Have Land Surface Processes in Earth System Models Improved Over Time?

Forrest M. Hoffman[1,2], Nathan Collier[1], Charles D. Koven[3], David M. Lawrence[4], Gretchen Keppel-Aleks[5], James T. Randerson[6], Mingquan Mu[6], William J. Riley[3], Qing Zhu[3], Jiafu Mao[1], Hyungjun Kim[7], J. Keith Moore[6], and Weiwei Fu[6]

[1]Oak Ridge National Laboratory (ORNL), [2]University of Tennessee Knoxville, [3]Lawrence Berkeley National Laboratory (LBNL), [4]National Center for Atmospheric Research (NCAR), [5]University of Michigan Ann Arbor, [6]University of California Irvine, and [7]University of Tokyo

University of Arizona
Department of Hydrology and Atmospheric Sciences Weekly Colloquium

February 18, 2021

Forrest M. Hoffman, Computational Earth Sciences Group Leader

- 32 years at ORNL; 27 years as staff in ESD, CSMD, and CSED
- B.S. (1991) and M.S. (2004) in Physics from University of Tennessee, Knoxville; M.S. (2012) and Ph.D. (2015) in Earth System Science from University of California, Irvine
- develop and apply Earth system models to study global biogeochemical cycles, including terrestrial & marine carbon cycle
- investigate methods for reconciling uncertainties in carbon cycle–climate feedbacks through comparison with observations
- apply artificial intelligence methods (machine learning and data mining) to environmental characterization, simulation, & analysis
- Joint Faculty Professor, University of Tennessee, Knoxville, Department of Civil & Environmental Engineering

Oak Ridge National Laboratory

The University of Tennessee Knoxville

US Dept. of Energy's RUBISCO Science Focus Area

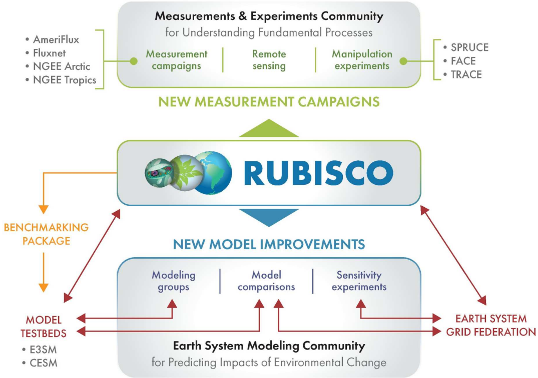
Forrest M. Hoffman (Laboratory Research Manager),
William J. Riley (Senoir Science Co-Lead), and
James T. Randerson (Chief Scientist)

Research Goals

- Identify and quantify interactions between biogeochemical cycles and the Earth system
- Quantify and reduce uncertainties in Earth system models (ESMs) associated with interactions

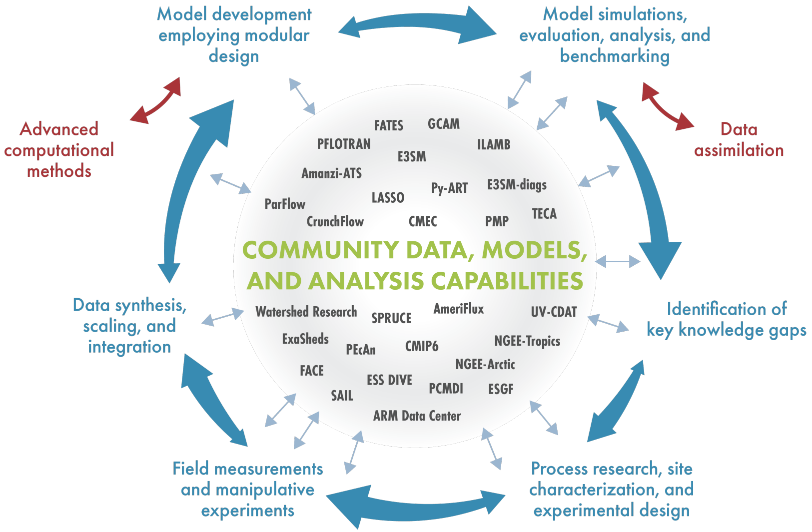
Research Objectives

- Perform hypothesis-driven analysis of biogeochemical & hydrological processes and feedbacks in ESMs
- Synthesize in situ and remote sensing data and design metrics for assessing ESM performance
- Design, develop, and release the International Land Model Benchmarking (ILAMB) and International Ocean Model Benchmarking (IOMB) packages for systematic evaluation of model fidelity
- Conduct and evaluate CMIP6 simulations with ESMs



The RUBISCO SFA works with the measurements and the modeling communities to use best-available data to evaluate the fidelity of ESMs. RUBISCO identifies model gaps and weaknesses, informs new model development efforts, and suggests new measurements and field campaigns.

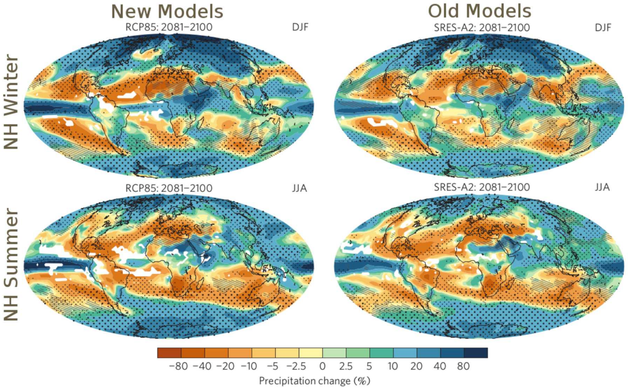
DOE's Model-Data-Experiment Enterprise



Problem: Model Uncertainty

Model uncertainty is one of the biggest challenges we face in Earth system science, yet comparatively little effort is devoted to fixing it (Carslaw et al., 2018)

- Model complexity is rapidly increasing as detailed process representations are added
- Evidence shows overall model uncertainty is reduced only slowly and is sometimes increased (Knutti and Sedláček, 2013)
- A balance must be struck between model "elaboration" and efforts to reduce model uncertainty



Patterns of precipitation change across two generations of models. Adapted from Knutti and Sedláček (2013).

Why is Reducing Uncertainty a Challenge?

- Ecosystems have complex responses to a wide range of forcing factors in heterogeneous spatial environments, requiring a highly multivariate approach
- The focus is on adding complexity (e.g., more detailed representations of plant traits, photosynthesis, nutrient limitation, respiration), assuming more processes is better
- However, model uncertainty may increase, even as predictions of states and fluxes improve
- Rigorous confrontation of models with independent observations and large ensembles of simulations are required to reduce uncertainty
- Modeling centers have a limited capacity to conduct sensitivity experiments and systematically assess model fidelity, especially in fully coupled Earth system models
- Community-developed benchmarking tools are beginning to address part of the solution

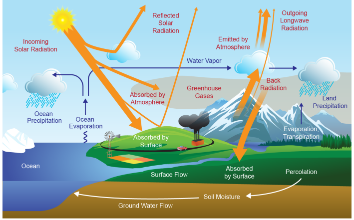
What is ILAMB?

Originally, ILAMB was a community activity designed to:

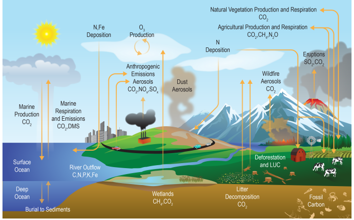
- **Develop internationally accepted benchmarks** for land model performance by drawing upon collaborative expertise
- **Promote the use of these benchmarks** for model intercomparison
- **Strengthen linkages between experimental, remote sensing, and climate modeling communities** in the design of new model tests
- **Support the development of open source benchmarking tools**

Now, ILAMB is a:

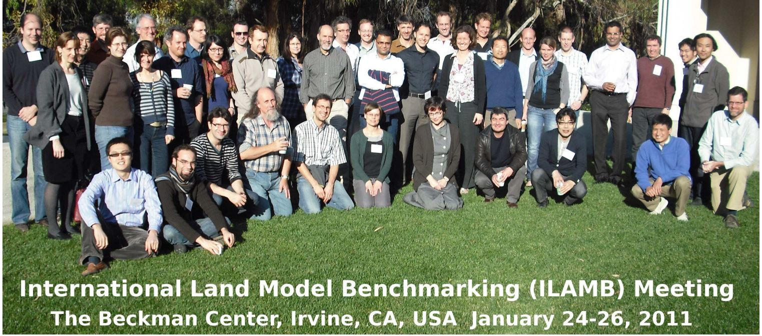
- **Community:** global group of modelers and scientists enthusiastic about benchmarking
- **Datasets:** curated collection of datasets formatted for easy data-model integration
- **Methods:** standard library of techniques for benchmarking models
- **Software:** an extensible open source Python package
- **Results:** an easy-to-use catalog of model-data comparisons



Energy and Water Cycles



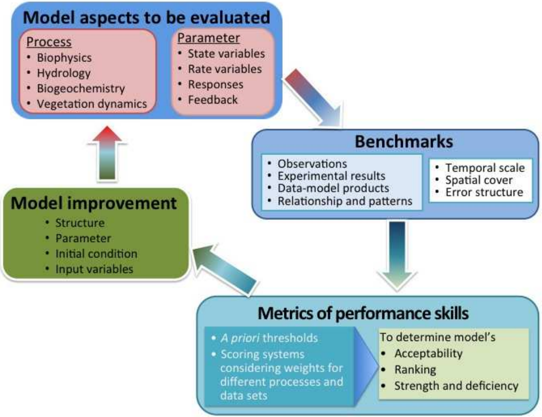
Carbon and Biogeochemical Cycles



- First ILAMB Meeting was held in Exeter, UK, on June 22–24, 2009
- Second ILAMB Meeting was held in Irvine, CA, USA, on January 24–26, 2011
  - ∼45 researchers participated from the United States, Canada, the United Kingdom, the Netherlands, France, Germany, Switzerland, China, Japan, and Australia
  - *Initial focus on CMIP5 models*
  - Developed methodology for model–data comparison and baseline standard for performance of land model process representations (Luo et al., 2012)

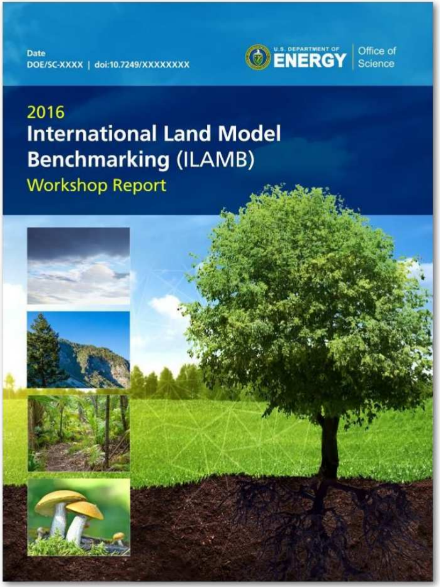
A Framework for Benchmarking Land Models

- A **benchmarking framework for evaluating land models** emerged and included (1) defining model aspects to be evaluated, (2) selecting benchmarks as standardized references, (3) developing a scoring system to measure model performance, and (4) stimulating model improvement
- Based on this methodology and prior work on the **Carbon-LAnd Model Intercomparison Project (C-LAMP)** (Randerson et al., 2009), a prototype model benchmarking package was developed for ILAMB



(Luo et al., 2012)

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



(Hoffman et al., 2017)

Third ILAMB Workshop was held to identify

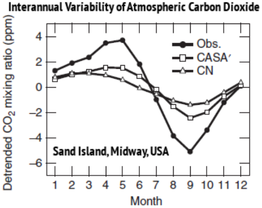
- New metrics for model benchmarking
- Model Intercomparison Project (MIP) evaluation needs
- Model development, test beds, and workflow requirements
- Observational datasets and needed measurements

Workshop Attendance

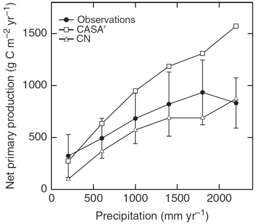
- 60+ participants from Australia, Japan, China, Germany, Sweden, Netherlands, UK, and US (10 modeling centers)
- ∼25 remote attendees at any time to enable participation by students and postdocs and enhance diversity and inclusion

What Is A Benchmark?

- A **benchmark** is a quantitative test of model function achieved through comparison of model results with observational data
- Acceptable performance on benchmarks **is a necessary but not sufficient condition** for a fully functioning model
- **Functional benchmarks** offer tests of model responses to forcings and yield insights into ecosystem processes
- Effective benchmarks must draw upon **a broad set of independent observations** to evaluate model performance at multiple scales



Models often fail to capture the amplitude of the seasonal cycle of atmospheric CO_2



Models may reproduce correct responses over only a limited range of forcing variables

(Randerson et al., 2009)

Why Benchmark Models?

- To **quantify and reduce uncertainties** in carbon cycle feedbacks to improve projections of future climate change
- To **quantitatively diagnose impacts of model development** on hydrological and carbon cycle process representations and their interactions
- To **guide synthesis efforts**, such as the Intergovernmental Panel on Climate Change (IPCC), by determining which models are broadly consistent with available observations (Eyring et al., 2019)
- To **increase scrutiny of key datasets** used for model evaluation
- To **identify gaps in existing observations** needed to inform model development
- To **accelerate delivery of new measurement datasets** for rapid and widespread use in model assessment

ILAMB Produces Diagnostics and Scores Models

- ILAMB generates a top-level **portrait plot** of model scores
- For every variable and dataset, ILAMB automatically produces
  - **Tables** containing individual metrics and metric scores (when relevant to the data), including
    - Reference and model **period mean**
    - **Bias** and **bias score** (S_{bias})
    - **Root-mean-square error (RMSE)** and **RMSE score** (S_{rmse})
    - **Phase shift** and **seasonal cycle score** (S_{phase})
    - **Interannual coefficient of variation** and **IAV score** (S_{iav})
    - **Spatial distribution score** (S_{dist})
    - **Overall score** (S_{overall})

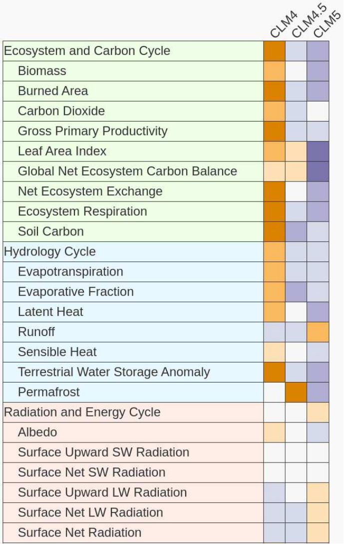
$$\Longrightarrow S_{\text{overall}} = \frac{S_{\text{bias}} + 2S_{\text{rmse}} + S_{\text{phase}} + S_{\text{iav}} + S_{\text{dist}}}{1+2+1+1+1}$$

  - **Graphical diagnostics**
    - Spatial contour maps
    - Time series line plots
    - Spatial Taylor diagrams (Taylor, 2001)
- Similar **tables** and **graphical diagnostics** for functional relationships
- ILAMB design, theory, and implementation are described in Collier et al. (2018)

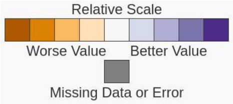
ILAMBv2.5 Package Current Variables

- **Biogeochemistry:** Biomass (Contiguous US, Pan Tropical Forest), Burned area (GFED4.1s), CO_2 (NOAA GMD, Mauna Loa), Gross primary production (Fluxnet, FLUXCOM), Leaf area index (AVHRR, MODIS), Global net ecosystem carbon flux (GCP, Khatiwala/Hoffman), Net ecosystem exchange (Fluxnet, FLUXCOM), Ecosystem respiration (Fluxnet, FLUXCOM), Soil C (HWSD, NCSCDv2, Koven)
- **Hydrology:** Evapotranspiration (GLEAM, MODIS), Evaporative fraction (FLUXCOM), Latent heat (Fluxnet, FLUXCOM, DOLCE), Permafrost (NSIDC), Runoff (Dai, LORA), Sensible heat (Fluxnet, FLUXCOM), Terrestrial water storage anomaly (GRACE)
- **Energy:** Albedo (CERES, GEWEX.SRB), Surface upward and net SW/LW radiation (CERES, GEWEX.SRB, WRMC.BSRN), Surface net radiation (CERES, GEWEX.SRB, WRMC.BSRN)
- **Forcing:** Surface air temperature (CRU, Fluxnet), Dirunal max/min/range temperature (CRU), Precipitation (CMAP, Fluxnet, GPCC, GPCP2), Surface relative humidity (ERA), Surface down SW/LW radiation (Fluxnet, CERES, GEWEX.SRB, WRMC.BSRN)

ILAMB Assessed Several Generations of CLM



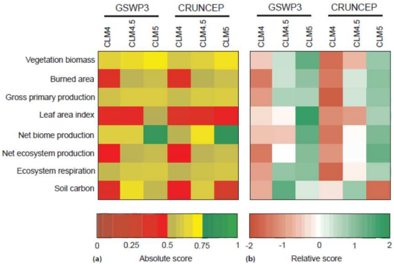
- Improvements in mechanistic treatment of hydrology, ecology, and land use with much more complexity in Community Land Model version 5 (CLM5)
- Simulations improved even with enhanced complexity
- Observational datasets are not always self-consistent
- Forcing uncertainty confounds assessment of model development



`http://webext.cgd.ucar.edu/I20TR/_build_set1F/`

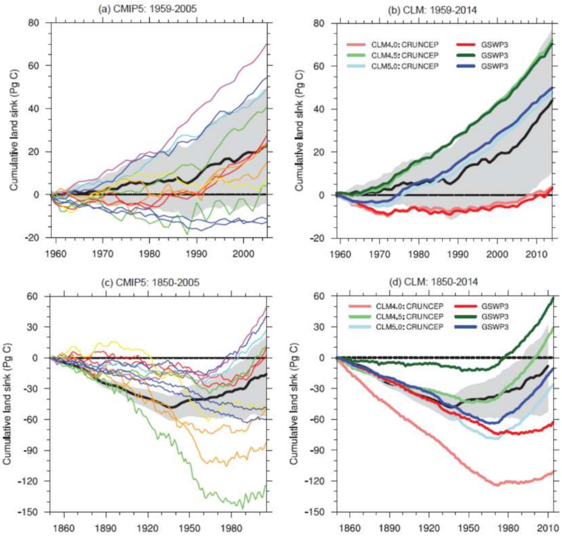
(Lawrence et al., 2019)

Land Model Performance Depends Strongly on Forcing



- Depending on the forcing used and the metric selected, different models may perform equally well
- ILAMB scores for CLM4, CLM4.5, and CLM5 forced with GSWP3 vs. CRUNCEP (above) and the cumulative land carbon sink for CMIP5 models vs. offline CLM (right).

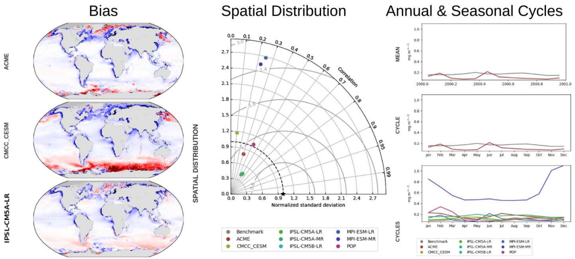
(Bonan et al., 2019)

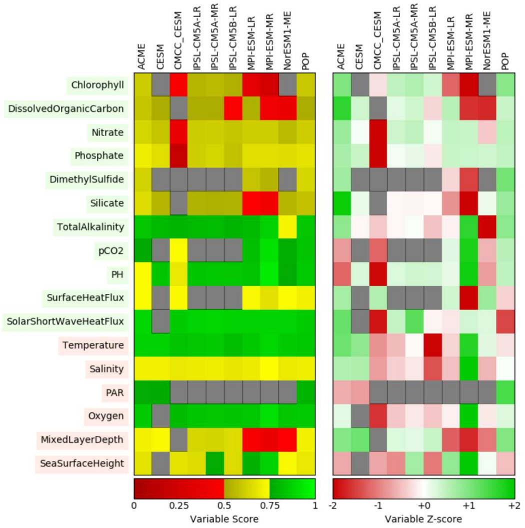


International Ocean Model Benchmarking (IOMB) Package

- Evaluates ocean biogeochemistry results compared with observations (global, point, ship tracks)
- Scores model performance across a wide range of independent benchmark data
- Leverages ILAMB code base, also runs in parallel
- Built on Python and open standards

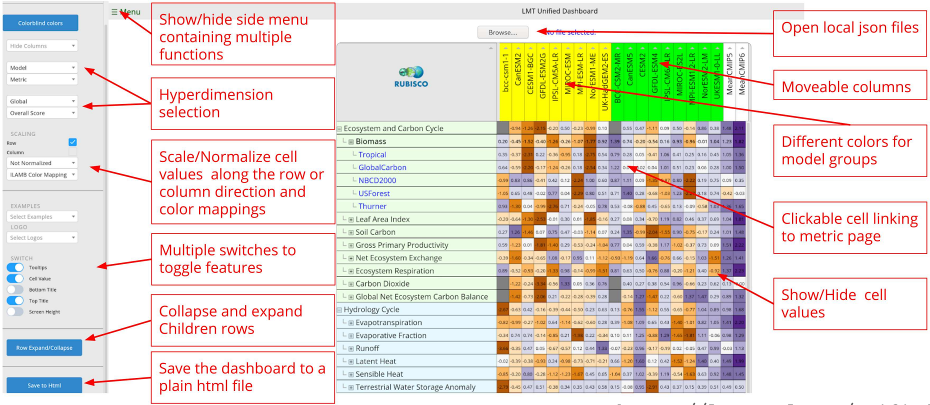
Chlorophyll / SeaWIFS





RUBISCO · Argonne National Laboratory · Brookhaven National Laboratory · Berkeley Lab · Los Alamos National Laboratory · NCAR · Oak Ridge National Laboratory · UCI · Michigan

Land Model Testbed (LMT) Unified Dashboard

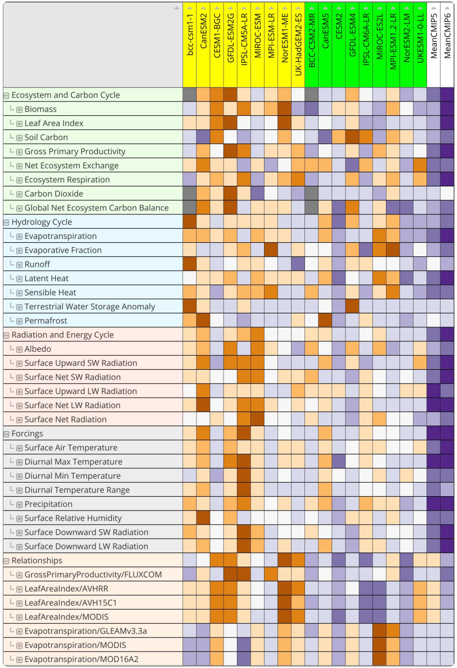


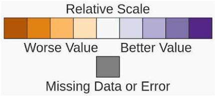
https://lmt.ornl.gov/unified-dashboard

- **Tooltips:** show scores when mouse hovers over the cells
- **Column hiding:** hides some models (columns) to focus on models of interest
- **Column sorting:** sort the scores along the columns/models to see the best metrics for each

CMIP5 vs. CMIP6 Land Models

- The performance of the CMIP6 suite of land models (on right with green headings) has improved over that of the CMIP5 suite of land models (on left with yellow headings)
- The multi-model mean (on far right with white headings) outperforms any single model for each suite of models
- The multi-model mean CMIP6 land model is the "best model" overall
- Why did CMIP6 land models improve over their CMIP5 progenitors?

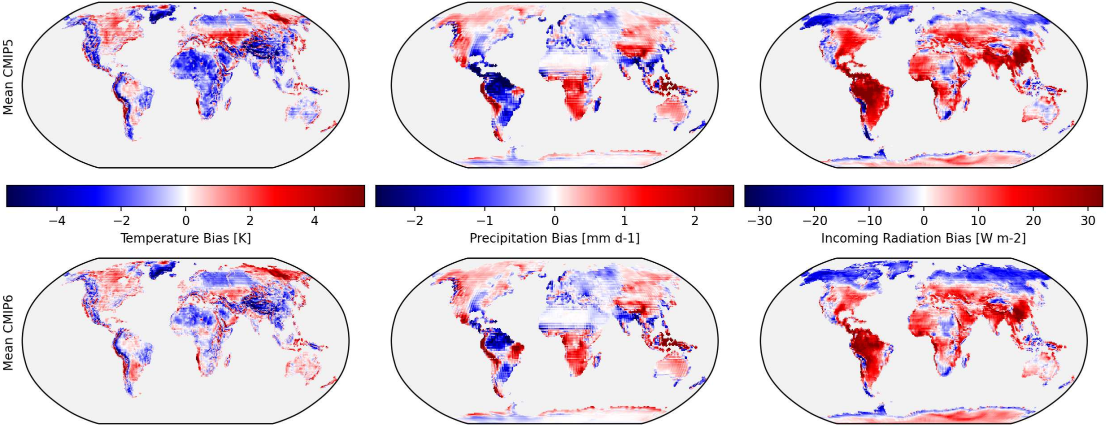




(Hoffman et al., in prep.)

Reasons for Land Model Improvements

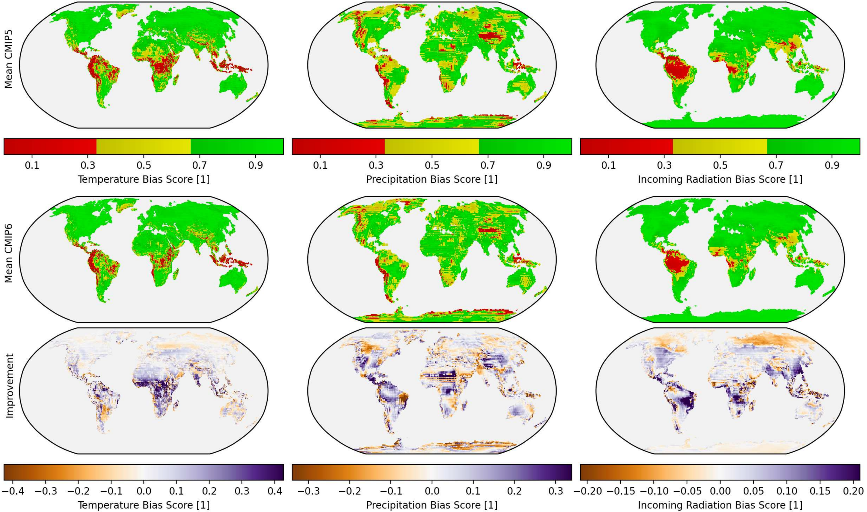
ESM improvements in climate forcings (temperature, precipitation, radiation) likely partially drove improvements exhibited by land carbon cycle models



(Hoffman et al., in prep.)

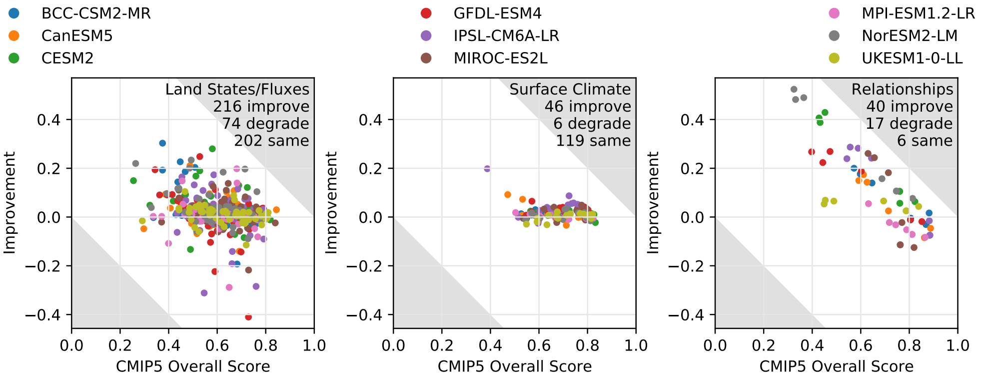
Reasons for Land Model Improvements

Differences in bias scores for temperature, precipitation, and incoming radiation were primarily positive, further indicating more realistic climate representation by the fully coupled ESMs



(Hoffman et al., in prep.)

Reasons for Land Model Improvements

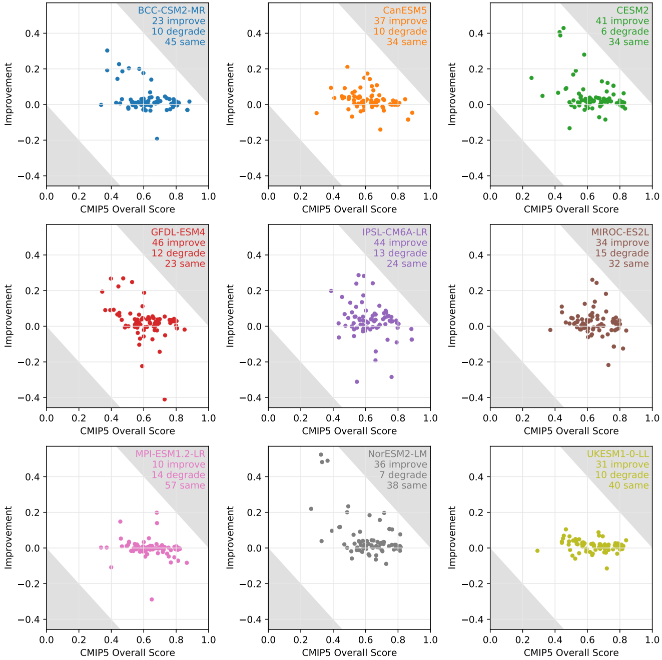


(Hoffman et al., in prep.)

Across all land models, scores for most state and flux variables improved (216) or remained nearly the same (202), although some were degraded (74). While atmospheric forcings from CMIP6 ESMs were improved over those from CMIP5 ESMs, the largest improvements were in land model **variable-to-variable relationships**, suggesting that increased land model development was also partially responsible for higher CMIP6 land model scores.

Improvements by Land Model

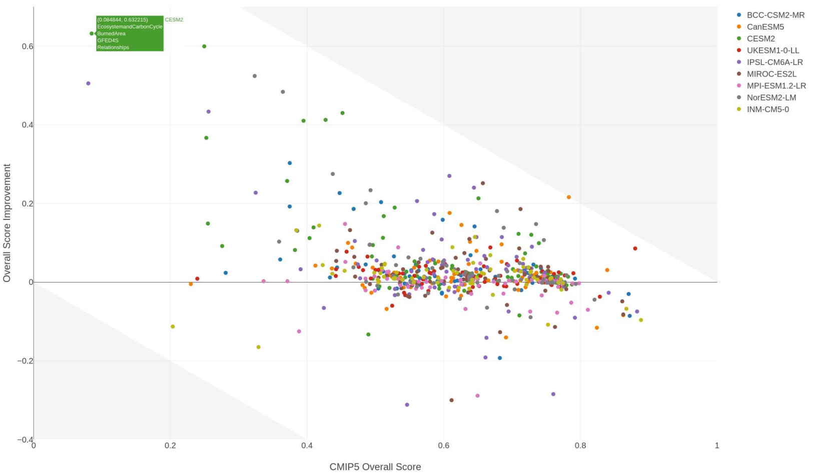
- Experience indicates that improvements in some model aspects will lead to degradation in some other aspects
- Here, all models except MPI-ESM1.2-LR showed more improvements than degredations
- CESM2 and NorESM2-LM had the largest ratio of improvements to degradations
- UKESM1-0-LL exhibited the smallest variation in scores between CMIP5 and CMIP6



(Hoffman et al., in prep.)

Interactive Exploration of Multi-Model Performance

https://www.ilamb.org/CMIP5v6/historical/chart.html



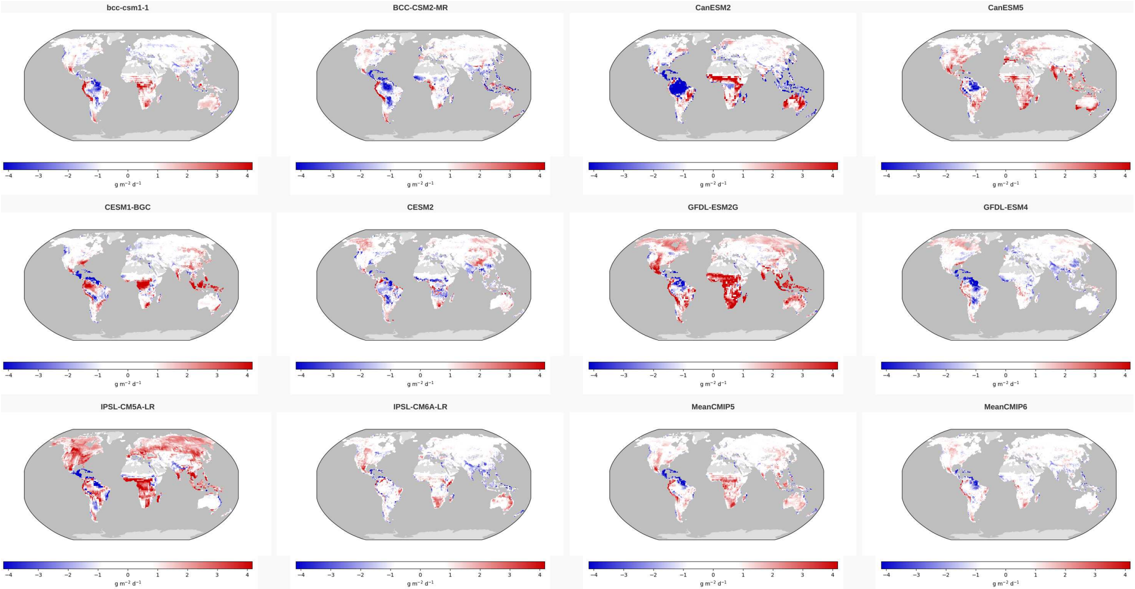
CMIP5 and CMIP6 Land Model Global GPP

	Download Data	Period Mean (original grids) [Pg yr-1]	Model Period Mean (intersection) [Pg yr-1]	Benchmark Period Mean (intersection) [Pg yr-1]	Model Period Mean (complement) [Pg yr-1]	Benchmark Period Mean (complement) [Pg yr-1]	Bias [g m-2 d-1]	RMSE [g m-2 d-1]	Phase Shift [months]	Bias Score [1]	RMSE Score [1]	Seasonal Cycle Score [1]	Spatial Distribution Score [1]	Overall Score [1]
Benchmark	[-]	114.												
bcc-csm1-1	[-]	123.	112.	114.	8.79	0.0945	0.238	1.51	1.01	0.484	0.435	0.830	0.955	0.628
BCC-CSM2-MR	[-]	114.	107.	113.	5.88	0.671	-0.0233	1.52	1.11	0.479	0.447	0.817	0.941	0.626
CanESM2	[-]	129.	117.	114.	9.54		0.0601	2.31	2.00	0.388	0.437	0.650	0.836	0.549
CanESM5	[-]	141.	128.	114.	10.1		0.730	1.87	1.60	0.449	0.418	0.710	0.948	0.589
CESM1-BGC	[-]	129.	123.	113.	5.55	0.660	0.379	1.66	1.20	0.426	0.468	0.765	0.889	0.603
CESM2	[-]	110.	104.	113.	5.57	0.642	-0.0542	1.62	1.32	0.458	0.466	0.774	0.933	0.619
GFDL-ESM2G	[-]	167.	152.	114.	12.4		1.26	2.78	1.38	0.377	0.288	0.735	0.897	0.517
GFDL-ESM4	[-]	105.	99.0	114.	6.18		-0.177	1.59	1.49	0.495	0.403	0.702	0.939	0.588
IPSL-CM5A-LR	[-]	165.	150.	113.	11.7	0.515	1.18	2.68	1.20	0.327	0.352	0.781	0.896	0.542
IPSL-CM6A-LR	[-]	115.	109.	113.	5.27	0.708	0.111	1.39	1.14	0.547	0.477	0.790	0.961	0.650
MeanCMIP5	[-]	121.	115.	114.	6.65		0.574	1.41	0.981	0.494	0.502	0.799	0.965	0.652
MeanCMIP6	[-]	116.	110.	114.	6.26		0.129	1.17	0.931	0.572	0.522	0.826	0.956	0.679
MIROC-ESM	[-]	129.	118.	102.	9.04	11.4	0.396	1.90	1.27	0.463	0.435	0.767	0.920	0.604
MIROC-ESM2L	[-]	116.	104.	113.	9.90	0.119	-0.0111	1.95	1.99	0.409	0.379	0.628	0.920	0.543
MPI-ESM-LR	[-]	169.	159.	104.	8.91	9.81	1.36	2.36	1.29	0.402	0.371	0.715	0.930	0.558
MPI-ESM1.2-LR	[-]	141.	133.	104.	6.89	9.81	0.725	2.06	1.13	0.409	0.393	0.769	0.925	0.578
NorESM1-ME	[-]	129.	120.	114.	7.82		0.386	1.86	1.25	0.387	0.456	0.761	0.856	0.583
NorESM2-LM	[-]	107.	97.5	114.	7.59		-0.0828	1.63	1.31	0.443	0.472	0.791	0.938	0.623
UK-HadGEM2-ES	[-]	137.	130.	113.	6.93	0.848	0.602	2.01	1.10	0.389	0.388	0.820	0.855	0.568
UKESM1-0-LL	[-]	126.	119.	113.	7.06	0.825	0.387	1.77	1.16	0.436	0.419	0.791	0.924	0.598

- Most models of the same lineage improved in various characteristics between CMIP5 and CMIP6
- The MeanCMIP5 and MeanCMIP6 models perform the best

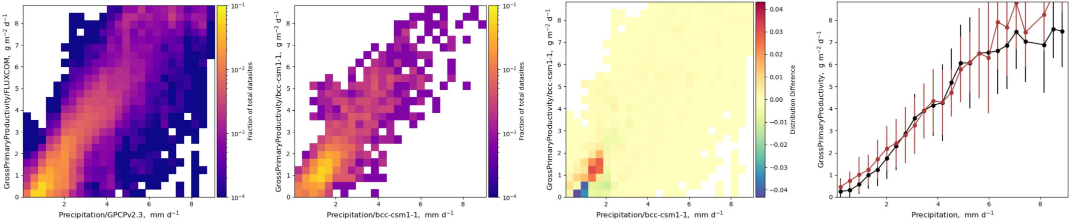
(Hoffman et al., in prep.)

Spatial Distribution of Global GPP Biases



Relationships of Global GPP with Precipitation and Temperature

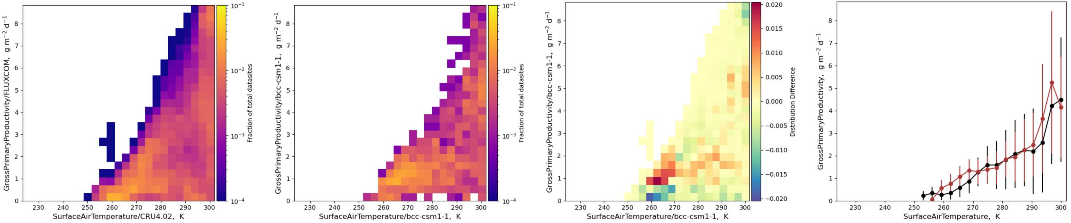
Precipitation/GPCPv2.3



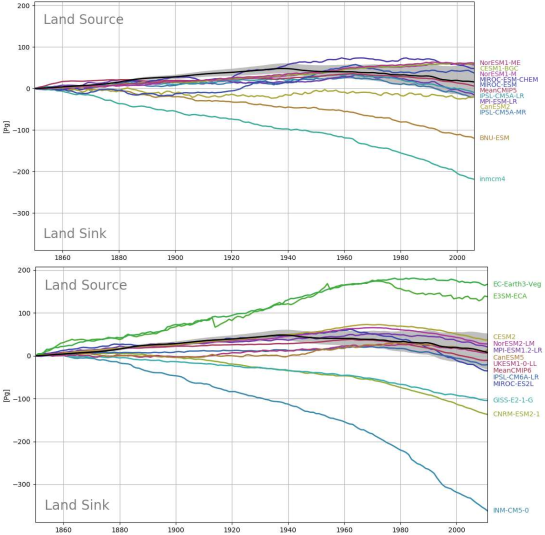
SurfaceDownwardSWRadiation/CERESed4.1

SurfaceNetSWRadiation/CERESed4.1

SurfaceAirTemperature/CRU4.02



Land Model Spread in Net Ecosystem Carbon Balance

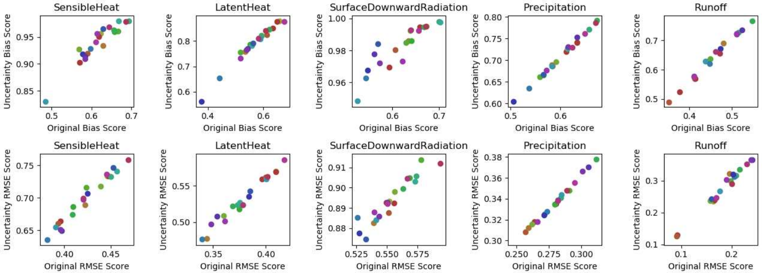


- The spread in the net ecosystem carbon balance increased between CMIP5 and CMIP6
  - CMIP5 at 2005: −215 Pg to 75 Pg → 290 Pg
  - CMIP6 at 2010: −360 Pg to 175 Pg → 535 Pg
- However, the range from most multi-generation models was reduced

(Hoffman et al., in prep.)

Addressing Observational Uncertainty

- Few observational datasets provide complete uncertainties
- ILAMB uses multiple datasets for most variables and allows users to weight them according to a rubric of uncertainty, scale mismatch, etc.
- ILAMB can also use:
  - Full spatial/temporal uncertainties provided with data
  - Fixed, expert-derived uncertainty for a dataset
  - Uncertainties derived from combining multiple datasets

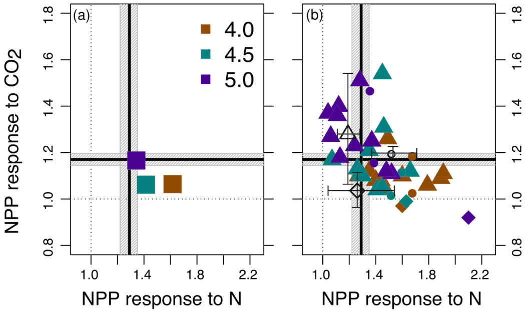


(Collier et al., in prep.)

- Experiments with CLASS self-consistent data (Hobeichi et al., 2020) demonstrates that while scores shift, including uncertainty rarely alters the rank ordering of models (figure)

Beyond Static Benchmarking

- To better support model development verification, we need to incorporate metrics from manipulative experiments
  - Simulated effect sizes of nitrogen versus CO_2 enrichment on rates of net primary production (NPP) calculated (a) globally or (b) for each plant functional type in CLM4, 4.5, and 5
  - Observational constraints for N response and CO_2 response are shown with vertical and horizontal polygons (mean ±95% confidence intervals)
  - In (b), observed (open symbols) and simulated (filled symbols) effect sizes of individual PFTs for woody vegetation, C_3 grasses, and C_4 grasses (triangles, circles, and diamonds, respectively)



(Wieder et al., 2019)

- Much more work is needed to foster land model ensemble simulations and benchmarking, including land model testbeds, diurnal and seasonal metrics, new synthesis datasets, ...

Conclusions and Future Research

- CMIP6 land models performed better than CMIP5 land models due to **(1) improved climate forcing from fully coupled ESMs** and **(2) improved process representation**
- **Variable-to-variable relationships** exhibited the largest improvements for some models
- CMIP6 model results are more valuable for impact and adaptation/mitigation analysis
- Land model performance depends strongly on imposed climate forcing
- Incorporating observational uncertainty in ILAMB analysis increases model scores, but rarely alters the rank ordering of models
- Model improvements in mean states and fluxes may not result in reduced uncertainty or projected model spread
- Upon further examination, will improved multi-model performance result in reduced spread in feedback sensitivities, projected land carbon storage, and future climate change?
- Can ILAMB scores be used to weight contributions to multi-model means to reduce contemporary biases, reduce projected uncertainties, or alter expected mitigation targets?

Acknowledgments

This research was supported by the *Reducing Uncertainties in Biogeochemical Interactions through Synthesis and Computation Science Focus Area (RUBISCO SFA)*, which is sponsored by the Regional and Global Model Analysis (RGMA) activity of the Earth & Environmental Systems Modeling (EESM) Program in the Climate and Environmental Sciences Division (CESD) of the Office of Biological and Environmental Research (BER) in the US Department of Energy (DOE) Office of Science. Additional support was provided by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC, for the US Department of Energy under contract DE-AC05-00OR22725.

We acknowledge the World Climate Research Programme (WCRP), which, through its Working Group on Coupled Modelling (WGCM), coordinated and promoted the Coupled Model Intercomparison Project phase 6 (CMIP6). We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies that support CMIP6 and ESGF. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. We thank DOE's RGMA activity, Data Management Program, and NERSC for making this coordinated CMIP6 analysis activity possible.

References

G. B. Bonan, D. L. Lombardozzi, W. R. Wieder, K. W. Oleson, D. M. Lawrence, F. M. Hoffman, and N. Collier. Model structure and climate data uncertainty in historical simulations of the terrestrial carbon cycle (1850–2014). *Global Biogeochem. Cycles*, 33(10):1310–1326, Oct. 2019. doi:10.1029/2019GB006175.

K. S. Carslaw, L. A. Lee, L. A. Regayre, and J. S. Johnson. Climate models are uncertain, but we can do something about it. *Eos Trans. AGU*, 99, Feb. 2018. doi:10.1029/2018EO093757.

N. Collier, F. M. Hoffman, D. M. Lawrence, G. Keppel-Aleks, C. D. Koven, W. J. Riley, M. Mu, and J. T. Randerson. The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. *J. Adv. Model. Earth Sy.*, 10(11):2731–2754, Nov. 2018. doi:10.1029/2018MS001354.

V. Eyring, P. M. Cox, G. M. Flato, P. J. Gleckler, et al. Taking climate model evaluation to the next level. *Nat. Clim. Change*, 9(2):102–110, Feb. 2019. doi:10.1038/s41558-018-0355-y.

S. Hobeichi, G. Abramowitz, and J. Evans. Conserving land–atmosphere synthesis suite (class). *J. Clim.*, 33(5):1821–1844, Mar. 2020. doi:10.1175/JCLI-D-19-0036.1.

F. M. Hoffman, C. D. Koven, G. Keppel-Aleks, D. M. Lawrence, W. J. Riley, J. T. Randerson, et al. International Land Model Benchmarking (ILAMB) 2016 workshop report. Technical Report DOE/SC-0186, U.S. Department of Energy, Office of Science, Germantown, Maryland, USA, Apr. 2017.

R. Knutti and J. Sedláček. Robustness and uncertainties in the new CMIP5 climate model projections. *Nat. Clim. Change*, 3(4):369–373, Apr. 2013. doi:10.1038/nclimate1716.

D. M. Lawrence, R. A. Fisher, C. D. Koven, K. W. Oleson, S. C. Swenson, et al. The Community Land Model version 5: Description of new features, benchmarking, and impact of forcing uncertainty. *J. Adv. Model. Earth Sy.*, 11(12):4245–4287, Dec. 2019. doi:10.1029/2018MS001583.

Y. Q. Luo, J. T. Randerson, et al. A framework for benchmarking land models. *Biogeosci.*, 9(10):3857–3874, Oct. 2012. doi:10.5194/bg-9-3857-2012.

J. T. Randerson, F. M. Hoffman, P. E. Thornton, N. M. Mahowald, K. Lindsay, Y.-H. Lee, C. D. Nevison, S. C. Doney, G. Bonan, R. Stöckli, C. Covey, S. W. Running, and I. Y. Fung. Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models. *Glob. Change Biol.*, 15(9):2462–2484, Sept. 2009. doi:10.1111/j.1365-2486.2009.01912.x.

K. E. Taylor. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res. Atmos.*, 106(D7):7183–7192, Apr. 2001. doi:10.1029/2000JD900719.

W. R. Wieder, D. M. Lawrence, R. A. Fisher, et al. Beyond static benchmarking: Using experimental manipulations to evaluate land model assumptions. *Global Biogeochem. Cycles*, 33:1289–1309, Oct. 2019. doi:10.1029/2018GB006141.