylin, X. Wang, A. Ahlström, A. Anav, J. G. Canadell, N. Cong, C. Huntingford, M. Jung, S. Levis, P. E. Levy, J. Li, X. Lin, M. R. Lomas, M. Lu, Y. Luo, Y. Ma, R. B. Myneni, B. Poulter, Z. Sun, T. Wang, N. Viovy, S. Zaehle, and N. Zeng (2013), Evaluation of terrestrial carbon cycle models for their response to climate variability and to CO_2 trends, *Glob. Change Biol., 19*(7), 2117–2132, doi:10.1111/gcb.12187.

Piao, S. L., A. Ito, S. G. Li, Y. Huang, P. Ciais, X. H. Wang, S. S. Peng, H. J. Nan, C. Zhao, A. Ahlström, R. J. Andres, F. Chevallier, J. Y. Fang, J. Hartmann, C. Huntingford, S. Jeong, S. Levis, P. E. Levy, J. S. Li, M. R. Lomas, J. F. Mao, E. Mayorga, A. Mohammat, H. Muraoka, C. H. Peng, P. Peylin, B. Poulter, Z. H. Shen, X. Shi, S. Sitch, S. Tao, H. Q. Tian, X. P. Wu, M. Xu, G. R. Yu, N. Viovy, S. Zaehle, N. Zeng, and B. Zhu (2012), The carbon budget of terrestrial ecosystems in East Asia over the last two decades, *Biogeosci., 9*(9), 3571–3586, doi:10.5194/bg-9-3571-2012.

Post, W. M., W. R. Emanuel, P. J. Zinke, and A. G. Stangenberger (1982), Soil carbon pools and world life zones, *Nature, 298*(5870), 156–159, doi:10.1038/298156a0.

Post, W. M., J. Pastor, P. J. Zinke, and A. G. Stangenberger (1985), Global patterns of soil nitrogen storage, *Nature, 317*(6038), 613–616, doi:10.1038/317613a0.

Poulter, B., P. Ciais, E. Hodson, H. Lischke, F. Maignan, S. Plummer, and N. E. Zimmermann (2011), Plant functional type mapping for Earth system models, *Geosci. Model Dev., 4*(4), 993–1010, doi:10.5194/gmd4-993-2011.

Powell, T. L., D. R. Galbraith, B. O. Christoffersen, A. Harper, H. M. A. Imbuzeiro, L. Rowland, S. Almeida, P. M. Brando, A. C. L. da Costa, M. H. Costa, N. M. Levine, Y. Malhi, S. R. Saleska, E. Sotta, M. Williams, P. Meir, and P. R. Moorcroft (2013), Confronting model predictions of carbon fluxes with measurements of Amazon forests subjected to experimental drought, *New Phytol., 200*(2), 350–365, doi: 10.1111/nph.12390.

Prigent, C., F. Papa, F. Aires, W. B. Rossow, and E. Matthews (2007), Global inundation dynamics inferred from multiple satellite observations, 1993–2000, *J. Geophys. Res. Atmos., 112*(D12), doi: 10.1029/2006JD007847.

Prihodko, L., A. S. Denning, N. P. Hanan, I. Baker, and K. Davis (2008), Sensitivity, uncertainty and time dependence of parameters in a complex land surface model, *Agr. Forest Meteorol., 148*(2), 268–287, doi:10.1016/j.agrformet.2007.08.006.

Pulliainen, J. (2006), Mapping of snow water equivalent and snow depth in boreal and sub-arctic zones by assimilating space-borne microwave radiometer data and ground-based observations, *Remote Sens. Environ., 101*(2), 257–269, doi:10.1016/j.rse.2006.01.002.

Purves, D., and S. Pacala (2008), Predictive models of forest dynamics, *Science, 320*(5882), 1452–1453, doi:10.1126/science.1155359.

Qu, X., and A. Hall (2006), Assessing snow albedo feedback in simulated climate change, *J. Clim., 19*(11), 2617–2630, doi:10.1175/JCLI3750.1.

Qu, X., and A. Hall (2007), What controls the strength of the snow-albedo feedback?, *J. Clim., 20*(15), 3971–3981, doi:10.1175/JCLI4186.1.

Quiquampoix, H., and R. G. Burns (2007), Interactions between proteins and soil mineral surfaces: Environmental and health consequences, *Elements, 3*(6), 401–406, doi:10.2113/GSELEMENTS.3.6.401.

Rafique, R., J. Xia, O. Hararuk, G. Leng, G. Asrar, and Y. Luo (2016), Comparing the performance of three land models in global C cycle simulations: A detailed structural analysis, *Land Degrad. Dev.*, doi: 10.1002/ldr.2506.

Randerson, J. T., F. M. Hoffman, P. E. Thornton, N. M. Mahowald, K. Lindsay, Y.-H. Lee, C. D. Nevison, S. C. Doney, G.Bonan, R. Stöckli, C. Covey, S. W. Running, and I. Y. Fung (2009), Systematic assessment of terrestrial biogeochemistry in coupled climate–carbon models, *Glob. Change Biol., 15*(9), 2462–2484, doi:10.1111/j.1365-2486.2009.01912.x.

Rasmussen, M., A. Hastings, M. J. Smith, F. B. Agusto, B. M. Chen-Charpentier, F. M. Hoffman, J. Jiang, K. E. O. Todd-Brown, Y. Wang, Y.-P. Wang, and Y. Luo (2016), Transit times and mean ages for nonautonomous and autonomous compartmental systems, *J. Math. Biol.*, pp. 1–20, doi:10.1007/s00285-0160990-8.

Raupach, M. R., P. J. Rayner, D. J. Barrett, R. S. DeFries, M. Heimann, D. S. Ojima, S. Quegan, and C. C. Schmullius (2005), Model–data synthesis in terrestrial carbon observation: Methods, data requirements and data uncertainty specifications, *Glob. Change Biol., 11*(3), 378–397, doi:10.1111/j.13652486.2005.00917.x.

Ray, J., Z. Hou, M. Huang, K. Sargsyan, and L. Swiler (2015), Bayesian calibration of the Community Land Model using surrogates, *SIAM/ASA J. Uncertain. Quantif., 3*(1), 199–233, doi:10.1137/140957998.

Reed, B. C., M. D. Schwartz, and X. Xiao (2009), Remote sensing phenology: Status and the way forward, in *Phenology of Ecosystem Processes: Applications in Global Change Research*, edited by A. Noormets, pp. 231–246, Springer New York, New York, NY, doi:10.1007/978-1-4419-0026-5 10.

Regis, R. G., and C. A. Shoemaker (2007), Improved strategies for radial basis function methods for global optimization, *J. Glob. Optim., 37*(1), 113–135, doi:10.1007/s10898-006-9040-1.

Reich, P. B., R. L. Rich, X. Lu, Y.-P. Wang, and J. Oleksyn (2014), Biogeographic variation in evergreen conifer needle longevity and impacts on boreal forest carbon cycle projections, *Proc. Nat. Acad. Sci., 111*(38), 13,703–13,708, doi:10.1073/pnas.1216054110.

Reichman, O. J., M. B. Jones, and M. P. Schildhauer (2011), Challenges and opportunities of open data in ecology, *Science, 331*(6018), 703–705, doi:10.1126/science.1197962.

Reichstein, M., M. Bahn, P. Ciais, D. Frank, M. D. Mahecha, S. I. Seneviratne, J. Zscheischler, C. Beer, N. Buchmann, D. C. Frank, D. Papale, A. Rammig, P. Smith, K. Thonicke, M. van der Velde, S. Vicca, A. Walz, and M. Wattenbach (2013), Climate extremes and the carbon cycle, *Nature, 500*(7462), 287–295, doi:10.1038/nature12350.

Reichstein, M., M. Bahn, M. D. Mahecha, J. Kattge, and D. D. Baldocchi (2014), Linking plant and ecosystem functional biogeography, *Proc. Nat. Acad. Sci., 111*(38), 13,697–13,702, doi: 10.1073/pnas.1216065111.

Ren, H., Z. Hou, M. Huang, J. Bao, Y. Sun, T. Tesfa, and L. R. Leung (2016), Classification of hydrological parameter sensitivity and evaluation of parameter transferability across 431 US MOPEX basins, *J. Hydrol., 536*, 92–108, doi:10.1016/j.jhydrol.2016.02.042.

Ricciuto, D. M., A. W. King, D. Dragoni, and W. M. Post (2011), Parameter and prediction uncertainty in an optimized terrestrial carbon cycle model: Effects of constraining variables and data record length, *J. Geophys. Res. Biogeosci., 116*(G1), doi:10.1029/2010JG001400.

Richter Jr., D., and R. A. Houghton (2011), Gross CO_2 fluxes from land-use change: Implications for reducing global emissions and increasing sinks, *Carbon Manag., 2*(1), 41–47, doi:10.4155/cmt.10.43.

Riley, W. J., F. Maggi, M. Kleber, M. S. Torn, J. Y. Tang, D. Dwivedi, and N. Guerry (2014), Long residence times of rapidly decomposable soil organic matter: Application of a multi-phase, multi-component, and vertically resolved model (BAMS1) to soil carbon dynamics, *Geosci. Model Dev., 7*(4), 1335–1355, doi: 10.5194/gmd-7-1335-2014.

Robinson, D. A., K. F. Dewey, and R. R. Heim Jr. (1993), Global snow cover monitoring: An update, *Bull. Am. Meteorol. Soc., 74*(9), 1689–1696, doi:10.1175/15200477(1993)074%3C1689:GSCMAU%3E2.0.CO;2.

Rougier, J., D. M. H. Sexton, J. M. Murphy, and D. Stainforth (2009), Analyzing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments, *J. Clim., 22*(13), 3540–3557, doi:10.1175/2008JCLI2533.1.

Running, S. W., R. R. Nemani, F. A. Heinsch, M. Zhao, M. Reeves, and H. Hashimoto (2004), A continuous satellite-derived measure of global terrestrial primary production, *Bioscience, 54*(6), 547–560, doi:10.1641/0006-3568(2004)054[0547:ACSMOG]2.0.CO;2.

Ryu, Y., D. D. Baldocchi, H. Kobayashi, C. van Ingen, J. Li, T. A. Black, J. Beringer, E. van Gorsel, A. Knohl, B. E. Law, and O. Roupsard (2011), Integration of MODIS land and atmosphere products with a coupled-process model to estimate gross primary productivity and evapotranspiration from 1 km to global scales, *Global Biogeochem. Cycles, 25*(4), doi:10.1029/2011GB004053.

Saatchi, S. S., N. L. Harris, S. Brown, M. Lefsky, E. T. A. Mitchard, W. Salas, B. R. Zutta, W. Buermann, S. L. Lewis, S. Hagen, S. Petrova, L. White, M. Silman, and A. Morel (2011), Benchmark map of forest carbon stocks in tropical regions across three continents, *Proc. Nat. Acad. Sci., 108*(24), 9899–9904, doi:10.1073/pnas.1019576108.

Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto (2004), *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*, John Wiley & Sons.

Saltelli, A., M. Ratto, S. Tarantola, and F. Campolongo (2006), Sensitivity analysis practices: Strategies for model-based inference, *Reliab. Eng. Syst. Safe., 91*(10–11), 1109–1125, doi:10.1016/j.ress.2005.11.014.

Sargsyan, K., C. Safta, H. N. Najm, B. J. Debusschere, D. Ricciuto, and P. Thornton (2014), Dimensionality reduction for complex models via Bayesian compressive sensing, *Int. J. Uncertain. Quantif., 4*(1), 63–93, doi:10.1615/Int.J.UncertaintyQuantification.2013006821.

Schaaf, C. B., F. Gao, A. H. Strahler, W. Lucht, X. Li, T. Tsang, N. C. Strugnell, X. Zhang, Y. Jin, J.P. Muller, P. Lewis, M. Barnsley, P. Hobson, M. Disney, G. Roberts, M. Dunderdale, C. Doll, R. P. d'Entremont, B. Hu, S. Liang, J. L. Privette, and D. Roy (2002), First operational BRDF, albedo nadir reflectance products from MODIS, *Remote Sens. Environ., 83*(1–2), 135–148, doi:10.1016/S00344257(02)00091-3.

Schädel, C., Y. Luo, R. D. Evans, S. Fei, and S. M. Schaeffer (2013), Separating soil CO_2 efflux into C-pool-specific decay rates via inverse analysis of soil incubation data, *Oecologia, 171*(3), 721–732, doi:10.1007/s00442-012-2577-4.

Schädel, C., E. A. G. Schuur, R. Bracho, B. Elberling, C. Knoblauch, H. Lee, Y. Luo, G. R. Shaver, and M. R. Turetsky (2014), Circumpolar assessment of permafrost C quality and its vulnerability over time using long-term incubation data, *Glob. Change Biol., 20*(2), 641–652, doi:10.1111/gcb.12417.

Schädel, C., M. K.-F. Bader, E. A. G. Schuur, C. Biasi, R. Bracho, P. Capek, S. De Baets, K. Diakova, J. Ernakovich, C. Estop-Aragones, D. E. Graham, I. P. Hartley, C. M. Iversen, E. Kane, C. Knoblauch, M. Lupascu, P. J. Martikainen, S. M. Natali, R. J. Norby, J. A. O'Donnell, T. R. Chowdhury, H. Santruckova, G. Shaver, V. L. Sloan, C. C. Treat, M. R. Turetsky, M. P. Waldrop, and K. P. Wickland (2016), Potential carbon emissions dominated by carbon dioxide from thawed permafrost soils, *Nature Clim. Change, advance online publication*, doi:10.1038/nclimate3054.

Schaefer, K., and E. Jafarov (2016), A parameterization of respiration in frozen soils based on substrate availability, *Biogeosci., 13*(7), 1991–2001, doi:10.5194/bg-13-1991-2016.

Schaefer, K., T. Zhang, L. Bruhwiler, and A. P. Barrett (2011), Amount and timing of permafrost carbon release in response to climate warming, *Tellus B, 63*(2), 165–180, doi:10.1111/j.1600-0889.2011.00527.x.

Scharlemann, J. P., E. V. Tanner, R. Hiederer, and V. Kapos (2014), Global soil carbon: Understanding and managing the largest terrestrial carbon pool, *Carbon Manag., 5*(1), 81–91, doi:10.4155/cmt.13.77.

Scheiter, S., L. Langan, and S. I. Higgins (2013), Next-generation dynamic global vegetation models: Learning from community ecology, *New Phytol., 198*(3), 957–969, doi:10.1111/nph.12210.

Schimel, D., R. Pavlick, J. B. Fisher, G. P. Asner, S. Saatchi, P. Townsend, C. Miller, C. Frankenberg, K. Hibbard, and P. Cox (2015), Observing terrestrial ecosystems and the carbon cycle from space, *Glob. Change Biol., 21*(5), 1762–1776, doi:10.1111/gcb.12822.

Schmid, B., S. Serbin, A. Vogelmann, G. de Boer, B. Dafflon, A. Guenther, and D. Moore (2015), Aerial observaton needs workshop (May 13–14, 2015), *Technical Report DOE/SC-0179*, U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Germantown, Maryland, USA.

Schmidt, M. W. I., M. S. Torn, S. Abiven, T. Dittmar, G. Guggenberger, I. A. Janssens, M. Kleber, I. Kogel-Knabner, J. Lehmann, D. A. C. Manning, P. Nannipieri, D. P. Rasse, S. Weiner, and S. E. Trumbore (2011), Persistence of soil organic matter as an ecosystem property, *Nature, 478*(7367), 49–56, doi: 10.1038/nature10386.

Schultz, N. M., X. Lee, P. J. Lawrence, D. M. Lawrence, and L. Zhao (2016), Assessing the use of subgrid land model output to study impacts of land cover change, *J. Geophys. Res. Atmos., 121*(11), 6133–6147, doi:10.1002/2016JD025094. Schuur, E. A. G., A. D. McGuire, C. Schadel, G. Grosse, J. W. Harden, D. J. Hayes, G. Hugelius, C. D. Koven, P. Kuhry, D. M. Lawrence, S. M. Natali, D. Olefeldt, V. E. Romanovsky, K. Schaefer, M. R. Turetsky, C. C. Treat, and J. E. Vonk (2015), Climate change and the permafrost carbon feedback, *Nature, 520*(7546), 171–179, doi:10.1038/nature14338.

Schwalm, C. R., D. N. Huntzinger, J. B. Fisher, A. M. Michalak, K. Bowman, P. Ciais, R. Cook, B. El-Masri, D. Hayes, M. Huang, A. Ito, A. Jain, A. W. King, H. Lei, J. Liu, C. Lu, J. Mao, S. Peng, B. Poulter, D. Ricciuto, K. Schaefer, X. Shi, B. Tao, H. Tian, W. Wang, Y. Wei, J. Yang, and N. Zeng (2015), Toward "optimal" integration of terrestrial biosphere models, *Geophys. Res. Lett., 42*(11), 4418–4428, doi:10.1002/2015GL064002.

Schwartz, M. D., J. L. Betancourt, and J. F. Weltzin (2012), From Caprio's lilacs to the USA National Phenology Network, *Front. Ecol. Environ., 10*(6), 324–327, doi:10.1890/110281.

Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling (2010), Investigating soil moisture–climate interactions in a changing climate: A review, *Earth-Sci. Rev., 99*(3–4), 125–161, doi:10.1016/j.earscirev.2010.02.004.

Seneviratne, S. I., N. Nicholls, D. Easterling, C. M. Goodess, S. Kanae, J. Kossin, Y. Luo, J. Marengo, K. McInnes, M. Rahimi, M. Reichstein, A. Sorteberg, C. Vera, and X. Zhang (2012), Changes in climate extremes and their impacts on the natural physical environment, in *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change*, edited by C. B. Field, V. Barros, T. F. Stocker, D. Qin, D. J. Dokken, K. L. Ebi, M. D. Mastrandrea, K. J. Mach, G.-K. Plattner, S. K. Allen, M. Tignor, and P. M. Midgley, pp. 109–230, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Seneviratne, S. I., M. Wilhelm, T. Stanelle, B. van den Hurk, S. Hagemann, A. Berg, F. Cheruy, M. E. Higgins, A. Meier, V. Brovkin, M. Claussen, A. Ducharne, J.-L. Dufresne, K. L. Findell, J. Ghattas, D. M. Lawrence, S. Malyshev, M. Rummukainen, and B. Smith (2013), Impact of soil moisture-climate feedbacks on CMIP5 projections: First results from the GLACE-CMIP5 experiment, *Geophys. Res. Lett., 40*(19), 5212–5217, doi:10.1002/grl.50956.

Serbin, S. P., A. Singh, A. R. Desai, S. G. Dubois, A. D. Jablonski, C. C. Kingdon, E. L. Kruger, and P. A. Townsend (2015), Remotely estimating photosynthetic capacity, and its response to temperature, in vegetation canopies using imaging spectroscopy, *Remote Sens. Environ., 167*, 78–87, doi: 10.1016/j.rse.2015.05.024.

Shao, P., X. Zeng, D. J. P. Moore, and X. Zeng (2013), Soil microbial respiration from observations and Earth system models, *Environ. Res. Lett., 8*(3), 034,034, doi:10.1088/1748-9326/8/3/034034.

Sheffield, J., G. Goteti, and E. F. Wood (2006), Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling, *J. Clim., 19*(13), 3088–3111, doi:10.1175/JCLI3790.1.

Shiklomanov, A. N., M. C. Dietze, T. Viskari, P. A. Townsend, and S. P. Serbin (2016), Quantifying the influences of spectral resolution on uncertainty in leaf trait estimates through a Bayesian approach to RTM inversion, *Remote Sens. Environ., 183*, 226–238, doi:10.1016/j.rse.2016.05.023.

Shugart, H. H., G. P. Asner, R. Fischer, A. Huth, N. Knapp, T. Le Toan, and J. K. Shuman (2015), Computer and remote-sensing infrastructure to enhance large-scale testing of individual-based forest models, *Front. Ecol. Environ., 13*(9), 503–511, doi:10.1890/140327.

Sierra, C. A., and M. Müller (2015), A general mathematical framework for representing soil organic matter dynamics, *Ecol. Monogr., 85*(4), 505–524, doi:10.1890/15-0361.1.

Singh, A., S. P. Serbin, B. E. McNeil, C. C. Kingdon, and P. A. Townsend (2015), Imaging spectroscopy algorithms for mapping canopy foliar chemical and morphological traits and their uncertainties, *Ecol. Appl., 25*(8), 2180–2197, doi:10.1890/14-2098.1.

Sitch, S., P. Friedlingstein, N. Gruber, S. D. Jones, G. Murray-Tortarolo, A. Ahlström, S. C. Doney, H. Graven, C. Heinze, C. Huntingford, S. Levis, P. E. Levy, M. Lomas, B. Poulter, N. Viovy, S. Zaehle, N. Zeng, A. Arneth, G. Bonan, L. Bopp, J. G. Canadell, F. Chevallier, P. Ciais, R. Ellis, M. Gloor, P. Peylin, S. L. Piao, C. Le Quéré, B. Smith, Z. Zhu, and R. Myneni (2015), Recent trends and drivers of regional sources and sinks of carbon dioxide, *Biogeosci., 12*(3), 653–679, doi:10.5194/bg-12-653-2015.

Slater, A. G., and D. M. Lawrence (2013), Diagnosing present and future permafrost from climate models, *J. Clim., 26*, 5608–5623, doi:10.1175/JCLI-D-12-00341.1.

Slater, A. G., A. P. Barrett, M. P. Clark, J. D. Lundquist, and M. S. Raleigh (2013), Uncertainty in seasonal snow reconstruction: Relative impacts of model forcing and image availability, *Adv. Water Resour., 55*, 165–177, doi:10.1016/j.advwatres.2012.07.006.

Slater, A. G., D. M. Lawrence, and C. D. Koven (2016), Process-level model evaluation: A snow and heat transfer metric, submitted.

Smith, N. G., V. L. Rodgers, E. R. Brzostek, A. Kulmatiski, M. L. Avolio, D. L. Hoover, S. E. Koerner, K. Grant, A. Jentsch, S. Fatichi, and D. Niyogi (2014), Toward a better integration of biological data from precipitation manipulation experiments into Earth system models, *Rev. Geophys., 52*(3), 412–434, doi:10.1002/2014RG000458.

Smith, P., J. I. House, M. Bustamante, J. Sobocká, R. Harper, G. Pan, P. C. West, J. M. Clark, T. Adhya, C. Rumpel, K. Paustian, P. Kuikman, M. F. Cotrufo, J. A. Elliott, R. McDowell, R. I. Griffiths, S. Asakawa, A. Bondeau, A. K. Jain, J. Meersmans, and T. A. M. Pugh (2016a), Global change pressures on soils from land use and management, *Glob. Change Biol., 22*(3), 1008–1028, doi:10.1111/gcb.13068.

Smith, W. K., S. C. Reed, C. C. Cleveland, A. P. Ballantyne, W. R. L. Anderegg, W. R. Wieder, Y. Y. Liu, and S. W. Running (2016b), Large divergence of satellite and Earth system model estimates of global terrestrial CO_2 fertilization, *Nature Clim. Change, 6*(3), 306–310, doi:10.1038/nclimate2879.

Sobol, I. M. (1993), Sensitivty estimates for nonlinear mathematical models, *Math. Modeling and Comput. Exper., 1*(4), 407–414.

Solonen, A., P. Ollinaho, M. Laine, H. Haario, J. Tamminen, and H. Järvinen (2012), Efficient MCMC for climate model parameter estimation: Parallel adaptive chains and early rejection, *Bayesian Anal., 7*(3), 715–736, doi:10.1214/12-BA724.

Stöckli, R., T. Rutishauser, D. Dragoni, J. O'Keefe, P. E. Thornton, M. Jolly, L. Lu, and A. S. Denning (2008), Remote sensing data assimilation for a prognostic phenology model, *J. Geophys. Res. Biogeosci., 113*(G4), doi:10.1029/2008JG000781.

Stöckli, R., T. Rutishauser, I. Baker, M. A. Liniger, and A. S. Denning (2011), A global reanalysis of vegetation phenology, *J. Geophys. Res. Biogeosci., 116*(G3), doi:10.1029/2010JG001545.

Sulman, B. N., R. P. Phillips, A. C. Oishi, E. Shevliakova, and S. W. Pacala (2014), Microbe-driven turnover offsets mineral-mediated storage of soil carbon under elevated CO_2, *Nature Clim. Change, 4*(12), 1099– 1102, doi:10.1038/nclimate2436.

Sun, Y., Z. Hou, M. Huang, F. Tian, and L. R. Leung (2013), Inverse modeling of hydrologic parameters using surface flux and runoff observations in the Community Land Model, *Hydrol. Earth Syst. Sci., 17*(12), 4995–5011, doi:10.5194/hess-17-4995-2013.

Sundareshwar, P. V., R. Murtugudde, G. Srinivasan, S. Singh, K. J. Ramesh, R. Ramesh, S. B. Verma, D. Agarwal, D. Baldocchi, C. K. Baru, K. K. Baruah, G. R. Chowdhury, V. K. Dadhwal, C. B. S. Dutt, J. Fuentes, P. K. Gupta, W. W. Hargrove, M. Howard, C. S. Jha, S. Lal, W. K. Michener, A. P. Mitra, J. T. Morris, R. R. Myneni, M. Naja, R. Nemani, R. Purvaja, S. Raha, S. K. S. Vanan, M. Sharma, A. Subramaniam, R. Sukumar, R. R. Twilley, and P. R. Zimmerman (2007), Environmental monitoring network for India, *Science, 316*(5822), 204–205, doi:10.1126/science.1137417.

Swenson, S. C., and D. M. Lawrence (2015), A GRACE-based assessment of interannual groundwater dynamics in the Community Land Model, *Water Resour. Res., 51*(11), 8817–8833, doi: 10.1002/2015WR017582.

Swenson, S. C., D. M. Lawrence, and H. Lee (2012), Improved simulation of the terrestrial hydrological cycle in permafrost regions by the Community Land Model, *J. Adv. Model. Earth Syst., 4*(3), doi: 10.1029/2012MS000165.

Swiler, L., J. Ray, M. Huang, and J. Hou (2015), Use of parallel MCMC methods with the Community Land Model, Abstract in the 2015 SIAM Symposium on Computational Science and Engineering Conference (March 14–18, 2015) Proceedings, Salt Lake City, Utah, USA.

Tang, J., 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, *Biogeosci., 10*(12), 8329– 8351, doi:10.5194/bg-10-8329-2013.

Tang, J., and W. J. Riley (2015), Weaker soil carbon–climate feedbacks resulting from microbial and abiotic interactions, *Nature Clim. Change, 5*(1), 56–60, doi:10.1038/nclimate2438.

Tang, J. Y., W. J. Riley, C. D. Koven, and Z. M. Subin (2013), CLM4-BeTR, a generic biogeochemical transport and reaction module for CLM4: Model development, evaluation, and application, *Geosci. Model Dev., 6*(1), 127–140, doi:10.5194/gmd-6-127-2013.

Tao, B., H. Tian, G. Chen, W. Ren, C. Lu, K. D. Alley, X. Xu, M. Liu, S. Pan, and H. Virji (2013), Terrestrial carbon balance in tropical Asia: Contribution from cropland expansion and land management, *Glob. Planet. Change, 100*, 85–98, doi:10.1016/j.gloplacha.2012.09.006.

Teixeira, J., D. Waliser, R. Ferraro, P. Gleckler, T. Lee, and G. Potter (2014), Satellite observations for CMIP5: The genesis of Obs4MIPs, *Bull. Am. Meteorol. Soc., 95*(9), 1329–1334, doi:10.1175/BAMS-D12-00204.1.

Teuling, A. J., S. I. Seneviratne, R. Stöckli, M. Reichstein, E. Moors, P. Ciais, S. Luyssaert, B. van den Hurk, C. Ammann, C. Bernhofer, E. Dellwik, D. Gianelle, B. Gielen, T. Grunwald, K. Klumpp, L. Montagnani, C. Moureaux, M. Sottocornola, and G. Wohlfahrt (2010), Contrasting response of European forest and grassland energy exchange to heatwaves, *Nat. Geosci., 3*(10), 722–727, doi:10.1038/ngeo950.

Thackeray, C. W., C. G. Fletcher, and C. Derksen (2015), Quantifying the skill of CMIP5 models in simulating seasonal albedo and snow cover evolution, *J. Geophys. Res. Atmos., 120*(12), 5831–5849, doi: 10.1002/2015JD023325.

Thomas, R. Q., and M. Williams (2014), A model using marginal efficiency of investment to analyze carbon and nitrogen interactions in terrestrial ecosystems (ACONITE Version 1), *Geosci. Model Dev., 7*(5), 2015–2037, doi:10.5194/gmd-7-2015-2014.

Thomas, R. Q., S. Zaehle, P. H. Templer, and C. L. Goodale (2013), Global patterns of nitrogen limitation: confronting two global biogeochemical models with observations, *Glob. Change Biol., 19*(10), 2986– 2998, doi:10.1111/gcb.12281.

Thornton, P. E., J. F. Lamarque, N. A. Rosenbloom, and N. M. Mahowald (2007), Influence of carbon– nitrogen cycle coupling on land model response to CO_2 fertilization and climate variability, *Global Biogeochem. Cycles, 21*(4), GB4018, doi:10.1029/2006GB002868.

Thurner, M., C. Beer, N. Carvalhais, M. Forkel, M. Santoro, M. Tum, and C. Schmullius (2016), Large-scale variation in boreal and temperate forest carbon turnover rate related to climate, *Geophys. Res. Lett., 43*(9), 4576–4585, doi:10.1002/2016GL068794.

Tian, H., C. Lu, J. Yang, K. Banger, D. N. Huntzinger, C. R. Schwalm, A. M. Michalak, R. Cook, P. Ciais, D. Hayes, M. Huang, A. Ito, A. K. Jain, H. Lei, J. Mao, S. Pan, W. M. Post, S. Peng, B. Poulter, W. Ren, D. Ricciuto, K. Schaefer, X. Shi, B. Tao, W. Wang, Y. Wei, Q. Yang, B. Zhang, and N. Zeng (2015), Global patterns and controls of soil organic carbon dynamics as simulated by multiple terrestrial biosphere models: Current status and future directions, *Global Biogeochem. Cycles, 29*(6), 775–792, doi: 10.1002/2014GB005021.

Tian, H., C. Lu, P. Ciais, A. M. Michalak, J. G. Canadell, E. Saikawa, D. N. Huntzinger, K. R. Gurney, S. Sitch, B. Zhang, J. Yang, P. Bousquet, L. Bruhwiler, G. Chen, E. Dlugokencky, P. Friedlingstein, J. Melillo, S. Pan, B. Poulter, R. Prinn, M. Saunois, C. R. Schwalm, and S. C. Wofsy (2016), The terrestrial biosphere as a net source of greenhouse gases to the atmosphere, *Nature, 531*(7593), 225–228, doi:10.1038/nature16946.

Tian, X., and Z. Xie (2008), A land surface soil moisture data assimilation framework in consideration of the model subgrid-scale heterogeneity and soil water thawing and freezing, *Sci. China Ser. D-Earth Sci., 51*(7), 992–1000, doi:10.1007/s11430-008-0069-5.

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, *Biogeosci., 10*(3), 1717–1736, doi:10.5194/bg-10-1717-2013.

Torn, M. S., S. E. Trumbore, O. A. Chadwick, P. M. Vitousek, and D. M. Hendricks (1997), Mineral control of soil organic carbon storage and turnover, *Nature, 389*(6647), 170–173, doi:10.1038/38260.

Toure, A. M., M. Rodell, Z.-L. Yang, H. Beaudoing, E. Kim, Y. Zhang, and Y. Kwon (2016), Evaluation of the snow simulations from the Community Land Model, version 4 (CLM4), *J. Hydrometeor., 17*(1), 153–170, doi:10.1175/JHM-D-14-0165.1.

Trumbore, S., E. S. D. Costa, D. C. Nepstadt, P. B. D. Camargo, L. A. Martinelli, D. Ray, T. Restom, and W. Silver (2006), Dynamics of fine root carbon in Amazonian tropical ecosystems and the contribution of roots to soil respiration, *Glob. Change Biol., 12*(2), 217–229, doi:10.1111/j.1365-2486.2005.001063.x.

Turner, D. P., W. D. Ritts, W. B. Cohen, S. T. Gower, M. Zhao, S. W. Running, S. C. Wofsy, S. Urbanski, A. L. Dunn, and J. W. Munger (2003), Scaling Gross Primary Production (GPP) over boreal and deciduous forest landscapes in support of MODIS GPP product validation, *Remote Sens. Environ., 88*(3), 256–270, doi:10.1016/j.rse.2003.06.005.

Ustin, S. L., D. A. Roberts, J. A. Gamon, G. P. Asner, and R. O. Green (2004), Using imaging spectroscopy to study ecosystem processes and properties, *Bioscience, 54*(6), 523–534, doi:10.1641/00063568(2004)054[0523:UISTSE]2.0.CO;2.

Utsumi, N., H. Kim, S. Seto, S. Kanae, and T. Oki (2014), Climatological characteristics of fronts in the western North Pacific based on surface weather charts, *J. Geophys. Res. Atmos., 119*(15), 9400–9418, doi:10.1002/2014JD021734.

Utsumi, N., H. Kim, S. Kanae, and T. Oki (2016), Which weather systems are projected to cause future changes in mean and extreme precipitation in CMIP5 simulations?, in revision.

van den Hurk, B., H. Kim, G. Krinner, S. I. Seneviratne, C. Derksen, T. Oki, H. Douville, J. Colin, A. Ducharne, F. Cheruy, N. Viovy, M. J. Puma, Y. Wada, W. Li, B. Jia, A. Alessandri, D. M. Lawrence, G. P. Weedon, R. Ellis, S. Hagemann, J. Mao, M. G. Flanner, M. Zampieri, S. Materia, R. M. Law, and J. Sheffield (2016), LS3MIP (v1.0) contribution to CMIP6: The Land Surface, Snow and Soil moisture Model Intercomparison Project – Aims, setup and expected outcome, *Geosci. Model Dev., 9*(8), 2809– 2832, doi:10.5194/gmd-9-2809-2016.

van der Werf, G. R., J. T. Randerson, L. Giglio, G. J. Collatz, M. Mu, P. S. Kasibhatla, D. C. Morton, R. S. DeFries, Y. Jin, and T. T. van Leeuwen (2010), Global fire emissions and the contribution of deforestation, savanna, forest, agricultural, and peat fires (1997–2009), *Atmos. Chem. Phys., 10*(23), 11,707–11,735, doi:10.5194/acp-10-11707-2010.

van Mantgem, P. J., N. L. Stephenson, J. C. Byrne, L. D. Daniels, J. F. Franklin, P. Z. Fulé, M. E. Harmon, A. J. Larson, J. M. Smith, A. H. Taylor, and T. T. Veblen (2009), Widespread increase of tree mortality rates in the western United States, *Science, 323*(5913), 521–524, doi:10.1126/science.1165000.

Verheijen, L. M., V. Brovkin, R. Aerts, G. Bönisch, J. H. C. Cornelissen, J. Kattge, P. B. Reich, I. J. Wright, and P. M. van Bodegom (2013), Impacts of trait variation through observed trait–climate relationships on performance of an Earth system model: A conceptual analysis, *Biogeosci., 10*(8), 5497–5515, doi: 10.5194/bg-10-5497-2013.

Viovy, N., and P. Ciais (2011), CRUNCEP data set for 1901–2008, *Tech. Rep. Version 4*, Laboratoire des Sciences du Climat et de l'Environnement, Gif Sur Yvette, France.

Viskari, T., B. Hardiman, A. R. Desai, and M. C. Dietze (2015), Model-data assimilation of multiple phenological observations to constrain and predict leaf area index, *Ecol. Appl., 25*(2), 546–558, doi:10.1890/140497.1.

Viskari, T., A. Shiklomanov, M. C. Dietze, P. A. Townsend, and S. P. Serbin (2016), Quantifying uncertainties in modeled canopy fluxes due to canopy radiative transfer, *Agr. Forest Meteorol.*, in preparation.

Walker, A. P., P. J. Hanson, M. G. De Kauwe, B. E. Medlyn, S. Zaehle, S. Asao, M. Dietze, T. Hickler, C. Huntingford, C. M. Iversen, A. Jain, M. Lomas, Y. Luo, H. McCarthy, W. J. Parton, I. C. Prentice, P. E. Thornton, S. Wang, Y.-P. Wang, D. Warlind, E. Weng, J. M. Warren, F. I. Woodward, R. Oren, and R. J. Norby (2014), Comprehensive ecosystem model–data synthesis using multiple data sets at two temperate forest free-air CO_2 enrichment experiments: Model performance at ambient CO_2 concentration, *J. Geophys. Res. Biogeosci., 119*(5), 2169–8961, doi:10.1002/2013JG002553.

Walker, A. P., S. Zaehle, B. E. Medlyn, M. G. De Kauwe, S. Asao, T. Hickler, W. Parton, D. M. Ricciuto, Y.-P. Wang, D. Wårlind, and R. J. Norby (2015), Predicting long-term carbon sequestration in response to CO_2 enrichment: How and why do current ecosystem models differ?, *Global Biogeochem. Cycles, 29*(4), doi:10.1002/2014GB004995.

Walton, D. B., F. Sun, A. Hall, and S. Capps (2015), A hybrid dynamical–statistical downscaling technique. Part I: Development and validation of the technique, *J. Clim., 28*(12), 4597–4617, doi:10.1175/JCLI-D14-00196.1.

Wang, Y. P., R. M. Law, and B. Pak (2010), A global model of carbon, nitrogen and phosphorus cycles for the terrestrial biosphere, *Biogeosci., 7*(7), 2261–2282, doi:10.5194/bg-7-2261-2010.

Wang, Y. P., B. C. Chen, W. R. Wieder, M. Leite, B. E. Medlyn, M. Rasmussen, M. J. Smith, F. B. Agusto, F. M. Hoffman, and Y. Q. Luo (2014), Oscillatory behavior of two nonlinear microbial models of soil carbon decomposition, *Biogeosci., 11*(7), 1817–1831, doi:10.5194/bg-11-1817-2014.

Wang, Y. P., J. Jiang, B. Chen-Charpentier, F. B. Agusto, A. Hastings, F. M. Hoffman, M. Rasmussen, M. J. Smith, K. Todd-Brown, Y. Wang, X. Xu, and Y. Q. Luo (2016), Responses of two nonlinear microbial models to warming and increased carbon input, *Biogeosci., 13*(4), 887–902, doi:10.5194/bg-13-887-2016.

Weedon, G. P., S. Gomes, P. Viterbo, W. J. Shuttleworth, E. Blyth, H. Österle, J. C. Adam, N. Bellouin, O. Boucher, and M. Best (2011), Creation of the WATCH forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century, *J. Hydrometeor., 12*(5), 823– 848, doi:10.1175/2011JHM1369.1.

Weedon, G. P., G. Balsamo, N. Bellouin, S. Gomes, M. J. Best, and P. Viterbo (2014), The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, *Water Resour. Res., 50*(9), 7505–7514, doi:10.1002/2014WR015638.

Wei, Y., S. Liu, D. N. Huntzinger, A. M. Michalak, N. Viovy, W. M. Post, C. R. Schwalm, K. Schaefer, A. R. Jacobson, C. Lu, H. Tian, D. M. Ricciuto, R. B. Cook, J. Mao, and X. Shi (2014a), The North American Carbon Program Multi-Scale Synthesis and Terrestrial Model Intercomparison Project — Part 2: Environmental driver data, *Geosci. Model Dev., 7*(6), 2875–2893, doi:10.5194/gmd-7-2875-2014.

Wei, Y., S. Liu, D. N. Huntzinger, A. M. Michalak, N. Viovy, W. M. Post, C. R. Schwalm, K. Schaefer, A. R. Jacobson, C. Lu, H. Tian, D. M. Ricciuto, R. B. Cook, J. Mao, and X. Shi (2014b), NACP MsTMIP: Global and North American driver data for multi-model intercomparison, Data set, available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA, doi:10.3334/ORNLDAAC/1220.

Weng, E. S., S. Malyshev, J. W. Lichstein, C. E. Farrior, R. Dybzinski, T. Zhang, E. Shevliakova, and S. W. Pacala (2015), Scaling from individual trees to forests in an Earth system modeling framework using a mathematically tractable model of height-structured competition, *Biogeosci., 12*(9), 2655–2694, doi:10.5194/bg-12-2655-2015.

Whitley, R., J. Beringer, L. B. Hutley, G. Abramowitz, M. G. De Kauwe, R. Duursma, B. Evans, V. Haverd, L. Li, Y. Ryu, B. Smith, Y.-P. Wang, M. Williams, and Q. Yu (2016), A model inter-comparison study to examine limiting factors in modelling Australian tropical savannas, *Biogeosci., 13*(11), 3245–3265, doi:10.5194/bg-13-3245-2016.

Wieder, W. R., G. B. Bonan, and S. D. Allison (2013), Global soil carbon projections are improved by modelling microbial processes, *Nature Clim. Change, 3*(10), 909–912, doi:10.1038/nclimate1951.

Wieder, W. R., S. D. Allison, E. A. Davidson, K. Georgiou, O. Hararuk, Y. He, F. Hopkins, Y. Luo, M. J. Smith, B. Sulman, K. Todd-Brown, Y.-P. Wang, J. Xia, and X. Xu (2015a), Explicitly representing soil microbial processes in Earth system models, *Global Biogeochem. Cycles, 29*(10), 1782–1800, doi:10.1002/2015GB005188.

Wieder, W. R., C. C. Cleveland, W. K. Smith, and K. Todd-Brown (2015b), Future productivity and carbon storage limited by terrestrial nutrient availability, *Nat. Geosci., 8*(6), 441–444, doi:10.1038/ngeo2413.

Wieder, W. R., A. S. Grandy, C. M. Kallenbach, P. G. Taylor, and G. B. Bonan (2015c), Representing life in the Earth system with soil microbial functional traits in the MIMICS model, *Geosci. Model Dev., 8*(6), 1789–1808, doi:10.5194/gmd-8-1789-2015.

Wilkinson, R. D. (2011), Bayesian calibration of expensive multivariate computer experiments, in *Large-Scale Inverse Problems and Quantification of Uncertainty*, chap. 10, pp. 195–215, John Wiley & Sons, Ltd., doi:10.1002/9780470685853.ch10.

Williams, C. A., M. Reichstein, N. Buchmann, D. Baldocchi, C. Beer, C. Schwalm, G. Wohlfahrt, N. Hasler, C. Bernhofer, T. Foken, D. Papale, S. Schymanski, and K. Schaefer (2012), Climate and vegetation controls on the surface water balance: Synthesis of evapotranspiration measured across a global network of flux towers, *Water Resour. Res., 48*(6), doi:10.1029/2011WR011586.

Williams, D. N. (2014), Visualization and analysis tools for ultrascale climate data, *Eos Trans. AGU, 95*(42), 377–378, doi:10.1002/2014EO420002.

Williams, D. N., G. Palanisamy, and K. Kleese-Van Dam (2016), Working group on virtual data integration: A report from the August 13–14, 2015, workshop, *Technical Report LLNL-TR–678127*, Lawrence Livermore National Laboratory (LLNL), Livermore, California, USA, doi:10.2172/1227017.

Williams, M., A. D. Richardson, M. Reichstein, P. C. Stoy, P. Peylin, H. Verbeeck, N. Carvalhais, M. Jung, D. Y. Hollinger, J. Kattge, R. Leuning, Y. Luo, E. Tomelleri, C. M. Trudinger, and Y.-P. Wang (2009), Improving land surface models with FLUXNET data, *Biogeosci., 6*(7), 1341–1359, doi:10.5194/bg-61341-2009.

Witze, A. (2015), Minnesota bog study turns up the heat on peat, *Nature, 524*(7566), 397, doi: 10.1038/524397a.

Wolf, A., P. Ciais, V. Bellassen, N. Delbart, C. B. Field, and J. A. Berry (2011), Forest biomass allometry in global land surface models, *Global Biogeochem. Cycles, 25*(3), doi:10.1029/2010GB003917.

Wright, S. J., O. Calderón, A. Hernandéz, and S. Paton (2004), Are lianas increasing in importance in tropical forests? A 17-year record from Panama, *Ecology, 85*(2), 484–489, doi:10.1890/02-0757.

Xia, J., Y. Luo, Y.-P. Wang, and O. Hararuk (2013), Traceable components of terrestrial carbon storage capacity in biogeochemical models, *Glob. Change Biol., 19*(7), 2104–2116, doi:10.1111/gcb.12172.

Xia, J., S. Niu, P. Ciais, I. A. Janssens, J. Chen, C. Ammann, A. Arain, P. D. Blanken, A. Cescatti, D. Bonal, N. Buchmann, P. S. Curtis, S. Chen, J. Dong, L. B. Flanagan, C. Frankenberg, T. Georgiadis, C. M. Gough, D. Hui, G. Kiely, J. Li, M. Lund, V. Magliulo, B. Marcolla, L. Merbold, L. Montagnani, E. J. Moors, J. E. Olesen, S. Piao, A. Raschi, O. Roupsard, A. E. Suyker, M. Urbaniak, F. P. Vaccari, A. Varlagin, T. Vesala, M. Wilkinson, E. Weng, G. Wohlfahrt, L. Yan, and Y. Luo (2015a), Joint control of terrestrial gross primary productivity by plant phenology and physiology, *Proc. Nat. Acad. Sci., 112*(9), 2788–2793, doi:10.1073/pnas.1413090112.

Xia, Y., M. B. Ek, Y. Wu, T. Ford, and S. M. Quiring (2015b), Comparison of NLDAS-2 simulated and NASMD observed daily soil moisture. Part I: Comparison and analysis, *J. Hydrometeor., 16*(5), 1962– 1980, doi:10.1175/JHM-D-14-0096.1.

Xiao, J., S. V. Ollinger, S. Frolking, G. C. Hurtt, D. Y. Hollinger, K. J. Davis, Y. Pan, X. Zhang, F. Deng, J. Chen, D. D. Baldocchi, B. E. Law, M. A. Arain, A. R. Desai, A. D. Richardson, G. Sun, B. Amiro, H. Margolis, L. Gu, R. L. Scott, P. D. Blanken, and A. E. Suyker (2014), Data-driven diagnostics of terrestrial carbon dynamics over North America, *Agr. Forest Meteorol., 197*, 142–157, doi: 10.1016/j.agrformet.2014.06.013.

Xu, C., R. Fisher, S. D. Wullschleger, C. J. Wilson, M. Cai, and N. G. McDowell (2012a), Toward a mechanistic modeling of nitrogen limitation on vegetation dynamics, *PLoS ONE, 7*(5), e37,914, doi: 10.1371/journal.pone.0037914.

Xu, L., and P. Dirmeyer (2011), Snow-atmosphere coupling strength in a global atmospheric model, *Geophys. Res. Lett., 38*(13), L13,401, doi:10.1029/2011GL048049.

Xu, X., Y. Luo, and J. Zhou (2012b), Carbon quality and the temperature sensitivity of soil organic carbon decomposition in a tallgrass prairie, *Soil Biology and Biochemistry, 50*, 142–148, doi: 10.1016/j.soilbio.2012.03.007.

Xu, X., W. J. Riley, C. D. Koven, D. P. Billesbach, R. Y.-W. Chang, R. Commane, E. S. Euskirchen, S. Hartery, Y. Harazono, H. Iwata, K. C. McDonald, C. E. Miller, W. C. Oechel, B. Poulter, N. Raz-Yaseef, C. Sweeney, M. Torn, S. C. Wofsy, Z. Zhang, and D. Zona (2016), A multi-scale comparison of modeled and observed seasonal methane emissions in northern wetlands, *Biogeosci., 13*(17), 5043–5056, doi:10.5194/bg-13-5043-2016.

Yamazaki, D., S. Kanae, H. Kim, and T. Oki (2011), A physically based description of floodplain inundation dynamics in a global river routing model, *Water Resour. Res., 47*(4), doi:10.1029/2010WR009726.

Yang, X., J. Tang, J. F. Mustard, J.-E. Lee, M. Rossini, J. Joiner, J. W. Munger, A. Kornfeld, and A. D. Richardson (2015), Solar-induced chlorophyll fluorescence that correlates with canopy photosynthesis on diurnal and seasonal scales in a temperate deciduous forest, *Geophys. Res. Lett., 42*(8), 2977–2987, doi:10.1002/2015GL063201.

Yang, X., P. E. Thornton, D. M. Ricciuto, and F. M. Hoffman (2016), Phosphorus feedbacks may constrain tropical ecosystem responses to changes in atmospheric CO_2 and climate, *Geophys. Res. Lett., 43*(13), 7205–7214, doi:10.1002/2016GL069241.

Yi, S., A. D. McGuire, J. Harden, E. Kasischke, K. Manies, L. Hinzman, A. Liljedahl, J. Randerson, H. Liu, V. Romanovsky, S. Marchenko, and Y. Kim (2009), Interactions between soil thermal and hydrological dynamics in the response of Alaska ecosystems to fire disturbance, *J. Geophys. Res. Biogeosci., 114*(G2), doi:10.1029/2008JG000841.

Zaehle, S., and A. D. Friend (2010), Carbon and nitrogen cycle dynamics in the O-CN land surface model: 1. Model description, site-scale evaluation, and sensitivity to parameter estimates, *Global Biogeochem. Cycles, 24*(1), GB1005, doi:10.1029/2009GB003521.

Zaehle, S., P. Friedlingstein, and A. D. Friend (2010), Terrestrial nitrogen feedbacks may accelerate future climate change, *Geophys. Res. Lett., 37*(1), L01,401, doi:10.1029/2009GL041345.

Zaehle, S., B. E. Medlyn, M. G. De Kauwe, A. P. Walker, M. C. Dietze, T. Hickler, Y. Luo, Y.-P. Wang, B. El-Masri, P. Thornton, A. Jain, S. Wang, D. Warlind, E. Weng, W. Parton, C. M. Iversen, A. Gallet-Budynek, H. McCarthy, A. Finzi, P. J. Hanson, I. C. Prentice, R. Oren, and R. J. Norby (2014), Evaluation of 11 terrestrial carbon–nitrogen cycle models against observations from two temperate free-air CO_2 enrichment studies, *New Phytol., 202*(3), 803–822, doi:10.1111/nph.12697.

Zeng, X., B. A. Drewniak, and E. M. Constantinescu (2013), Calibration of the Crop model in the Community Land Model, *Geosci. Model Dev. Discuss., 6*, 379–398, doi:10.5194/gmdd-6-379-2013.

Zhang, D., D. Hui, Y. Luo, and G. Zhou (2008), Rates of litter decomposition in terrestrial ecosystems: Global patterns and controlling factors, *J. Plant Ecol., 1*(2), 85–93, doi:10.1093/jpe/rtn002.

Zhu, Q., W. J. Riley, J. Tang, and C. D. Koven (2016), Multiple soil nutrient competition between plants, microbes, and mineral surfaces: Model development, parameterization, and example applications in several tropical forests, *Biogeosci., 13*(1), 341–363, doi:10.5194/bg-13-341-2016.

Zscheischler, J., M. D. Mahecha, S. Harmeling, and M. Reichstein (2013), Detection and attribution of large spatiotemporal extreme events in Earth observation data, *Ecol. Inform., 15*, 66–73, doi: 10.1016/j.ecoinf.2013.03.004.

Zscheischler, J., A. M. Michalak, C. Schwalm, M. D. Mahecha, D. N. Huntzinger, M. Reichstein, G. Berthier, P. Ciais, R. B. Cook, B. El-Masri, M. Huang, A. Ito, A. Jain, A. King, H. Lei, C. Lu, J. Mao, S. Peng, B. Poulter, D. Ricciuto, X. Shi, B. Tao, H. Tian, N. Viovy, W. Wang, Y. Wei, J. Yang, and N. Zeng (2014), Impact of large-scale climate extremes on biospheric carbon fluxes: An intercomparison based on MsTMIP data, *Global Biogeochem. Cycles, 28*(6), 585–600, doi:10.1002/2014GB004826.

Appendix H. Acronyms and Abbreviations

ACME	Accelerated Climate Modeling for Energy
AGU	American Geophysical Union
ALM	ACME Land Model
ALMA	Assistance for Land-surface Modeling Activities convention for NetCDF files
AMIP	Atmospheric Model Intercomparison Project
API	application programming interface
ASCAT	Advanced SCATterometer
BCS	Bayesian Compressive Sensing
C	carbon
CESM	Community Earth System Model
CF	Climate and Forecast convention for NetCDF files
C-LAMP	Carbon-Land Model Intercomparison Project
CLM	Community Land Model
C^4MIP	Coupled Climate-Carbon Cycle MIP
CMIP	Coupled Model Intercomparison Project
CRU	Climate Research Unit
CTFS	Center for Tropical Forest Science
CZO	Critical Zone Observatory
DA	data assimilation
DECK	Diagnostic, Evaluation, and Characterization of Klima
DGVM	dynamic global vegetation model
DOE	U.S. Department of Energy
DVM	dynamic vegetation model
ECV	essential climate variable
ESGF	Earth System Grid Federation
ESM	Earth System Model
ESM-SnowMIP	Earth System Model Snow Model Intercomparison Project
ESMValTool	Earth System Model Evaluation Tool
ET	evapotranspiration
FACE	Free-Air Carbon dioxide Enrichment
FIA	Forest Inventory and Analysis
FLUXNET	Global eddy covariance flux network of regional networks
ForestGEO	Forest Global Earth Observatory
GEDI	Global Ecosystem Dynamics Investigation
GEM	Global Ecosystem Monitoring network
GFDL	Geophysical Fluid Dynamics Laboratory
GLACE	Global Land-Atmosphere Coupling Experiment
GPP	gross primary production
GRDC	Global Runoff Data Center
GRACE	Gravity Recovery And Climate Experiment
GSA	global sensitivity analysis
GSWP3	Global Soil Wetness Project 3
GUI	graphical user interface
HPC	high-performance computing

ICOS	Integrated Carbon Observation System
ILAMB	International Land Model Benchmarking
IS	imaging spectroscopy
ITCZ	Inter-Tropical Convergence Zone
ITEX	International Tundra Experiment
JPL	Jet Propulsion Laboratory
JSON	JavaScript Object Notation
KL	Karhunen-Loeve
LAI	leaf area index
LH	latent heat
LiDAR	Light Detection And Ranging
LIS	Land Information System
LS3MIP	Land Surface, Snow and Soil Moisture Model Intercomparison Program
LSM	land surface model
LUCID	Land-Use and Climate, IDentification of robust impacts
LULCC	land use and land cover change
LUMIP	Land Use Model Intercomparison Project
LVT	Land surface Verification Toolkit
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MDF	model–data fusion
MIP	model intercomparison project
ModEx	Model–data experimentation
MOPEX	Model Parameter Estimation Experiment
MSE	mean-square error
MsTMIP	Multi-scale Synthesis & Terrestrial Model Intercomparison Project
MTE	model tree ensemble
NACP	North American Carbon Program
NASA	National Aeronautics and Space Administration
NBP	net biosphere productivity
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
NCL	NCAR Command Language
NDVI	Normalized Difference Vegetation Index
NEE	net ecosystem exchange
NEON	National Ecological Observatory Network
NEP	net ecosystem productivity
NetCDF	Network Common Data Form
NGEE	Next Generation Ecosystem Experiments
NOAA	National Oceanic and Atmospheric Administration
NSE	Nash-Sutcliffe Efficiency
NPP	net primary productivity
OAT	one at a time
PalEON	Paleo-Ecological Observatory Network
PALS	Protocol for the Analysis for Land Surface models
PC	polynomial chaos
PCMDI	Program for Climate Model Diagnosis and Intercomparison
PCN	Permafrost Carbon Network
PDA	Parameter data assimilation
PDF	probability density function
PEcAn	Predictive Ecosystem Analyzer

PF	particle filter
PFT	plant functional type
PLUMBER	PALS Land Surface Model Benchmarking Evaluation Project
PLUME	Processes Linked to Uncertainties Modelling Ecosystems
PMP	PCMDI Metrics Package
QOIs	quantities of interest
RAINFOR	Amazon Forest Inventory Network
Re	ecosystem respiration
RECCAP	REgional Carbon Cycle Assessment and Processes
RMSE	root-mean-square error
RTM	radiative transfer model
SA	sensitivity analysis
SAChES	Scalable Adaptive Chain Ensemble Sampling
SavMIP	MIP focused on Australian savannas
SDA	state-variable data assimilation
SFA	Scientific Focus Area
SH	sensible heat
SIF	solar-induced fluorescence
SMAP	Soil Moisture Active Passive mission
SMOS	Soil Moisture and Ocean Salinity mission
SOM	soil organic matter
SPRUCE	Spruce and Peatland Responses Under Climatic and Environmental Change
SST	sea surface temperature
SWE	snow water equivalent
TBM	terrestrial biosphere model
TES	Terrestrial Ecosystem Science
TF	Traceability Framework
TIR	thermal infrared
TRACE	Tropical Responses to Altered Climate Experiment
TRIP	Total Runoff Integrating Pathways
TRMM	Tropical Rainfall Measurement Mission
TWS	total water storage
TWSA	total water storage anomaly
UK	United Kingdom
UQ	uncertainty quantification
US	United States
USA	United States of America
USDA	US Department of Agriculture
VD	Variance Decomposition
VDM	vegetation demographic model
WCE	weather and climate extreme
WIBCS	Weighted Iterative Bayesian Compressive Sensing
WMO	World Meteorological Organization
XML	eXtensible Markup Language
ZPW	Zero Power Warming

For More Information

CLIMATE AND ENVIRONMENTAL SCIENCES DIVISION
http://science.energy.gov/ber/research/cesd
Gary Geernaert, gerald.geernaert@science.doe.gov

REGIONAL AND GLOBAL CLIMATE MODELING
https://climatemodeling.science.energy.gov/rgcm
Renu Joseph, renu.joseph@science.doe.gov

EARTH SYSTEM MODELING
https://climatemodeling.science.energy.gov/esm
Dorothy Koch, dorothy.koch@science.doe.gov