2016 International Land Model Benchmarking (ILAMB) Workshop: Comprehensive evaluation of the carbon cycle, hydrology, and terrestrial ecosystem processes in Earth system models

Workshop Co-Chairs:

Forrest M. Hoffman[1], William J. Riley[2], James T. Randerson[3], Gretchen Keppel-Aleks[4], David M. Lawrence[5], Renu Joseph[6], and Dorothy Koch[6]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[2]Lawrence Berkeley National Laboratory, Berkeley, California, USA
[3]University of California Irvine, Irvine, California, USA
[4]University of Michigan Ann Arbor, Ann Arbor, Michigan, USA
[5]National Center for Atmospheric Research, Boulder, Colorado, USA
[6]U.S. Department of Energy, Germantown, Maryland, USA

Updated May 11, 2016

1 Prospectus

1.1 Workshop Dates:

May 16–18, 2016

1.2 Workshop Location:

DoubleTree by Hilton Hotel Washington DC
1515 Rhode Island Avenue NW
Washington, District of Columbia 20005-5595, USA
+1-202-232-7000

1.3 Workshop Rationale:

As Earth system models (ESMs) become increasingly complex, there is a growing need for comprehensive and multi-faceted evaluation of model predictions. To advance understanding of biogeochemical processes and their interactions with hydrology and climate under conditions of increasing atmospheric carbon dioxide (CO_2), new methods are needed 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 ESMs are essential for reducing uncertainties associated with projections of climate change during the remainder of the 21st century. In an effort sponsored by the U.S. Department of Energy (DOE), a diverse team of National Laboratory and university researchers is developing new diagnostic approaches for evaluating ESM hydrological and biogeochemical process representations. This research effort supports the International Land Model Benchmarking (ILAMB) project (http://www.ilamb.org/) by creating an open source benchmarking system that leverages the growing collection of laboratory, field, and remote sensing data. This benchmarking system performs comparisons of 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. Next generation benchmarking priorities will focus on design of new perturbation experiments (e.g., atmospheric CO_2 enrichment, water exclusion, nutrient addition, soil/plant warming) and resulting model evaluation metrics, new metrics from extreme events (e.g., drought, floods), and process-specific experiments (e.g., litterbags, 14C tracers). This benchmarking system is expected to become an integral part of model verification for future rapid model development cycles. Moreover, it will contribute model analysis and evaluation capabilities to phase 6 of the Coupled Model Intercomparison Project (CMIP) and future model and model–data intercomparison experiments.

The complexity of today's process-rich ESMs poses a verification challenge to developers, implementing new parameterizations or tuning process representations. Model developers and software engineers require a systematic means for evaluating changes in model results to assure that their developments improve the fidelity of the target process representations while not adversely affecting results in other parts of the model. In addition to objectively assessing the performance of ESMs and identifying model weaknesses, the ILAMB benchmarking system can also provide a framework for verifying and tuning model developments, supporting a more rapid development cycle and providing continuous and documented evaluation of model skill. DOE's Accelerated Climate Modeling for Energy (ACME) project has adopted ILAMB for this purpose and is implementing new metrics for more advanced features currently under development and testing.

CMIP provides essential information about future climate scenarios and ESM behavior that is used by the Intergovernmental Panel on Climate Change (IPCC) for periodic assessment and more broadly by policy makers and resource managers for the design of effective climate mitigation and adaptation strategies. Planning for CMIP6 is well underway, and the sophistication of model representations of biosphere and carbon cycle processes will far exceed levels achieved in past efforts. The complexity and volume of archived simulation output will be unparalleled, creating new opportunities for increasing our knowledge of biosphere–climate interactions, yet creating important challenges with respect to effectively harvesting this information to reduce uncertainties and improve our understanding of fundamental Earth system processes. Coordination among several MIPs is needed, as well as enhanced information flow among the modeling, evaluation, and observational communities.

We propose to continue and expand DOE's role in coordination of CMIP activities and to improve upon DOE's capabilities to assess the fidelity of model development in the ACME project by supporting comprehensive terrestrial model evaluation and benchmarking through the ILAMB project. Specifically, we propose to convene a three-day workshop to be held in downtown Washington, DC, USA, during May 16–18, 2016. Following more than five years after the first ILAMB Workshop in the U.S.—which was co-sponsored by DOE, NASA, and the International Geosphere–Biosphere Programme's (IGBP's) Analysis, Integration and Modeling of the Earth System (AIMES) project in January 2010—this second U.S. workshop is designed to accomplish the following objectives:

- To highlight new techniques for model evaluation that can reduce uncertainties with respect to biosphere processes and biogeochemical feedbacks with the climate system;
- To enable coordination among the Coupled Climate–Carbon Cycle Model Intercomparison Project (C^4MIP); the Land Surface, Snow, and Soil Moisture Model Intercomparison Project (LS3MIP); and the LandUse Model Intercomparison Project (LUMIP) activities, particularly with respect to synergies that may exist for model evaluation and analysis;
- To increase awareness of new data streams that will be available for model verification and benchmarking from remote sensing, in situ measurements, and synthesis activities;
- To increase the use and sharing of information and community tools for model evaluation and benchmarking, including the ILAMB software package;
- To design new metrics and evaluation approaches for integration into the next generation ILAMB system; and
- To create new metrics that integrate across carbon, surface energy, hydrology, and land use disciplines.

1.4 Workshop Outcomes:

To meet these objectives for DOE's climate research mission and to provide needed international climate science leadership, the workshop will invite participation from approximately 40 leading Earth system modelers, remote sensing experts, ecosystem ecologists, and data producers from around the world, in addition to team members of the Biogeochemistry–Climate Feedbacks Scientific Focus Area (SFA) and Program Managers from DOE and other federal agencies. A list of candidate participants from 11 countries is provided below. The outcome of the workshop will include a formal written report summarizing invited presentations, breakout group findings, and discussions regarding model evaluation strategies, gaps, and synergies. We also plan to have demonstrations on how to install and run version 1.0 of the ILAMB software system. In particular, the workshop report is expected address the topics following the draft outline shown here.

A. Model evaluation and benchmarking concepts and principles

B. Benchmarking tools

   1. PALS / PLUMBER
   2. ESMValTool
   3. NASA LIS Evaluation
   4. ILAMB
   5. Existing model evaluation capabilities in use at modeling centers
   6. Synergies between different benchmarking activities

C. Existing and new metrics for carbon, water, energy, and ecosystem processes

   1. Ecosystem processes and states
   2. Hydrology
   3. Atmospheric CO_2
   4. Soil carbon and nutrient biogeochemistry
   5. Surface fluxes (energy and carbon)
   6. Vegetation dynamics

D. Model Intercomparison Project (MIP) benchmarking needs and evaluation priorities

   1. CMIP6 historical and DECK
   2. C^4MIP
   3. LS3MIP
   4. LUMIP
   5. TRENDY
   6. MsTMIP
   7. PLUME-MIP

E. Next generation benchmarking challenges and priorities

   1. Process-specific experiments
   2. Metrics from extreme events
   3. Design of new perturbation experiments
   4. High latitude processes
   5. Tropical processes
   6. Global remote sensing
   7. Fluxnet and other surface hydrology and ecosystem networks

F. Model benchmarking gaps and synergies

   1. Integration with uncertainty quantification frameworks
   2. Computational requirements for post-processing and workflow
   3. Frameworks, Open Model Benchmarking Architecture (OpenMBA)
   4. Integration with archival and distribution cyberinfrastructure

5. Evaluating new process representations and rapid model development

G. Conclusions and next steps (5 and 10 y long term goals)

H. Appendix A. ILAMB Tutorial Materials and Data

The workshop will also serve as a venue for communicating recent scientific results in applying the emergent constraints approach for limiting the range of model predictions and reducing uncertainty, as well as new methodologies for generating and applying metrics in the assessment of model fidelity. Opportunities for information exchange will be through invited plenary presentations held over the course of the workshop, focused breakout group discussions, informal breakaway group meetings on the evening of the second day, and a poster session to be conducted on the first evening of the workshop in conjunction with a hosted dinner, consisting of food items suitable for eating while standing. Moreover, the workshop will provide the opportunity for key international participants to discuss the organization and proposal of special issues of scientific journals addressing new model analyses, evaluation and benchmarking, and verification and validation for new process representations and parameter optimization.

We expect the ILAMB benchmarking software development, supported primarily by DOE, to be of keen interest to workshop participants. Access to these software tools and related data sets will be provided at the workshop, and time will be set aside for a tutorial or training session focused on installing, using, and extending these tools for model development and verification, model validation exercises, and model–data intercomparison studies. This training session will be led by the software package developers from the BGC Feedbacks SFA to assure that participants receive hands-on assistance and to enable direct feedback from users on bugs, analytical issues, and desired features.

2016 International Land Model Benchmarking (ILAMB) Workshop

Final Agenda
(Updated May 10, 2016)

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 Land-use and land-cover change model performance metrics for LUMIP – *David M. Lawrence*	
	9:30 P.13 Multi-scale Synthesis & Terrestrial Model Intercomparison Project: From cohort to insight – *Christopher R. Schwalm*	
	9:45 P.14 Processes Linked to Uncertainties Modelling Ecosystems (PLUME-MIP) – *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)	Terrace Ballroom
	14:00 C^4MIP	
	14:10 LS3MIP	
	14:20 LUMIP	
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*	Terrace 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**	Terrace Ballroom
	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*	
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*	Terrace Ballroom
	Metrics from extreme events – *Hyungjun Kim and Maoyi Huang*	Directors Room
	Design of new perturbation experiments – *Martin De Kauwe and Ankur Desai*	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*	Terrace Ballroom
	Tropical processes – *Nathan McDowell and Paul Moorcroft*	Directors Room
	Global remote sensing – *David Schimel and Shawn Serbin*	Congressional Room
12:10	**Breakout Group Reports** (1–3 datasets, 1–3 new metrics, and bibliographies)	Terrace Ballroom
	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	
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**	

3 Plenary Presentation Abstracts

3.1 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@metoffice.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.

3.2 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.

3.3 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.

3.4 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 Office, 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 fill 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^{4}MIP)

Forrest M. Hoffman[1,†], *James T. Randerson*[2], *Charles D. Koven*[3], *and the C*^{4}*MIP 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^{4}MIP) 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 udnerstand 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^{4}MIP 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^{4}MIP 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.

3.5 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 bemchmark 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 ILAMB 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.

3.6 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.

3.7 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 ^{15}N 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° land surface 2 m temperature ($T_{2\mathrm{m}}$) 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 $T_{2\mathrm{m}}$ 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 $T_{2\mathrm{m}}$ 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 feedback 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.

3.8 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, Washingtom, 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 input 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.

4 Poster Presentation Abstracts

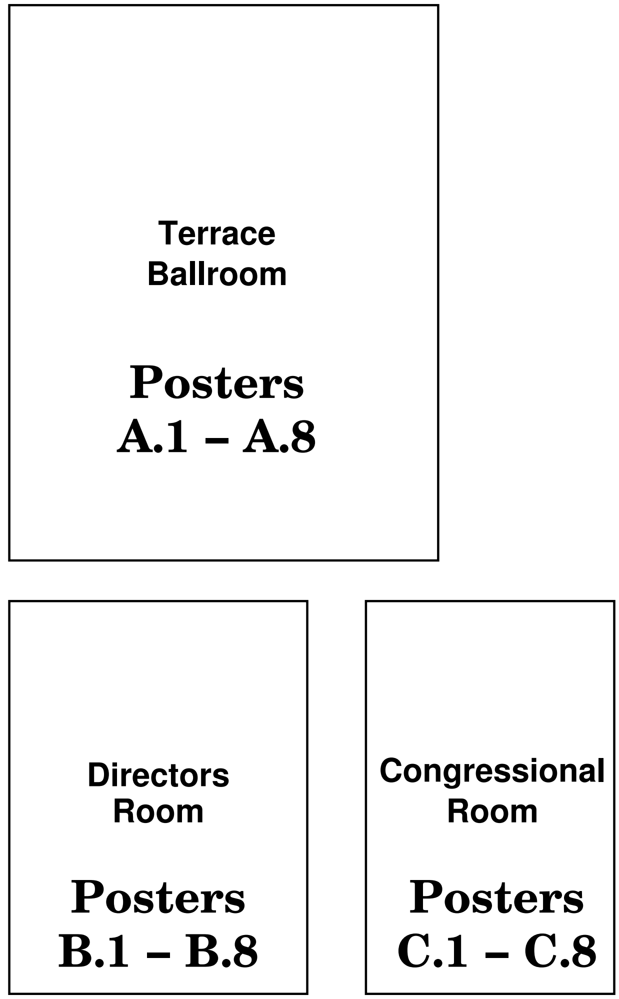
The poster session is scheduled for 6:00 p.m. to 7:00 p.m. on Monday, May 16, and a reception with light refreshments will be provided during that time. Poster viewing will be preceded by a set of lightning presentations in which each presenting author will be given 90 seconds to describe his or her poster with a single accompanying slide.

Posters will be hung on boards around the edges of the three available meeting rooms as follows:

- Posters A.1 through A.8 in the Terrace Ballroom
- Posters B.1 through B.8 in the Directors Room
- Posters C.1 through C.8 in the Congressional Room

Posters can remain up throughout the entire 3-day workshop.

Posters may be up to 4 feet $\times$ 4 feet in size, and two posters will occupy a single 8 foot $\times$ 4 foot board.



4.1 Benchmarking Packages

A.1 Towards an online web-based environment for land model evaluation and benchmarking

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

If we had an online web-based environment for land model evaluation and benchmarking that could incorporate the ILAMB package (and others, such as LVT), how would it work? Such an environment would offer many benefits, such as:

- 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

This poster will build on the Protocol for the Analysis of Land Surface models (PALS) plenary presentation by outlining the use-cases of such a web application, as well as detailing how provenance information, intellectual property and privacy can be incorporated.

A.2 The Python-based ILAMB Benchmarking System

Nathan Collier[1,†], *Forrest M. Hoffman*[1], *Dave M. Lawrence*[2], *and Andrew G. Slater*[3]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[2]National Center for Atmospheric Research, Boulder, Colorado, USA
[3]National Snow and Ice Data Center, University of Colorado, Boulder, Colorado, USA

[†]Author to whom correspondence should be addressed; e-mail: nathaniel.collier@gmail.com

In this poster we give an overview of the python-based ILAMB benchmarking system. The package centers around the abstraction of a benchmark into what we term a 'confrontation'. This abstraction pairs the benchmark dataset with the routines that (1) extract a comparable quantity from model results and (2) perform the relevant analysis. This notion of a confrontation makes community contribution of benchmarks possible in the form of plugins to the existing system. Confrontations may be developed as stand-alone files, which are supported/hosted by their respective groups and advertised to be compatible with the ILAMB system. In this poster, we give a flavor for how this may be accomplished and present an example of a custom confrontation, namely a permafrost benchmark.

A.3 The PCMDI Metrics Package

Peter J. Gleckler[1,†], *Charles Doutriaux*[1], *Paul J. Durack*[1], *Erik Mason*[2], *Ji-Woo Lee*[1], *Jrôme Servonnat*[3], *Karl E. Taylor*[1], *Yuying Zhang*[1], *and Dean N. Williams*[1]

[1]Lawrence Livermore National Laboratory, Livermore, California, USA
[2]Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA
[3]Laboratoire des Sciences du Climat et de l'Environnement, Gif sur Yvette Cedex, France

[†]Author to whom correspondence should be addressed; e-mail: gleckler1@llnl.gov

The PCMDI Metrics Package (PMP) leverages the vast CMIP data archive and uses established statistical error measures to gauge the consistency between observations and physical aspects of the simulated climate. It consists of four components: analysis software, an observationally based collection of global or near-global observations built on obs4MIPs, a database of performance metrics computed from all models contributed to CMIP, and usage documentation. The PMP will provide rapid and comprehensive feedback to modeling groups contributing simulations to the CMIP DECK and related experiments suitable for benchmarking purposes. The PMP (Gleckler et al., 2016, *EOS*, in press) uses the Python programming language in conjunction with the DOE supported Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT), a powerful software tool kit that provides cutting-edge diagnostic and visualization capabilities. Testing new simulations with the PMP does not require prior experience with Python or UV-CDAT, however those with some Python experience will have access to a wide range of state-of-the-art analysis capabilities. Community users of the PMP can develop and include additional tests of model behavior and work with the PCMDI team to integrate these into the package for others to use. The current release includes well-established large- to global-scale mean climate and variability performance metrics, while near future releases will include summary statistics for sea ice distribution, three-dimensional structure of ocean temperature and salinity, land surface vegetation characteristics (in collaboration with the ILAMB package development team), monsoon onset characteristics, major modes of climate variability, forced behavior such as the diurnal cycle, and selected "emergent constraints". The PCMDI and ILAMB packages are being coordinated to provide a unified DOE capability for gauging the consistency between state-of-the-art climate simulations and observations.

4.2 Emergent Constraints

A.4 Causes and implications of persistent atmospheric carbon dioxide biases in Earth system models

Forrest M. Hoffman[1,†], *James T. Randerson*[2],, *Vivek K. Arora*[3], *Qing Bao*[4], *Patricia Cadule*[5], *Duoying Ji*[6], *Chris D. Jones*[7], *Michio Kawamiya*[8], *Samar Khatiwala*[9], *Keith Lindsay*[10], *Atsushi Obata*[11], *Elena Shevliakova*[12], *Katharina D. Six*[13], *Jerry F. Tjiputra*[14], *Evgeny M. Volodin*[15], *and Tongwen Wu*[16]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[2]University of California Irvine, Irvine, California, USA
[3]Meteorological Service of Canada, Victoria, BC, V8W 3V6, Canada
[4]Institute of Atmospheric Physics, Chinese Academy of Science, Beijing, China
[5]Laboratoire des Sciences du Climat et de l'Environnement, Gif sur Yvette Cedex, France
[6]Beijing Normal University, Beijing 100875, China
[7]UK Met Office, Exeter, EX1 3PB, UK
[8]Japan Agency for Marine-Earth Science and Technology, Yokosuka-city, Kanagawa, Japan
[9]Columbia University, Palisades, New York, USA
[10]National Center for Atmospheric Research, Boulder, Colorado, USA
[11]Japan Meteorological Agency, Tsukuba-city, Ibaraki, Japan
[12]Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA
[13]Max Planck Institute for Meteorology, Hamburg, Germany
[14]Uni Climate, Uni Research, Allégaten 70, 5007 Bergen, Norway
[15]Russian Academy of Science, Moscow, 119333, Russia
[16]Beijing Climate Center, China Meteorological Administration, Beijing 100081, China

[†]Author to whom correspondence should be addressed; e-mail: forrest@climatemodeling.org

The strength of feedbacks between a changing climate and future CO_2 concentrations are uncertain and difficult to predict using Earth System Models (ESMs). We analyzed emission-driven simulations—in which atmospheric CO_2 levels were computed prognostically—for historical (1850–2005) and future periods (RCP 8.5 for 2006–2100) produced by 15 ESMs for the Fifth Phase of the Coupled Model Intercomparison Project (CMIP5). Comparison of ESM prognostic atmospheric CO_2 over the historical period with observations indicated that ESMs, on average, had a small positive bias in predictions of contemporary atmospheric CO_2. A key driver of this persistent bias was weak ocean carbon uptake exhibited by the majority of ESMs, based on comparisons with observations of ocean and atmosphere anthropogenic carbon inventories. We found a significant linear relationship between contemporary atmospheric CO_2 biases and future CO_2 levels for the multi-model ensemble. We used this relationship to create a contemporary CO_2 tuned model (CCTM) estimate of the atmospheric CO_2 trajectory for the 21$^{\text{st}}$ century. The CCTM yielded CO_2 estimates of 600 ± 14 ppm at 2060 and 947 ± 35 ppm at 2100, which were 21 ppm and 32 ppm below the multi-model mean during these two time periods. Uncertainty estimates derived from this approach were almost 6 times smaller at 2060 and almost 5 times smaller at 2100 than those from the ESM ensemble. The CCTM also significantly narrowed the range of CO_2-induced radiative forcing and temperature increases during the remainder of the 21$^{\text{st}}$ century. Because many

processes contributing to contemporary carbon cycle biases persist over decadal timescales, our analysis suggests uncertainties in future climate scenarios may be considerably reduced by tuning models to the long-term time series of CO_2 from Mauna Loa and other atmospheric monitoring stations.

A.5 Hydrological metrics for Earth system modeling

Hongyi Li[1,†], *Wei Wang*[1], *and L. Ruby Leung*[1]

[1]Pacific Northwest National Laboratory, Richland, Washington, USA

[†]Author to whom correspondence should be addressed; e-mail: hongyi.li@pnnl.gov

In hydrological simulations using earth system models, uncertainties could come from inputs, model structures, parameters etc. Hydrological metrics, when designed and interpreted carefully, contain rich information on the dominant hydrological processes, hence can be used to diagnose the major sources of uncertainties. We propose a small set of hydrological metrics derived from a mathematical framework based on the Budyko formula, which essentially bridge the annual, inter- and intra-annual variability of hydrological responses with the competition between water and energy potentials. We demonstrate the proposed metrics at over 400 natural watershed in the United States, then apply them at the global runoff and streamflow simulations with CLM-MOSART driven by four different forcing datasets. In addition, potential linkages between the proposed metrics and hydrological extremes are explored. Finally current limitations and future directions are discussed.

A.6 Reducing the uncertainty in the projection of the terrestrial carbon cycle by fusing models with remote sensing data

Shawn P. Serbin[1,†], *Toni Viskari*[1], *Phil Townsend*[2], *Alexey Shiklomanov*[3], *and Michael C. Dietze*[3]

[1]Brookhaven National Laboratory, Upton, New York, USA
[2]University of Wisconsin, Madison, Wisconsin, USA
[3]Boston University, Boston, Massachusetts, USA

[†]Author to whom correspondence should be addressed; e-mail: sserbin@bnl.gov

Modeling global change requires accurate representation of terrestrial carbon (C), energy and water fluxes. In particular, capturing the properties of vegetation canopies that describe the radiation regime are a key focus for global change research because the properties related to radiation utilization and penetration within plant canopies provide an important constraint on terrestrial ecosystem productivity, as well as the fluxes of water and energy from vegetation to the atmosphere. As such, optical remote sensing observations present an important, and as yet relatively untapped, source of observations that can be used to inform modeling activities. In particular, high-spectral resolution optical data at the leaf and canopy scales offers the potential for an important and direct data constraint on the parameterization and structure of the radiative transfer model (RTM) scheme within ecosystem models across diverse vegetation types, disturbance and management histories. In this presentation we highlight ongoing work to integrate optical remote sensing observations, specifically leaf and imaging spectroscopy (IS) data across a range of forest ecosystems, into complex ecosystem process models within an efficient computational assimilation framework as a means to improve the description of canopy optical properties, vegetation composition, and modeled radiation balance. Our work leverages the Predictive Ecosystem Analyzer (PEcAn; http://www.pecanproject.org/) ecoinformatics toolbox together with a RTM module designed for efficient assimilation of leaf and IS observations to inform vegetation optical properties as well as associated plant traits. Ultimately, an improved understanding of the radiation balance of ecosystems will provide a better constraint on model projections of energy balance, vegetation composition, and carbon pools and fluxes thus allowing for a better diagnosis of the vulnerability of terrestrial ecosystems in response to global change.

A.7 Benchmarking Earth system model carbon dynamics: An example using carbon use efficiency

Mathew Williams[1,†], *Jean-François Exbrayat*[1], *A. Anthony Bloom*[2], *T. Luke Smallman*[1], *and Chris D. Jones*[3]

[1]University of Edinburgh, Edinburgh, EH9 3JN, UK
[2]NASA Jet Propulsion Laboratory, Pasadena, California, USA
[3]UK Met Office, Exeter, EX1 3PB, UK

[†]Author to whom correspondence should be addressed; e-mail: mat.williams@ed.ac.uk

There is a pressing need to evaluate and improve process models used to make projections of terrestrial carbon cycling under expected forcings. We show how CARDAMOM, a novel system for retrieving carbon cycle parameters based on global datasets on LAI, burned area, biomass (tropical) and soil C, can provide independent benchmarks for process parameters. We focus here on carbon use efficiency (CUE), the ratio of net primary production (NPP) to gross primary production (GPP). CARDAMOM outputs show clear spatial patterns in CUE, with lowest values found in the moist tropics. Higher values are found in the E USA, Canadian prairies, Russian steppes, eastern Siberia, and parts of the Sahel and NW Australia. Fire disturbance may be a significant correlate with the retrievals, with more disturbed areas having higher CUE. An analysis of CMIP5 models was used to produce CUE maps. The analyses show clear differences among models, both in mean values and their spatial variation. A comparison with the independent CARDAMOM outputs is valuable. Plots of CUE estimates in climate phase space (mean annual temperature and annual precipitation) show differing temperature sensitivities (strong in HadGEM2 and MIROC, versus weak in CESM). CARDAMOM phase space analyses show smoother transitions than the models, suggestive of the categorical parameterisation via plant functions types in CMIP5 models. Zonal means of CUE for all models are compared against CARDAMOM values to indicate potential biases and differing latitudinal sensitivities. We suggest that the independent benchmark of CARDAMOM (which includes robust confidence intervals on estimates) provides an important assessment method for ESMs. A key question is: why is there a tendency for CUE to increase at high northern latitudes, which is not suggested by CARDAMOM? The implications of potential biases should be explored in further model projections. Finally, we suggest that the CARDAMOM analyses of CUE and other parameters can be used to provide a baseline for attributing the sensitivity of model projections to process variations.

A.8 Vegetation dynamics benchmarking based on forest inventory data

Chonggang Xu[1,†], *Daniel Johnson*[1], *and Nate McDowell*[1]

[1]Los Alamos National Laboratory, Los Alamos, New Mexico, USA

[†]Author to whom correspondence should be addressed; e-mail: xuchongang@gmail.com

Next generation land surface models in Earth system models are now incorporating dynamic vegetation components. The challenge of incorporating dynamic vegetation lies in constraining the large number of parameters in these complex models. The benefit of dynamic vegetation offsets the complexity resides in being able to better to capture successional dynamics and plant coexistence. While it is still possible to constrain the model using remote sensing data such as leaf area index, it is more likely to fit the model for the wrong reason and to provide unrealistic predictions under future climate conditions.

To make the model prediction more robust under future climate conditions, we propose to use the rich forest inventory data to further constrain the simulated canopy structure, growth, and survival. Here we provide an example of how mortality and growth can be used to constrain the DOE sponsored CLMED simulations using USFS forest inventory data. Disturbance may alter forest demographic processes. To test the model behavior under various disturbances, we also provide an example of functional model behaviors that could be used for testing model fidelity using the Forest-GEO network and FIA data. A future challenge is to compile a globe dataset for dynamic vegetation model benchmarking utilizing existing forest inventory data from different countries and ecosystems.

4.3 Model–Data Intercomparison Projects

B.1 Beyond benchmarking: Using CO_2 experiments to improve models

Martin G. De Kauwe[1,†] *and Belinda E. Medlyn*[2]

[1]Macquarie University, Sydney NSW 2109, Australia
[2]Western Sydney University, Richmond NSW 2753, Australia

[†]Author to whom correspondence should be addressed; e-mail: mdekauwe@gmail.com

Free-Air CO_2 Enrichment (FACE) experiments provide a unique opportunity to assess and reduce uncertainty in future predictions. The FACE Model–Data Synthesis (FACE MDS) project exploited two, decade long datasets, from the Duke forest and Oak Ridge sites, to test a suite of state-of-the-art ecosystem models.

As an alternative to standard model-observation benchmarking, an 'assumption-centred' model intercomparison approach was used. Following this approach, the underlying assumptions that models used to represent key ecosystem processes were identified and then these processes were evaluated against the experimental data.

Not only did this approach provide a clear roadmap to model improvement, the approach could readily be applied to other datasets as a more instructive method to improve predictive understanding in earth system modelling.

B.2 Decadal trends in the seasonal-cycle amplitude of terrestrial CO_2 exchange: An analysis of Multi-scale Terrestrial Model Intercomparison Project ensemble of terrestrial biosphere models

Akihiko Ito[1,†] *and MsTMIP model groups*

[1]National Institute for Environmental Studies, Tsukuba, Ibaraki Prefecture 305-0053, Japan

[†]Author to whom correspondence should be addressed; e-mail: itoh@nies.go.jp

The seasonal-cycle amplitude (SCA) of the atmosphere–ecosystem carbon dioxide (CO_2) exchange rate is a useful metric of the responsiveness of the terrestrial biosphere to environmental variations. It is unclear, however, what underlying mechanisms are responsible for the observed increasing trend of SCA in atmospheric CO_2 concentration. Using output data from the Multi-scale Terrestrial Model Intercomparison Project (MsTMIP), we investigated how well the SCA of atmosphere–ecosystem CO_2 exchange was simulated with 15 contemporary terrestrial ecosystem models during the period 1901–2010. Also, we made attempt to evaluate the contributions of potential mechanisms such as atmospheric CO_2, climate, land-use, and nitrogen deposition, through factorial experiments using different combinations of forcing data. Under contemporary conditions, the simulated global-scale SCA of the cumulative net ecosystem carbon budget of most models was comparable in magnitude with the SCA of atmospheric CO_2 concentrations. Through factorial experiments, it was shown that elevated atmospheric CO_2 exerted a strong influence on the seasonality amplification.

When the model considered not only climate change but also land-use and atmospheric CO_2 changes, the majority of the models showed amplification trends of the SCAs of photosynthesis, respiration, and net ecosystem production (+0.19% to +0.50% yr^{-1}). In the case of land-use change, it was difficult to separate the contribution of the effect of agricultural management because of inadequacies in both the data and models. The simulated amplification of SCA was approximately consistent with the observational evidence of the SCA in atmospheric CO_2 concentrations. Large inter-model differences remained, however, in the simulated global tendencies and spatial patterns of CO_2 exchanges. Further studies are required to identify a consistent explanation for the simulated and observed amplification trends, including their underlying mechanisms. Nevertheless, this study implied that monitoring of ecosystem seasonality would provide useful insights concerning ecosystem dynamics.

B.3 Height-structured vegetation and the carbon cycle in the NASA GISS Earth System Model/Ent Terrestrial Biosphere Model

Nancy Y. Kiang[1,†], *Igor Aleinov*[2], *Anastasia Romanou*[2], *Yeonjoo Kim*[3], *Crystal Schaaf*[4], *Qingsong Sun*[4], *Tian Yao*[5], *Feng Zhao*[6], *and Zhuosen Wang*[6]

[1]NASA Goddard Institute for Space Studies, New York, New York, USA
[2]Columbia University, Palisades, New York, USA
[3]Yonsei University, Seoul, South Korea
[4]University of Massachusetts Boston, Boston, Massachusetts, USA
[5]Universities Space Research Association (USRA)/NASA Goddard Space Flight Center, Greenbelt, Maryland, USA
[6]University of Maryland, College Park Maryland, USA

[†]Author to whom correspondence should be addressed; e-mail: Nancy.Y.Kiang@nasa.gov

The Ent Terrestrial Biosphere Model (Ent TBM) is a canopy height-stratified, subgrid patch community dynamic global vegetation model that provides the land surface albedo, water vapor conductance, and atmospheric exchange of CO_2 to the NASA Goddard Institute for Space Studies Earth System Model (GISS ESM) (Schmidt et al., 2014; Schmidt et al., 2006). Its distinctive features are: plant demographics after the Ecosystem Demography model; and the Analytical Clumped Two-Stream (ACTS) canopy radiative transfer model developed specially for the Ent TBM to account for foliage clumping in the vertical stratification of incident light and the diurnal variation in canopy albedo in heterogeneous canopies (Ni-Meister et al., 2010; Yang et al., 2010). It currently has a phenology scheme well tested at the field-site scale based on a combination of carbon balance and climatological control on timing of leafout and senescence (Kim et al., 2015), and ecological dynamics are slated for development with the Perfect Plasticity Approximation for plant competition for light and nutrients (Weng et al., 2015). The soil biogeochemistry model is that of CASA′ (Doney et al., 2006), with alternative temperature and moistures responses of soil respiration derived from data from Del Grosso et al. (2005).

We present here a survey of the different components of the Ent TBM and recent results from coupled carbon cycle simulations in the GISS ESM. Global scale simulations utilize the

Ent Global Vegetation Structure Dataset (Ent GVSD) version 1.0, which specifies satellite-derived mosaicked subgrid cover type, leaf area index, soil albedo, and geographic distributions in forest heights. From the Ent GVSD, plant densities derived from Ent allometric relations between plant heights and leaf area then provide the cohort-based canopy structure in which height- and PFT-classified cohorts are ensembles of identical individuals. This poster shows field site performance of light profile and canopy albedo calculations with the ACTS model, and seasonality of a variety of plant functional types (PFTs) and their water vapor and carbon fluxes. At the global scale, utilizing the Ent GVSD as boundary conditions, we show global land surface albedo predicted by the ACTS model. Also with the Ent GVSD, we show recent results of simulated atmospheric CO_2, in which soil carbon and plant carbohydrate reserves are the dynamic pools tuned to equilibrate with climate and observed vegetation structure.

Ongoing work involves development of plant physiological and allometric parameter sets from the TRY database, and introduction of crop timing and ecological dynamics.

B.4 Plant temperature acclimation in Earth system models: Scaling from the leaf to the canopy to the globe

Nicholas G. Smith[1,†], *Danica L. Lombardozzi*[2], *and Jeffrey S. Dukes*[1]

[1]Purdue University, West Lafayette, Indiana, USA
[2]National Center for Atmospheric Research, Boulder, Colorado, USA

[†]Author to whom correspondence should be addressed; e-mail: ngsmith@purdue.edu

Photosynthesis and respiration on land are the two largest carbon fluxes between the atmosphere and Earth's surface. The parameterization of these processes represent major uncertainties in the terrestrial component of the Earth System Models (ESMs) used to project future climate change. Research has shown that much of this uncertainty is due to the parameterization of the temperature responses of leaf photosynthesis and autotrophic respiration, which are typically based on short-term empirical responses. Here, we show that including longer-term responses to temperature, such as temperature acclimation, can help to reduce this uncertainty and improve model performance, leading to large changes in future land-atmosphere carbon and energy feedbacks across multiple models. However, contemporary acclimation formulations have flaws, including an underrepresentation of many important global flora. In addition, these parameterizations were done using multiple leaf-level studies that employed differing methodology. In a follow-up study, we used a consistent methodology to quantify model-relevant short- and long-term temperature responses of photosynthetic and respiratory parameters in multiple species representing each of the plant functional types used in ESMs. Our analyses indicated that the instantaneous responses were highly sensitive to longer-term temperature changes, but that this sensitivity was larger in species whose leaves typically experience a greater range of temperatures over the course of their lifespan. Additionally, our analyses indicated that gross leaf carbon uptake increases when plants are grown at warmer temperatures. Nonetheless, these data still need tested at larger scales. As such, we are actively working on constructing datasets and developing protocols to examine these parameterizations at increasingly larger scales. This includes developing simple leaf scale models, based on those employed in ESMs to test new parameterizations. We also are developing ways to test parameterizations at canopy and global scales using larger-scale models, such as the Community Land Model (CLM). While out data indicate that plants species are likely to increase their carbon uptake (and sink capacity) with future warming, model evaluation at scales beyond which the parameters were evaluated are needed for elucidating mechanisms that may act at larger scales. The integration of datasets at varying scales can improve the reliability of future model simulations and our study on plant temperature acclimation could serve as a model for doing this.

B.5 The Model–Data Integration Framework for NASA's Arctic Boreal Vulnerability Experiment (ABoVE)

Eric J. Stofferahn[1,†], *Joshua B. Fisher*[1], *Daniel J. Hayes*[2], *Deborah N. Huntzinger*[3], *and Christopher R. Schwalm*[4]

[1]NASA Jet Propulsion Laboratory, Pasadena, California, USA
[2]University of Maine, Orono, Maine, USA
[3]Northern Arizona University, Flagstaff, Arizona, USA
[4]Woods Hole Research Center, Falmouth, Massachusetts, USA

[†]Author to whom correspondence should be addressed; e-mail: ericstofferahn@gmail.com

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. A model–data integration framework is under development 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.

4.4 Remote Sensing or In Situ Data Sets

B.6 Diagnosing the downstream performance of the European Center for Meteorological Weather Forecasting (ECMWF) land data assimilation system

Concepcion Arroyo[1,†], *Thomas R. H. Holmes*[1], *Wade T. Crow*[1], *and Martha Anderson*[1]

[1]U.S. Department Agriculture, Beltsville, Maryland, USA

[†]Author to whom correspondence should be addressed; e-mail: arroyo.concha@gmail.com

Soil moisture (SM) is an essential variable for numerical weather and climate predictions. It controls the partitioning of available energy between sensible and latent heat flux at the soil-atmosphere interface and influences near-surface climate. The assimilation of remote sensing brightness temperatures (TB) into a soil moisture analysis should improve short-term numerical weather prediction (NWP) air temperature (T_{air}) forecasts. However, current results from European Centre for Medium-Range Weather Forecasts (ECMWF) experiments assimilating L-band TB do not support that conclusion that 24 hr T_{air} forecasts are enhanced over the United States. This region is particularly known for demonstrating a positive response when assimilating SM for weather forecasts. A more realistic SM initialization should lead to better estimates of evapotranspiration (ET) and, in turn, improved 24 hr T_{air} forecasts. The aim of this poster is to diagnose where the breakdown in performance between SM and T_{air} occurs using ECMWF experiments assimilating separately screen-level variables (CTRL) and SMOS (EXPT) at 0.25-degree spatial resolution from 2012 to 2013 across the contiguous US. For the SM analysis, we compare ECMWF experiments to a total of 273 SM ground stations from The Soil Climate Analysis Network (SCAN), U.S Climate Reference Network (USCRN) and two super sites local networks. For ET, a total of 41 Eddy covariance tower measurements together with remote sensing estimates of ET from thermal infrared ALEXI (Atmosphere-Land Exchange Inverse) model are available to compare the performance of the two ECMWF experiments. Together, this analysis of SM and ET allows us to diagnose what fails in the SM-ET-T_{air} causal chain for 24 hr T_{air} forecasts. Results show that the EXPT slightly improve the SM analysis. On the other hand, even if SM improves, ET forecast results still demonstrate a reduction in correlation for the EXPT when looking at the Corn Belt region, explaining the negative results for the 24 hr T_{air} forecast. In addition, we confirm ALEXI's capability to estimate ET fields as well as an excellent tool for diagnosing breakdown processes for SM assimilation studies.

This indicates that, for this particular region, the break-down in forecast performance can be traced back to the land-surface model's inability to accurately characterize SM/ET coupling rather than the assimilation of SM itself. With this work we demonstrate the use of ground and remote sensing technology to assess the hydrological response of a data assimilation system.

B.7 Obs4MIPs

Peter J. Gleckler[1,†] *and Duane E. Waliser*[2]

[1]Lawrence Livermore National Laboratory, Livermore, California, USA
[2]NASA Jet Propulsion Laboratory, Pasadena, California, USA

[†]Author to whom correspondence should be addressed; e-mail: gleckler1@llnl.gov

Obs4MIPs refers to a collection of well-established and documented datasets with demonstrated value for evaluating climate model simulations. Obs4MIPs data are aligned with the data conventions of the Coupled Model Intercomparison Project and made available on the Earth System Grid Federation (ESGF). Each obs4MIPs dataset can be compared to a field that is output in one or more of the CMIP5 or CMIP6 experiments. This technical alignment of observational products with climate model output can greatly facilitate model data comparisons; the traceability it provides is invaluable for benchmarking purposes and it prepares the CMIP community for "server side" capabilities such as the ILAMB and PCMDI metrics packages. Guidelines have also been developed for obs4MIPs product documentation that is of particular relevance for model evaluation. Obs4MIPs was initiated with support from DOE and NASA, and as a result of its success has now expanded to include contributions from a broader community including NOAA, ESA, EUMETSAT and projects such as CFMIP-OBS. Obs4MIPs is now a centralizing activity of the WCRP's Data Advisory Council (WDAC) as a mechanism to better coordinate the diverse suite of observations use to test climate models.

B.8 Global 0.5° hourly land surface 2 m air temperature datasets for model evaluations

Xubin Zeng[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

Land surface 2 m temperature (T_{2m}) is one of the most important variables in weather and climate studies, and its diurnal cycle is also needed for a variety of applications. Global long-term hourly T_{2m} observational data, however, do not exist. While such hourly products could be obtained from global reanalyses, they are found to be unrealistic in representing the T_{2m} diurnal cycle.

We have developed global hourly 0.5° T_{2m} datasets based on four reanalysis products (MERRA from 1979–2009, ERA-40 from 1958–2001, ERA-Interim from 1979–2009, and NCEP-NCAR from 1948–2009) 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).

Using these datasets, we have quantified the differences between the monthly mean [$T_m = (T_x + T_n)/2$] and the true monthly mean (T_{m24}) using 24 hourly values. Furthermore, the monthly diurnal temperature range $[DTR = T_x - T_n]$ is substantially different from the range of monthly averaged hourly temperature diurnal cycle ($RMDT$) over some regions (e.g., northern high latitudes) (Wang and Zeng 2014).

The polar amplification ratio of average temperature trend north of 65°N to that over global land (excluding Greenland and Antarctica) is weaker in summer (June–August) than in other seasons. Based on the probability distribution functions from the monthly anomalies of different variables, the coldest tenth percentile of temperature in each decade overall increases with time, while the warmest tenth percentile does not vary much from 1950–1979, followed by a rapid increase from 1980–2009 (Wang and Zeng 2015).

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 (T_{m24}) 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.

4.5 Model Evaluation Metrics

C.1 Performance of the new soil carbon module in JSBACH

Daniel Goll[1,‡], *Victor Brovkin*[1,†], *Thomas Raddatz*[1], *Jari Liski*[2], *and Tea Thum*[3]

[1]Max Planck Institute for Meteorology, Hamburg, Germany
[2]Finnish Environment Institute, Helsinki, Finland
[3]Finnish Meteorological Institute, Helsinki, Finland

[‡]Now at Laboratoire des Sciences du Climat et de l'Environnement, Gif sur Yvette Cedex, France
[†]Author to whom correspondence should be addressed; e-mail: victor.brovkin@mpimet.mpg.de

The response of soil organic carbon (C) to climate change, land use and land cover changes (LULCC) or other disturbances is a major uncertainty in estimating the land C balance. JSBACH has particular low skills in reproducing current soil organic C (SOC) stocks compared to other models. We exchanged the decomposition module in JSBACH with a data-based model and evaluated how the exchange affects the simulated C balance. The new soil module improves spatial variability in SOC, responses of decomposition to warming, historical land C balance, and reduces uncertainty in the net LULCC flux. Nonetheless, future changes in SOC remain highly uncertain.

C.2 A benchmark and diagnostic of climatological temperature control on soil carbon turnover

Charles D. Koven[1,†], *Gustaf Hugelius*[2], *and Dave M. Lawrence*[3]

[1]Lawrence Berkeley National Laboratory, Berkeley, California, USA
[2]Stockholm University, SE-106 91 Stockholm, Sweden
[3]National Center for Atmospheric Research, Boulder, Colorado, USA

[†]Author to whom correspondence should be addressed; e-mail: cdkoven@lbl.gov

We present a metric of soil carbon turnover based on the ratio of soil carbon in the top meter of soils to the net primary productivity. We show that the way in which this metric varies as a function of climatological temperature may be used as both a benchmark and a diagnostic of long-term sensitivity of soil carbon decomposition to climate. We further show that the sensitivity of soil carbon turnover to climatological temperature increases with decreasing temperatures, and that the benchmark suggests the existence of two emergent regimes, in which long-term climatological temperature sensitivity differs from short-term sensitivity, as well as a broad non-emergent regime in which long-term and short-term sensitivities largely agree. We show that the CMIP5 models fail to capture the basic dynamics of this benchmark, and present a simple scaling theory to explain the increased sensitivity at cold climates, and show how this can be included in both simple and more complex model representations.

C.3 A framework of detecting and attributing terrestrial ecosystem dynamics

Jiafu Mao[1,†], *Whitney Forbes*[2], *Daniel M. Ricciuto*[1], *Mingzhou Jin*[2], *Xiaoying Shi*[1], *Peter E. Thornton*[1], *and Forrest M. Hoffman*[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[2]University of Tennessee, Knoxville, Tennessee, USA

[†]Author to whom correspondence should be addressed; e-mail: maoj@ornl.gov

The statistical methods of detection and attribution (D&A) have been widely used in studies of climate change and quantifications of causes underlying the multi-year changes. Their successful applications in the terrestrial ecosystems, however, are limited, mainly due to the lack of long-term and broad-scale observational records, and the lack of suitable simulations from both coupled and uncoupled models. We will overcome these challenges by proposing a framework that includes the development of effective D&A algorithms, the design of factorial land model ensemble simulations, and the assembling of observational and observation-based datasets at relevant scales. This framework is expected to increase the efficiency and our confidence in attributing observed changes in carbon and water fluxes, and vegetation activities to extensive natural and anthropogenic factors.

C.4 JSBACH performance in comparison to observations and other models

Thomas Raddatz[1], *Victor Brovkin*[1,†], *Alexander Loew*[1], *Stefan Hagemann*[1], *Christian Reick*[1], *Daniela Dalmonech*[2], *and Soenke Zaehle*[2]

[1]Max Planck Institute for Meteorology, Hamburg, Germany
[2]Max Planck Institute for Biogeochemistry, Jena, Germany

[†]Author to whom correspondence should be addressed; e-mail: victor.brovkin@mpimet.mpg.de

A comprehensive evaluation of JSBACH, MPI-ESM land surface scheme, is difficult as observational constraints to some key results of global land surface models exhibit a large and not well quantified uncertainty (e.g. large scale amount of soil carbon). Furthermore, some land surface processes, that are important for feedbacks in the Earth System (e.g. strength of snow-masking by boreal forest) cannot be observed directly. Therefore, we aim at understanding how biases arise on the process level. MPI for Meteorology and MPI for Biogeochemistry are working together on a set of standard tests that are robust with respect to uncertainties in observations. Here we present some examples (vegetation, albedo, hydrology, phenology) based on CMIP5 results.

C.5 A Snow Heat Transfer Metric for process-level model evaluation

Andrew G. Slater[1,†], *David M. Lawrence*[2], *and Charles D. Koven*[3]

[1]National Snow and Ice Data Center, University of Colorado, Boulder, Colorado, USA
[2]National Center for Atmospheric Research, Boulder, Colorado, USA
[3]Lawrence Berkeley National Laboratory, Berkeley, California, USA

[†]Author to whom correspondence should be addressed; e-mail: aslater@colorado.edu

The ability of models to capture fundamental processes is assessed by looking at the functional relationships among model variables rather than the more common path of evaluating how well a model simulates an independently observed time series or spatial field. Snow is the most variable of terrestrial surface condition on the planet with the seasonal extent of snow cover varying by about 48% of land area in the Northern Hemisphere. Among its many properties of that snow is an effective insulator and during winter this typically restricts the amount of heat lost from the underlying ground/soil to the atmosphere. In turn, hydrologic and biogeochemical fluxes may be impacted. Here we present a diagnostic measure of snow insulation that is rooted in the physics of heat transfer. By assessing model capabilities in this way we can largely eliminate the uncertainties associated with model forcing. The proficiency of models to correctly expel heat from the terrestrial subsurface to the atmosphere in cold regions, particularly permafrost zones, is assessed by quantifying the relationship between seasonal amplitudes in soil and air temperatures while accounting for insulation from seasonal snow. Observations demonstrate the anticipated exponential relationship of attenuated soil temperature amplitude with snow depth. The marginal influence of snow insulation diminishes beyond a defined effective depth of 50 cm. Land models used within the CMIP5 experiment vary in their ability to reproduce the correct relationship and deficiencies can often be traced back to architectural or parameterization weaknesses of the models.

C.6 Benchmark nutrient competition in Earth system land models

Qing Zhu[1,†], *William J. Riley*[1], *Jinyun Tang*[1], *and Charles D. Koven*[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, California, USA

[†]Author to whom correspondence should be addressed; e-mail: qzhu@lbl.gov

Plant–microbe nutrient competition strongly affects plant productivity, soil microbial activity, and land surface CO_2 exchanges with the atmosphere. Here, we benchmark existing plant–microbe competition theories that have been implemented in Earth System Models (ESMs) using $^{15}N/^{33}P$ tracer experiments. We also applied a new competition theory based on Equilibrium Chemistry Approximation (ECA). We show that competition theories in current ESMs fail to capture observed patterns, while our new approach robustly simplifies the complex nature of nutrient competition and quantitatively matches the observations. Since plant carbon dynamics are strongly modulated by soil nutrient acquisition, we conclude that (1) existing ESM carbon-cycle predictions may be biased and (2) our new theory can improve

carbon–climate feedback predictions by mechanistically considering plant–microbe nutrient competition.

4.6 Sensitive Ecosystems (high latitudes, tropics, etc.)

C.7 New approaches to comparing permafrost soil carbon pools from Earth system models to observational data

Gustaf Hugelius[1,†], *J. W. Harden*[2], *A. D. McGuire*[3], *T. J. Bohn*[4], *E. J. Burke*[5], *S. Chadburn*[6], *G. Chen*[7], *X. Chen*[8], *D. J. Hayes*[7], *E. E. Jafarov*[9], *C. D. Koven*[10], *S. Peng*[11], *and K. M. Schaefer*[12]

[1]Stockholm University, Stockholm, Sweden

[2]U.S. Geological Survey, Menlo Park, California, USA and Stanford University, Stanford, California, USA

[3]U.S. Geological Survey, Fairbanks, Alaska, USA and University of Alaska, Fairbanks, Alaska, USA

[4]University of Washington, Seattle, Washington, USA and Arizona State University, Arizona, USA

[5]UK Met Office, Exeter, EX1 3PB, UK

[6]University of Exeter, Exeter, UK

[7]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

[8]University of Washington, Seattle, Washington, USA

[9]University of Colorado, Boulder, Colorado, USA

[10]Lawrence Berkeley National Laboratory, Berkeley, California, USA

[11]Laboratoire des Sciences du Climat et de l'Environnement, Gif sur Yvette Cedex, France and Laboratoire de Glaciologie et Gophysique de l'Environnement, CNRS and Universit Grenoble Alpes, France

[12]National Snow and Ice Data Center, University of Colorado, Boulder, Colorado, USA

[†]Author to whom correspondence should be addressed; e-mail: gustaf.hugelius@natgeo.su.se

Permafrost region soils contain very large stocks of soil organic carbon, which have accumulated over long time periods. These large stocks are a result of environmental factors and soil forming processes that are particularly prevalent or unique in cold region soils, including, peat formation, cryoturbation and the presence of near-surface permafrost. The extent and detail to which these processes are represented in Earth System Models varies, but is typically very limited. Here we compare the soil carbon pools generated by spin-up runs of a suite of land surface models from the Permafrost Carbon Network model intercomparison to observational data; (1) the empirically based maps of soil carbon storage in the Northern Circumpolar Soil Carbon Database (NCSCD) and (2) an extensive set of pedon data made available through the International Soil Carbon Network. Circumpolar spatial correlations reveal very weak linear relationships between modeled and empirically mapped soil carbon stocks. Based on the mapped soil classification data in the NCSCD (following U.S.D.A Soil Taxonomy), we assigned the mapped soil carbon stocks to the following categories: permafrost peatlands (Histels), cryoturbated soils (Turbels), other permafrost soils (Orthels) non-permafrost peatlands (Histosols) and non-permafrost mineral soils (remaining soil orders). Following this scheme, we analyze regional-scale patterns of deviations between empirical and modelled soil carbon stocks. We place a particular emphasis on model benchmarking in regions where soil carbon storage is dominated by the particular soil forming

processes affecting these different soil types. By focusing the spatial analyses to specific regions and soil types we can isolate which soil types are well represented in models and which are not. These types of comparisons reveal interesting trends that can provide semi-quantitative relationships between soil forming processes and observed model biases, which may in turn inform and improve future model development.

4.7 Uncertainty Quantification

C.8 Divergent ecosystem carbon dynamics resulting from ambiguous numerical implementation of nitrogen limitation

Jinyun Tang[1,†] *and William J. Riley*[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, California, USA

[†]Author to whom correspondence should be addressed; e-mail: jinyuntang@lbl.gov

For the given mathematical formulation of nitrogen dynamics in the land biogeochemical (BGC) module of the version-0 accelerated climate model for energy (ACME) earth system model, we suggest there are at least three possible numerical implementations of nitrogen limitation. Although they predict similar global terrestrial carbon and nitrogen distributions by year 2000, simulations with the Representative Concentration Pathway 4.5 (RCP4.5) scenario atmospheric CO_2 over year 2001–2300 diverge: (1) the global land varies from being a strong carbon sink to a weak source; and (2) tropical and artic ecosystems are either a strong carbon sink or source. We attribute this uncertainty to the lack of consistent interpretation of nitrogen limitation in land BGC models, making many of these models lacking the calibratability, i.e. models calibrated under nitrogen limitation will assign wrong values to parameters that are related to non-nitrogen dynamics. We give our recommendations on how to systematically remove this uncertainty.

5 ILAMB v2 Package Tutorial

ILAMB 2016 Workshop Tutorial Outline

Nathan Collier[1,†]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

[†]Author to whom correspondence should be addressed; e-mail: nathaniel.collier@gmail.com

The tutorial for the python ILAMB package is split into two parts, across two days. In the following I give an outline of the tutorial session, providing a high level description of the content as well as links to written documents. The written documents are not meant to be a script, but rather cover the concepts we will be discussing and serve as a resource for users to review at a later time.

Part I: Getting Started

The first part addresses basic usage: installation, getting data, running the code, adding models, and adding benchmarks or new datasets. This session targets those who plan to use the software and does not require much python knowledge.

- Installation – While I will cover this in the tutorial session, I encourage participants to try going through the text version on their own. This is because installation issues take time, and while I want to address questions, this could eat up the whole session time. I will make myself available to help users with issues all day Monday through Wednesday. `http://climate.ornl.gov/~ncf/ILAMB/docs/install.html`
- First Steps – This tutorial shows how the system works by a small example. The users will see how to organize their data and setup a configuration file which defines a set of benchmarks to run though the mean state analysis we have developed. We will then go over the results and explain how to navigate the output information. `http://climate.ornl.gov/~ncf/ILAMB/docs/first_steps.html`
- Adding a Model – In this tutorial the users will learn how their own model results may be added to the sample data from the previous tutorial. `http://climate.ornl.gov/~ncf/ILAMB/docs/add_model.html`
- Adding a Dataset – In this tutorial we will add a benchmark dataset to the sample data from the First Steps tutorial. If the dataset we wish to add maps to a variable which models output, either directly or via an algebraic manipulation, then this is a matter of adding two to three lines to the configuration file. `http://climate.ornl.gov/~ncf/ILAMB/docs/add_data.html`

Part II: Advanced usage

The second part addresses adding new benchmark comparisons to the system which are not simply algebraic manipulations of model output variables as well as including analysis which differs from our mean state analysis. While basic usage requires little knowledge of how the ILAMB package works, a greater understanding will be needed for these more complex systems. These tutorials are meant to help make the learning curve a bit less daunting.

- Overview of the ILAMB Package – What exactly is the package? How is it organized? How does it work? The goal here is to provide context for the users so they understand how their future additions will fit into the overall package functionality. `http://climate.ornl.gov/~ncf/ILAMB/docs/overview.html`
- Writing your own Confrontation – In this tutorial I will explain the Confrontation concept which is implemented as a python class and show how you can create your own by simply reimplementing the functions which your special purpose code does differently from the base Confrontation. I will show an example using the Global Net Ecosystem Carbon Balance code which exists in our system. `http://climate.ornl.gov/~ncf/ILAMB/docs/confront.html`

6 Participant List

Gab Abramowitz	gabriel at unsw dot edu dot au
Anders Ahlstrom	anders dot ahlstrom at nateko dot lu dot se
Igor Aleinov	igor dot aleinov at nasa dot gov
Concepcion Arroyo	arroyo dot concha at gmail dot com
Dennis Baldocchi	baldocchi at berkeley dot edu
Martin Best	Martin dot Best at metoffice dot com
Benjamin Bond-Lamberty	BondLamberty at pnnl dot gov
Victor Brovkin	victor dot brovkin at mpimet dot mpg dot de
Nuno Carvalhais	ncarval at bgc-jena dot mpg dot de
Philippe Ciais	nathalie dot chedeville at cea dot fr
Nathan Collier	nate at climatemodeling dot org
Stuart J. Davies	DaviesS at si dot edu
Martin De Kauwe	mdekauwe at gmail dot com
Alberto M. de la Torre	albmar at ceh dot ac dot uk
Scott Denning	denning at atmos dot colostate dot edu
Ankur Desai	desai at aos dot wisc dot edu
Paul A. Dirmeyer	pdirmeye at gmu dot edu
Richard Ellis	rjel at ceh dot ac dot uk
Rosie Fisher	rfisher at ucar dot edu
Peter J. Gleckler	gleckler1 at llnl dot gov
Hirofumi Hashimoto	hirofumi dot hashimoto at gmail dot com
Forrest M. Hoffman	forrest at climatemodeling dot org
Thomas Holmes	thomas dot holmes at ars dot usda dot gov
Maoyi Huang	maoyi dot huang at pnnl dot gov
Gustaf Hugelius	gustaf dot hugelius at natgeo dot su dot se
Akihiko Ito	itoh at nies dot go dot jp
Atul K. Jain	jain1 at illinois dot edu
Gretchen Keppel-Aleks	gkeppela at umich dot edu
Nancy Y. Kiang	nancy dot y dot kiang at nasa dot gov
Hyungjun Kim	hjkim at iis dot u-tokyo dot ac dot jp
Randal D. Koster	randal dot d dot koster at nasa dot gov
Charles D. Koven	cdkoven at lbl dot gov
Sujay Kumar	sujay dot v dot kumar at nasa dot gov
David M. Lawrence	dlawren at ucar dot edu
Hongyi Li	Hongyi dot Li at pnnl dot gov
Chris Lu	chris dot lu at csiro dot au
Yiqi Luo	yluo at ou dot edu
Jiafu Mao	maoj at ornl dot gov
Hank A. Margolis	hank dot a dot margolis at nasa dot gov

Nathan McDowell	mcdowell at lanl dot gov
Umakant Mishra	umishra at anl dot gov
Paul Moorcroft	paul_moorcroft at harvard dot edu
Mingquan Mu	mmu at uci dot edu
George Pau	gpau at lbl dot gov
Gregory Quetin	gquetin at atmos dot washington dot edu
James T. Randerson	jranders at uci dot edu
Dan Ricciuto	ricciutodm at ornl dot gov
William J. Riley	wjriley at lbl dot gov
Kevin Schaefer	kevin dot schaefer at nsidc dot org
David S. Schimel	David dot Schimel at jpl dot nasa dot gov
Christopher R. Schwalm	schwalm dot christopher at gmail dot com
Shawn Serbin	sserbin at bnl dot gov
Elena Shevliakova	Elena dot Shevliakova at noaa dot gov
James Simkins	jsimkins2 at wisc dot edu
Andrew Slater	aslater at nsidc dot org
Nicholas G. Smith	ngsmith at purdue dot edu
Eric J. Stofferahn	estoffer at masonlive dot gmu dot edu
Kaoru Tachiiri	tachiiri at jamstec dot go dot jp
Jinyun Tang	jinyuntang at lbl dot gov
Dali Wang	wangd at ornl dot gov
Ying Ping Wang	Yingping dot Wang at csiro dot au
Mathew Williams	mat dot williams at ed dot ac dot uk
Jianyang Xia	jyxia at des dot ecnu dot edu dot cn
Chonggang Xu	xuchongang at gmail dot com
Xubin Zeng	xubin at atmo dot arizona dot edu
Qian Zhang	qian dot zhang at bnu dot edu dot cn
Sha Zhou	shazhou at princeton dot edu
Qing Zhu	qzhu at lbl dot gov

Presentation Index

Author Index