U.S. DEPARTMENT OF ENERGY

Office of Science

International Land Model Benchmarking (ILAMB)

Forrest M. Hoffman[1,2], Nathan Collier[1], Mingquan Mu[3], Min Xu[1], Weiwei Fu[3], Cheng-En Yang[2,1], Gretchen Keppel-Aleks[4], David M. Lawrence[5], Charles D. Koven[6], William J. Riley[6], and James T. Randerson[3]

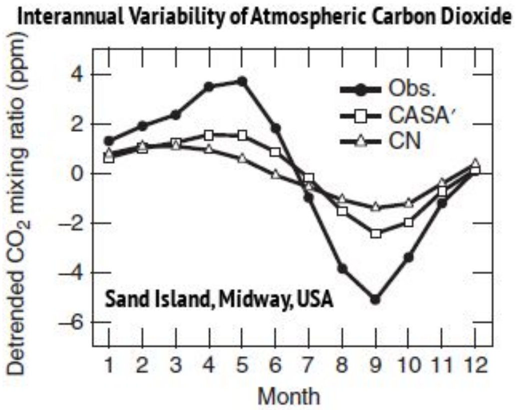
[1]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2]University of Tennessee, Knoxville, TN, USA
[3]University of California, Irvine, CA, USA
[4]University of Michigan, Ann Arbor, MI, USA
[5]National Center for Atmospheric Research, Boulder, CO, USA
[6]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Coupling of Land and Atmospheric Subgrid Parameterizations (CLASP) Project Meeting

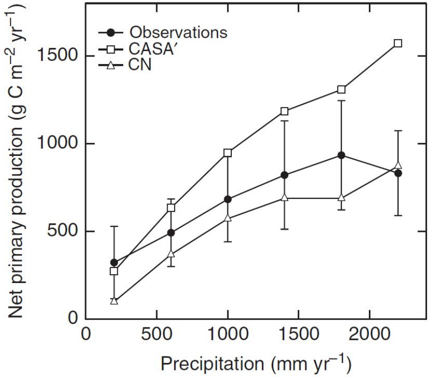
August 5, 2020

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 a benchmark **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



(Randerson et al., 2009)

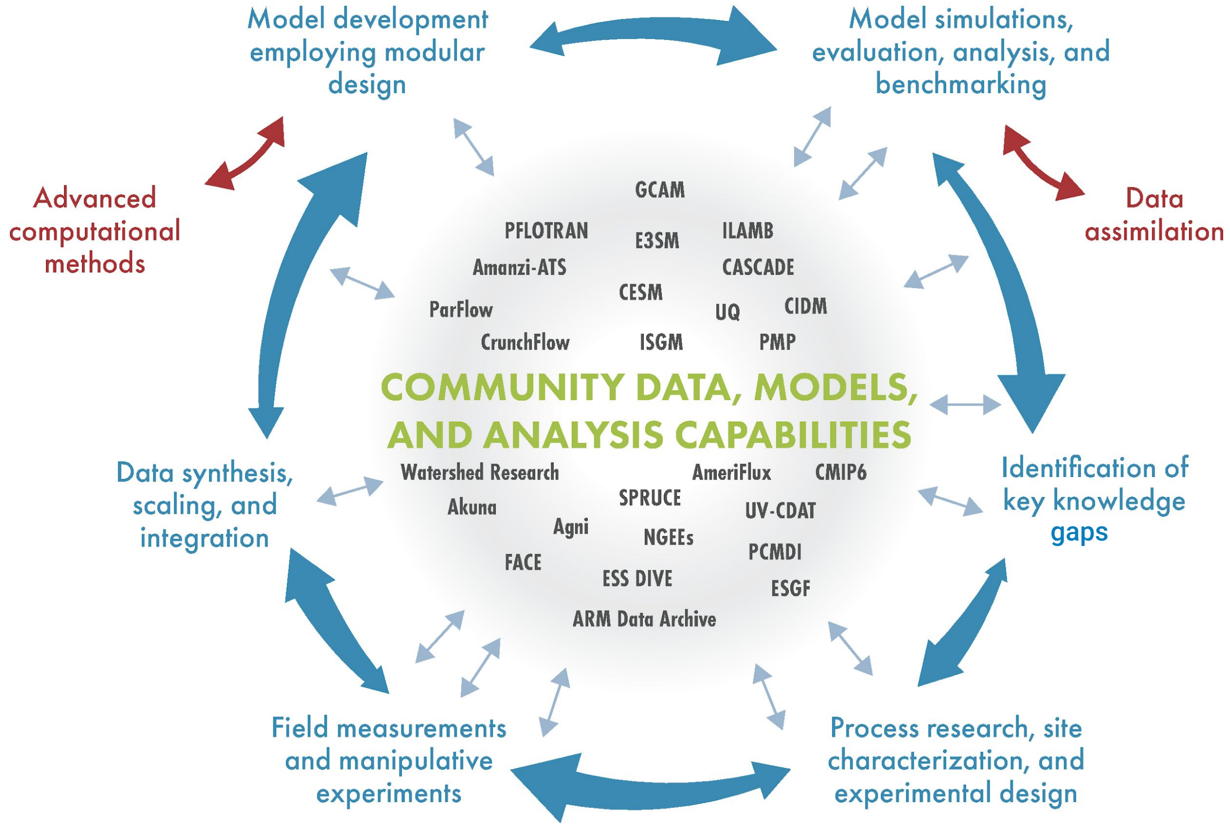
Models may reproduce correct responses over only a limited range of forcing variables

Argonne National Laboratory · Brookhaven National Laboratory · Berkeley Lab · Los Alamos National Laboratory · NCAR National Center for Atmospheric Research · Oak Ridge National Laboratory · UCI · Michigan

Why Benchmark Models?

- To **quantify and reduce uncertainties** in carbon cycle feedbacks to improve projections of future climate change (Eyring et al., 2019; Collier et al., 2018)
- 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

DOE's Model-Data-Experiment Enterprise (aka MODEX)

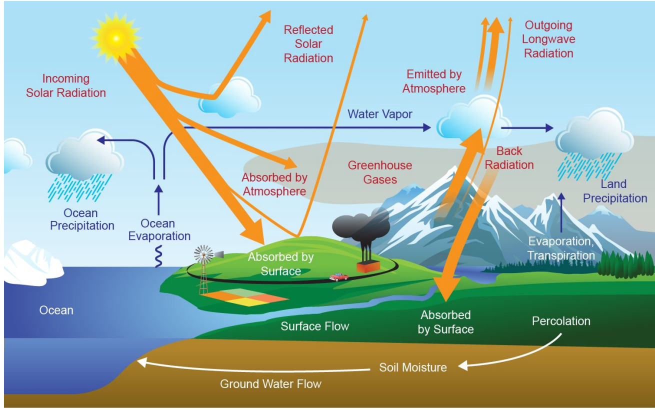


RUBISCO

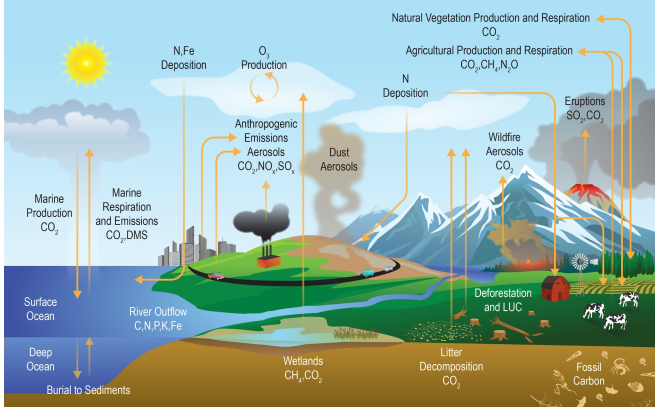
What is ILAMB?

A community coordination activity created 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 Earth system modeling communities** in the design of new model tests and new measurement programs
- **Support the design and development of open source benchmarking tools** (Luo et al., 2012)



Energy and Water Cycles



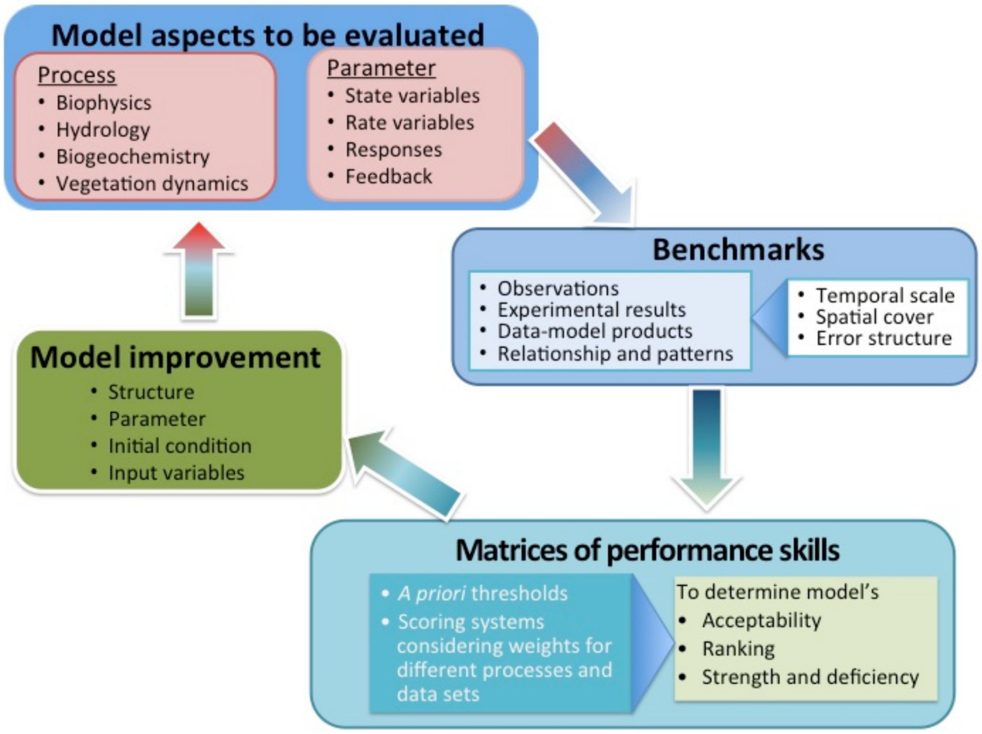
Carbon and Biogeochemical Cycles

Argonne National Laboratory · Brookhaven National Laboratory · Berkeley Lab · Los Alamos National Laboratory · NCAR National Center for Atmospheric Research · Oak Ridge National Laboratory · UCI · Michigan

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

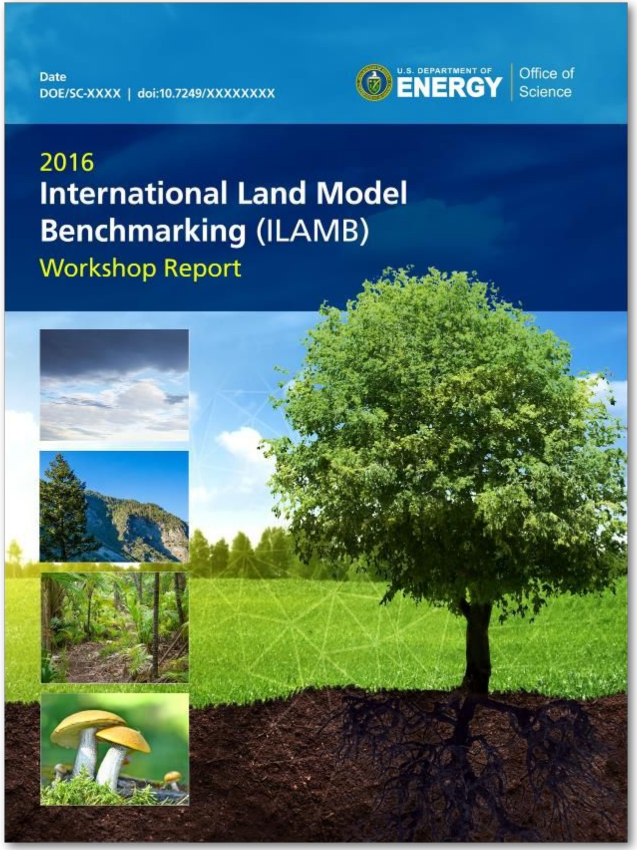
A Framework for Benchmarking Land Models

- 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)

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



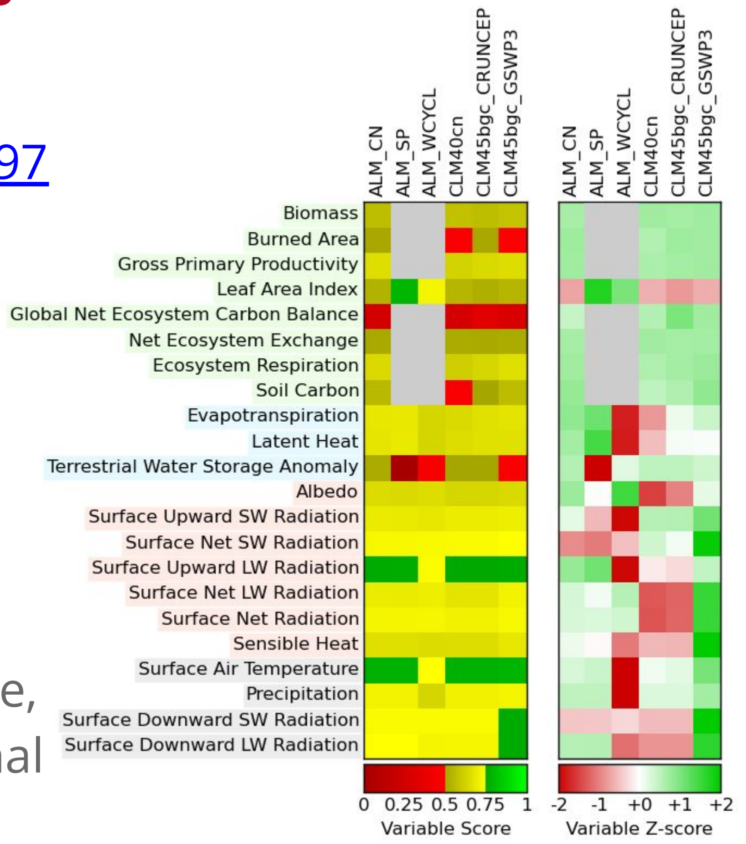
Third ILAMB Workshop was held May 16–18, 2016

- Workshop Goals
  - Design of new metrics for model benchmarking
  - Model Intercomparison Project (MIP) evaluation needs
  - Model development, testbeds, and workflow processes
  - Observational data sets 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

Argonne NATIONAL LABORATORY | BROOKHAVEN NATIONAL LABORATORY | BERKELEY LAB | Los Alamos NATIONAL LABORATORY | NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH | OAK RIDGE National Laboratory | UCI | MICHIGAN

Development of ILAMB Packages

- **ILAMBv1** released at 2015 AGU Fall Meeting Town Hall, doi:10.18139/ILAMB.v001.00/1251597
- **ILAMBv2** released at 2016 ILAMB Workshop, doi:10.18139/ILAMB.v002.00/1251621
- Open Source software freely distributed
- Routinely used for E3SM and CESM evaluation during development
- Employed to evaluate CMIP5 models
- Models are scored based on statistical comparisons (bias, RMS error, phase, amplitude, spatial distribution, Taylor scores) and functional response metrics



ILAMB Produces Diagnostics and Scores Models

- ILAMB generates a top-level **portrait plot** of models scores
- For every variable and dataset, ILAMB can automatically produce
  - **Tables** containing individual metrics and metric scores (when relevant to the data), including
    - Benchmark 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** ($S_{overall}$) →

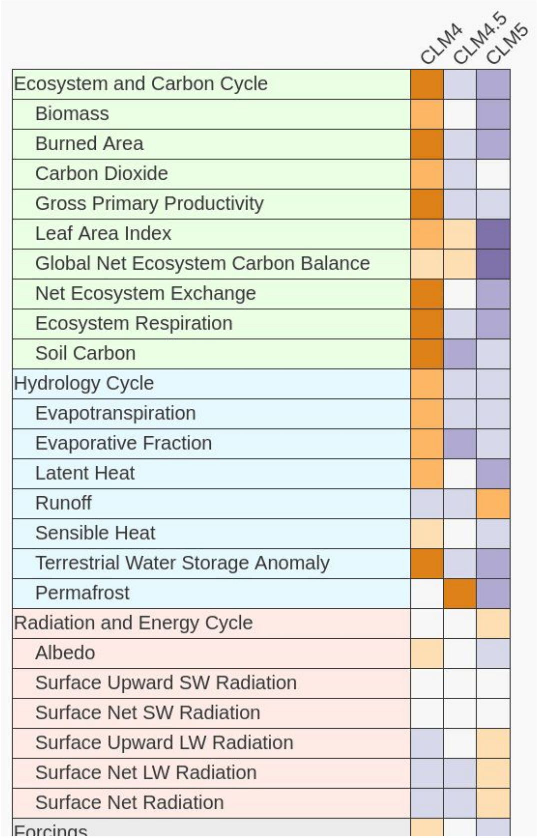
$$S_{\text{overall}} = \frac{S_{\text{bias}} + 2S_{\text{rmse}} + S_{\text{phase}} + S_{\text{iav}} + S_{\text{dist}}}{1+2+1+1+1}$$

  - **Graphical diagnostics**
    - Spatial contour maps
    - Time series line plots
    - Spatial Taylor diagrams (Taylor, 2001)
- Similar **tables** and **graphical diagnostics** for functional relationships

ILAMBv2.5 Package Current Variables

- **Biogeochemistry:** Biomass (Contiguous US, Pan Tropical Forest), Burned area (GFED3), CO_2 (NOAA GMD, Mauna Loa), Gross primary production (Fluxnet, GBAF), Leaf area index (AVHRR, MODIS), Global net ecosystem carbon balance (GCP, Khatiwala/Hoffman), Net ecosystem exchange (Fluxnet, GBAF), Ecosystem Respiration (Fluxnet, GBAF), Soil C (HWSD, NCSCDv22, Koven)
- **Hydrology:** Evapotranspiration (GLEAM, MODIS), Evaporative fraction (GBAF), Latent heat (Fluxnet, GBAF, DOLCE), Runoff (Dai, LORA), Sensible heat (Fluxnet, GBAF), Terrestrial water storage anomaly (GRACE), Permafrost (NSIDC)
- **Energy:** Albedo (CERES, GEWEX.SRB), Surface upward and net SW/LW radiation (CERES, GEWEX.SRB, WRMC.BSRN), Surface net radiation (CERES, Fluxnet, GEWEX.SRB, WRMC.BSRN)
- **Forcing:** Surface air temperature (CRU, Fluxnet), Diurnal max/min/range temperature (CRU), Precipitation (CMAP, Fluxnet, GPCC, GPCP2), Surface relative humidity (ERA), Surface down SW/LW radiation (CERES, Fluxnet, GEWEX.SRB, WRMC.BSRN)

ILAMB Assessing Several Generations of CLM

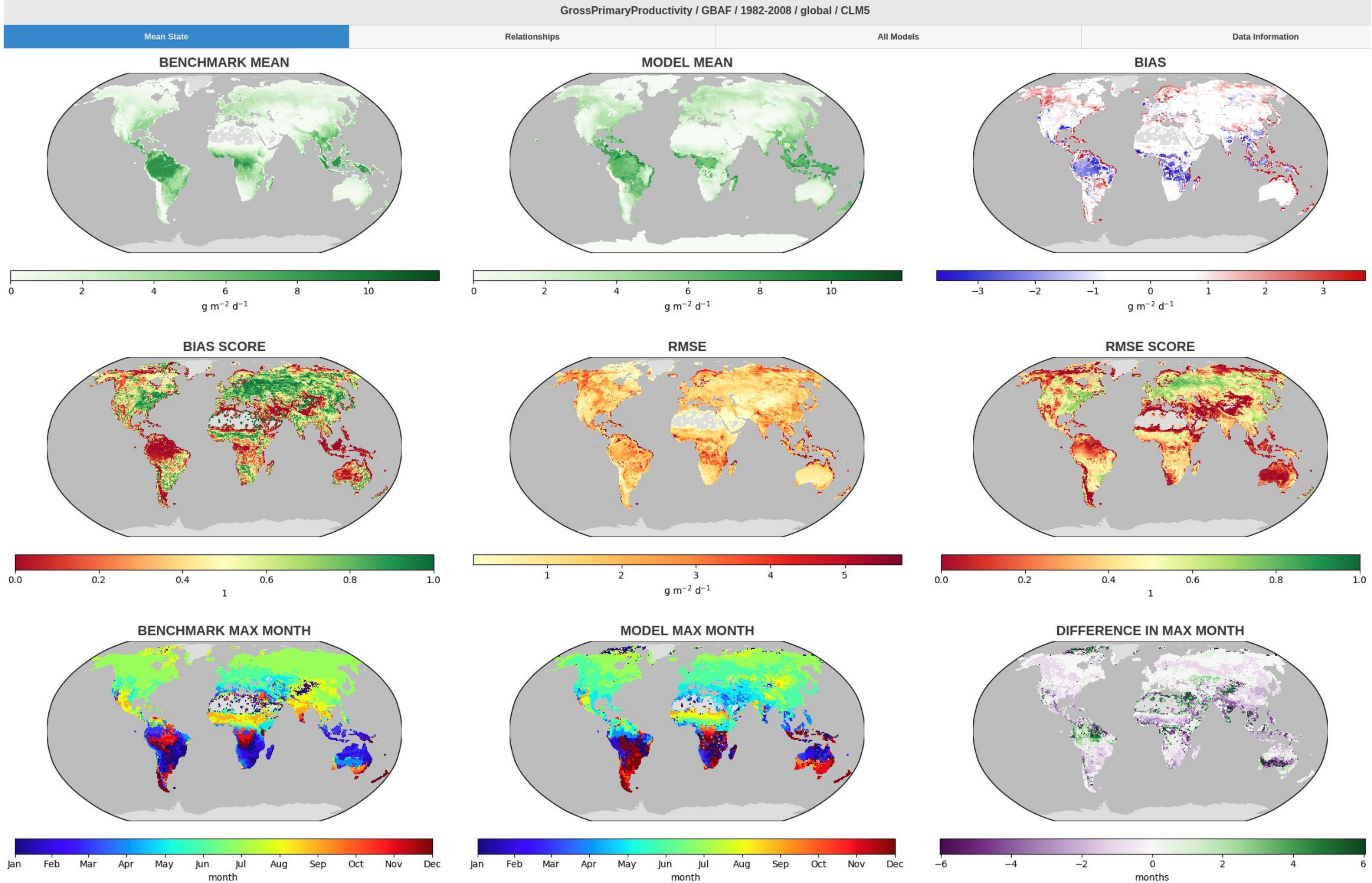


- Improvements in mechanistic treatment of hydrology, ecology, and land use with much more complexity in Community Land Model version 5 (CLM5)
- Simulations improved even with enhanced complexity
- Observational datasets not always self-consistent
- Forcing uncertainty confounds assessment of model development

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

(Lawrence et al., 2019)

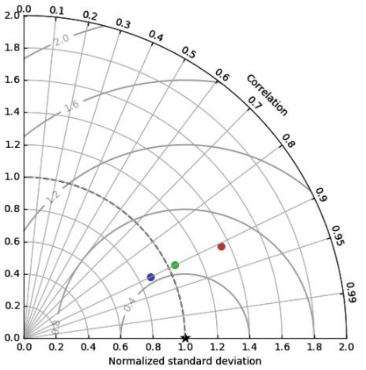
ILAMB Graphical Diagnostics



ILAMB Graphical Diagnostics

RUBISCO

SPATIAL TAYLOR DIAGRAM



MODEL COLORS

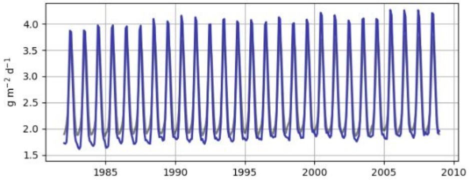
- Benchmark
- CLM4
- CLM4.5
- CLM5

Spatially integrated regional mean

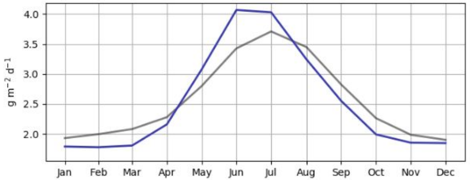
MODEL COLORS

- Benchmark
- CLM4
- CLM4.5
- CLM5

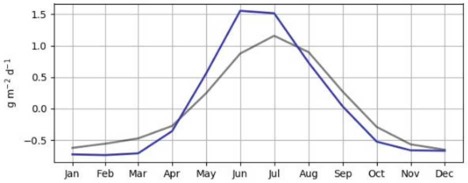
REGIONAL MEAN



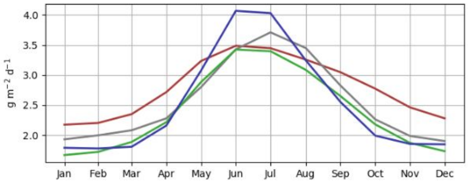
ANNUAL CYCLE



MONTHLY ANOMALY



ANNUAL CYCLE



Argonne NATIONAL LABORATORY
BROOKHAVEN NATIONAL LABORATORY
BERKELEY LAB
Los Alamos NATIONAL LABORATORY
NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
OAK RIDGE National Laboratory
UCI
MICHIGAN

ILAMB Graphical Diagnostics

New PEcAn-ILAMB site-level diagnostics

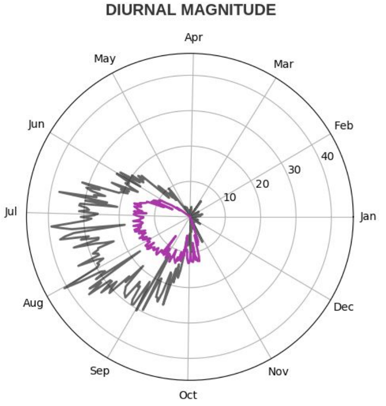
GrossPrimaryProductivity / AMF_US-WCr / global / CLM5

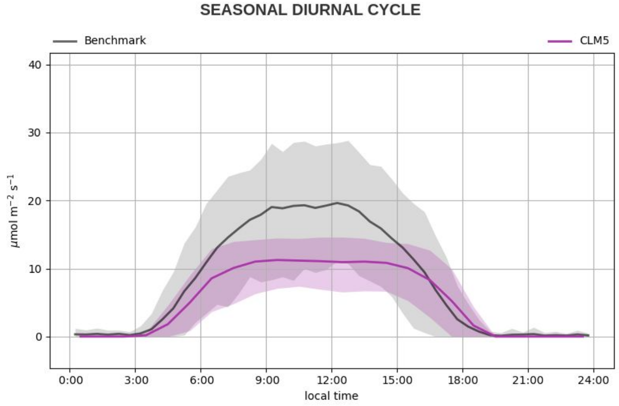
Mean State | All Models | Data Information

	Download Data	Number of Years [1]	Season Length [d]	Diurnal Peak Timing [h]	Mean Season Uptake [1e-6 mol m-2 s-1]	Season Beginning [d]	Season Ending [d]	Diurnal Peak Timing Score [1]	Diurnal Uptake Score [1]	Season Beginning Score [1]	Season Ending Score [1]	Season Strength Score [1]	Overall Score [1]
Benchmark	[-]	14.0	120.	10.6	8.07	144.	264.						
CLM5	[-]	14.0	153.	10.4	5.62	138.	291.	0.896	0.756	0.681	0.361	0.668	0.672
ED2a	[-]	2.00	140.	12.8	3.77	138.	278.	0.750	0.570	0.712	0.616	0.601	0.650
ED2b	[-]	6.00	161.	10.8	4.35	120.	281.	0.910	0.610	0.370	0.517	0.630	0.607
SIPNETa	[-]	7.00	136.	9.79	6.86	145.	281.	0.908	0.818	0.801	0.510	0.835	0.774
SIPNETb	[-]	2.00	178.	4.50	5.76	104.	282.	0.521	0.670	0.205	0.400	0.850	0.529
SIPNETc	[-]	7.00	128.	8.64	8.81	144.	273.	0.830	0.769	0.811	0.716	0.736	0.773

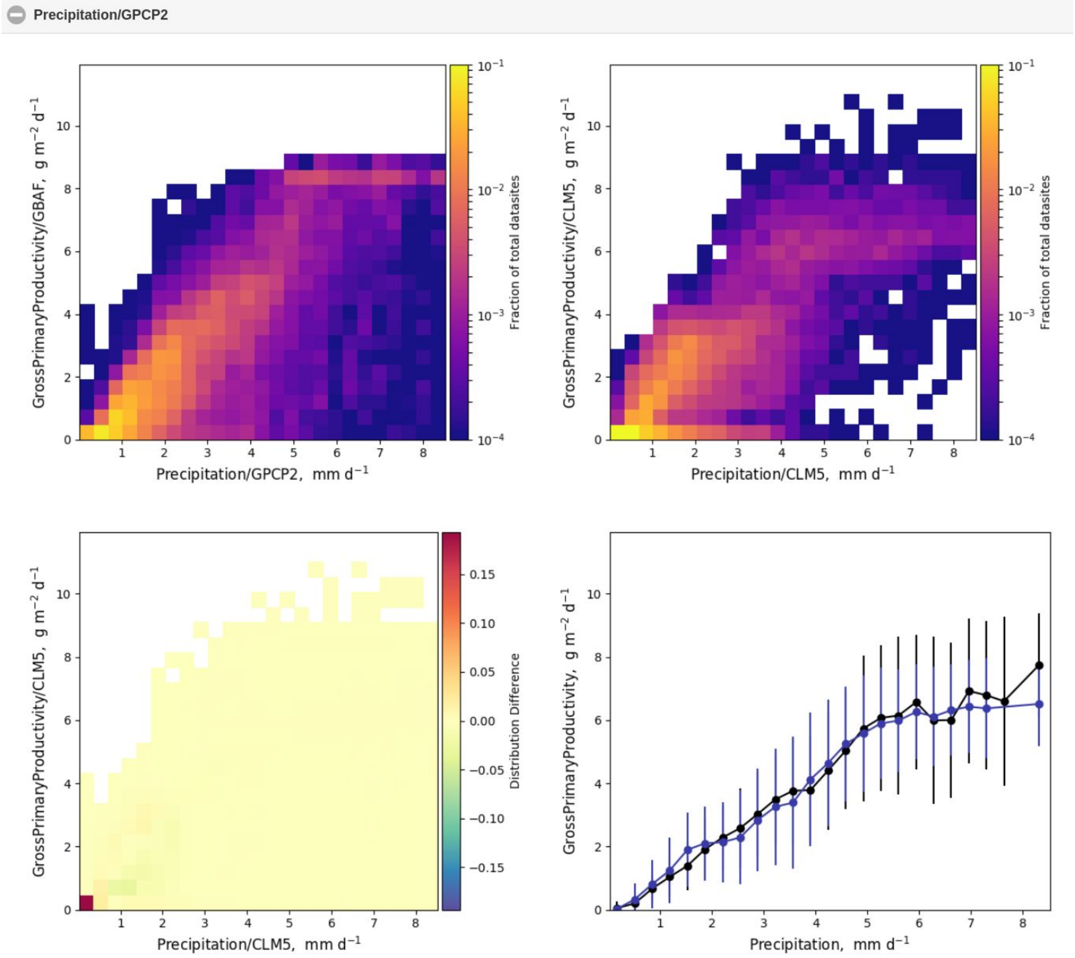
1999

2000

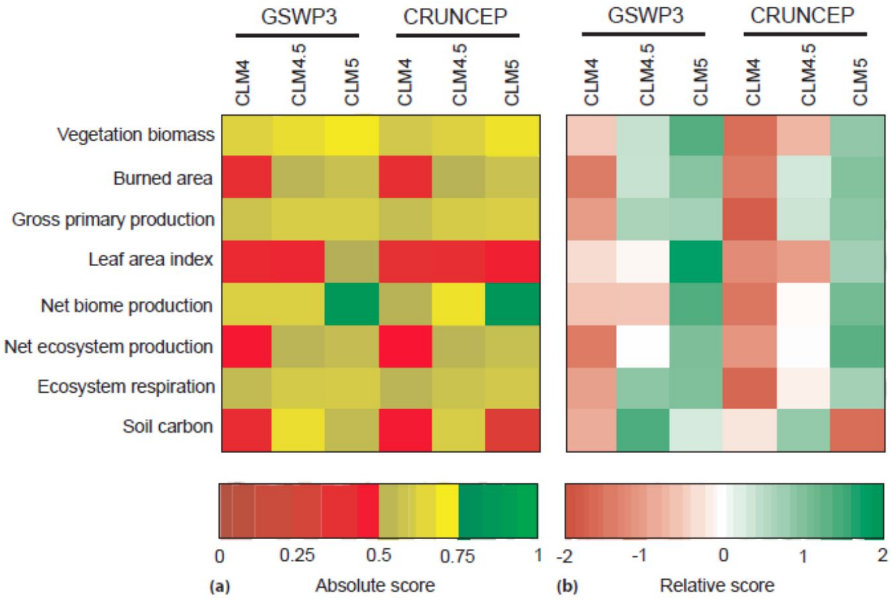


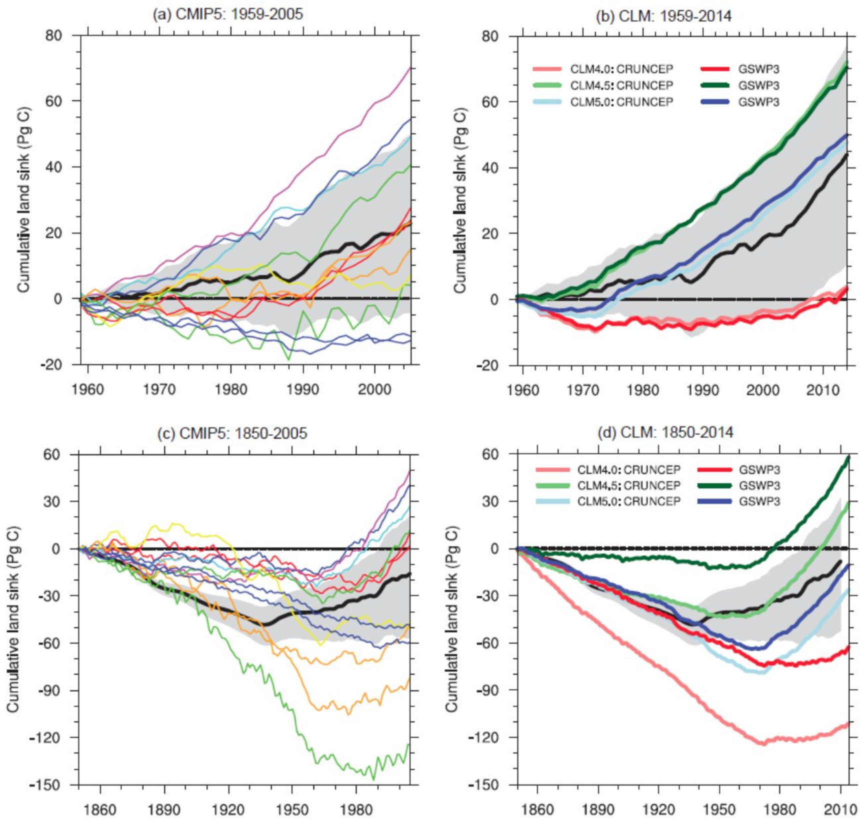


Variable-to-Variable Comparisons



Land Model Performance Depends Strongly on Forcing



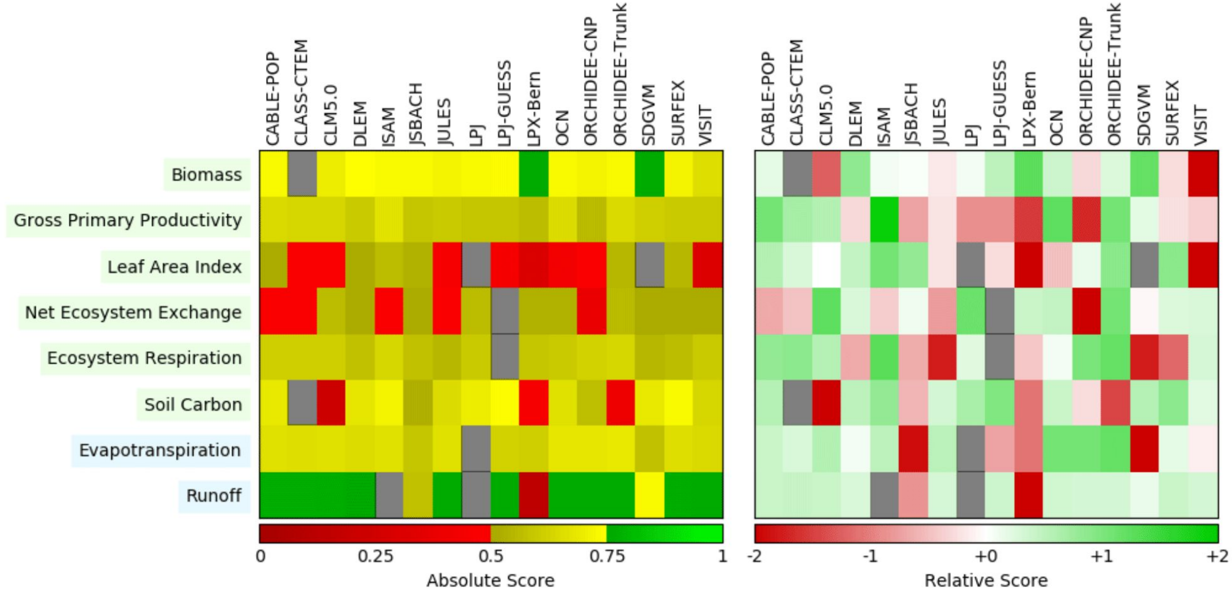


ILAMB performance for CLM4, CLM4.5, and CLM5 forced with GSWP3 vs. CRUNCEP (left) and the cumulative land carbon sink for CMIP5 vs. CLM offline models (right).

Bonan et al. (2019)

Global Carbon Budget 2018 - TRENDY Models

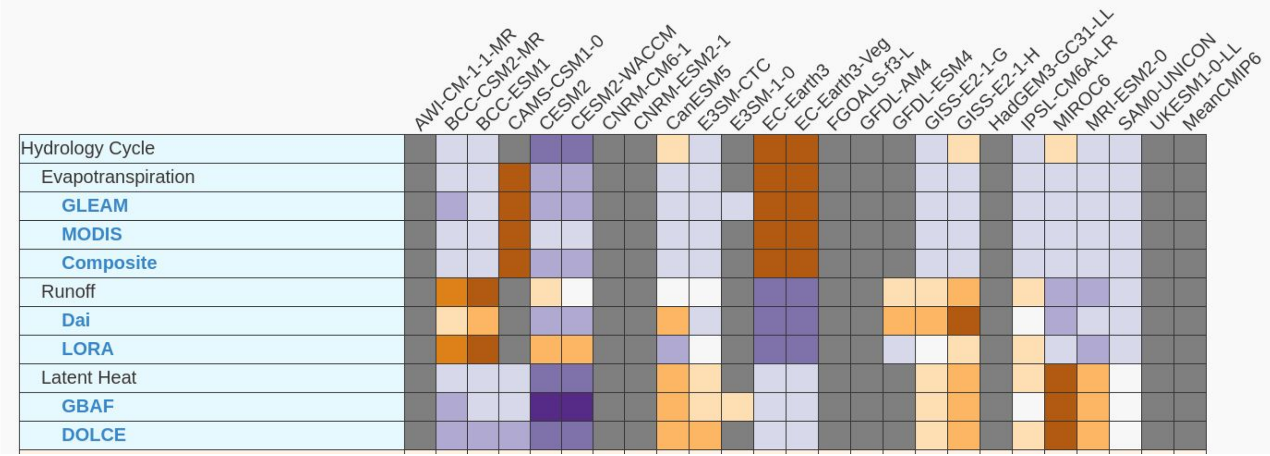
Evaluation of the DGVMs using the International Land Model Benchmarking system (ILAMB; Collier et al., 2018) (left) absolute skill scores and (right) skill scores relative to other models for a subset of ILAMB variables.



Le Quéré et al. (2018)

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. (Table 1, Collier et al., 2018)
- ILAMB can also use:
  - Full spatial/temporal uncertainties provided with the data
  - Fixed, expert-derived uncertainty for a dataset
  - Uncertainties derived from combining multiple datasets



(Collier et al., in prep)

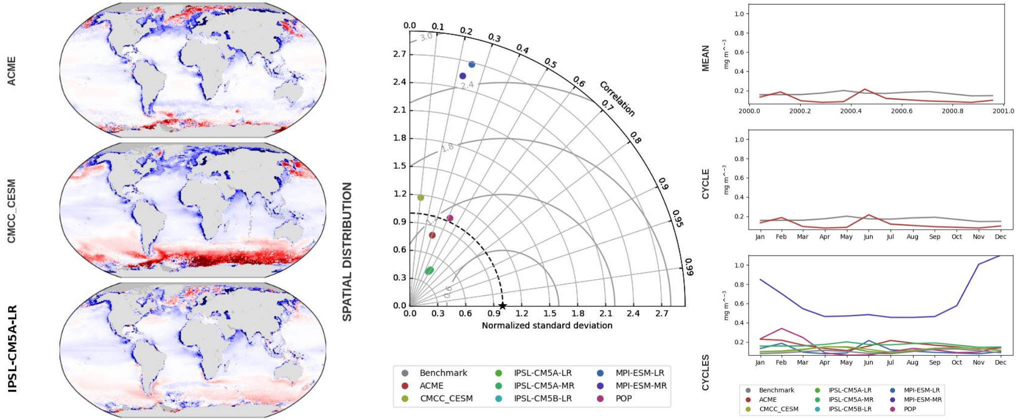
RUBISCO · Argonne National Laboratory · Brookhaven National Laboratory · Berkeley Lab · Los Alamos National Laboratory · NCAR National Center for Atmospheric Research · Oak Ridge National Laboratory · UCI · Michigan

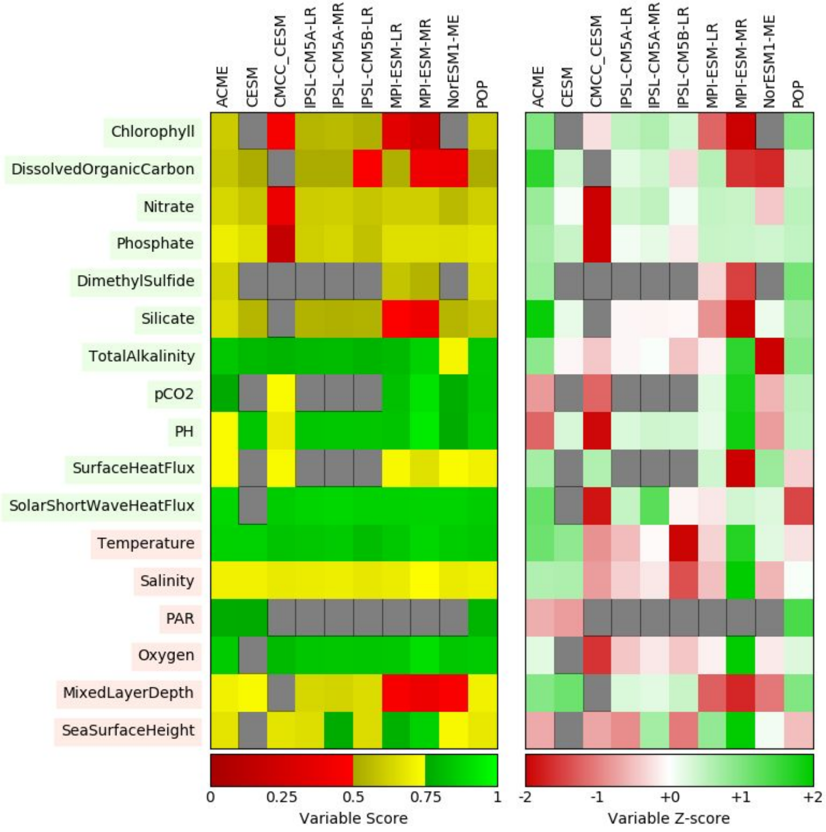
International Ocean Model Benchmarking (IOMB) Package

- 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
- Is also open source and will be released soon

Chlorophyll / SeaWIFS

Bias | Spatial Distribution | Annual & Seasonal Cycles

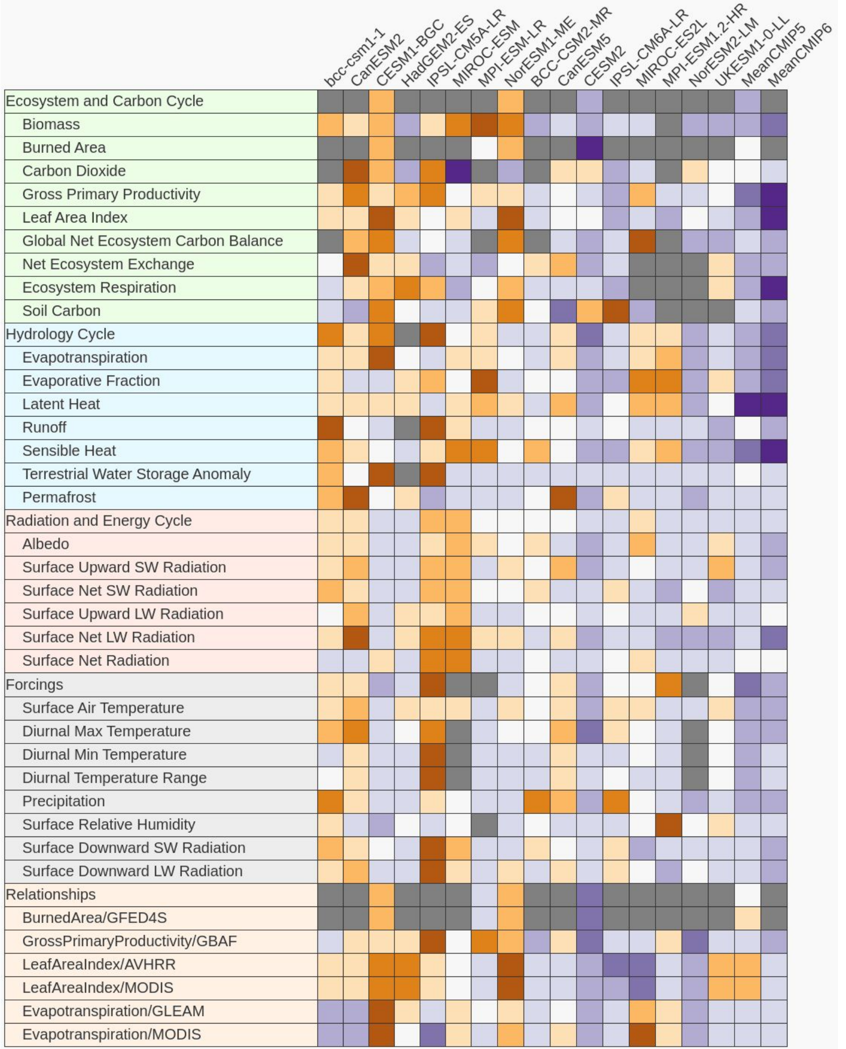




CMIP5 vs. CMIP6 Models

- The CMIP6 suite of land models (right) has improved over the CMIP5 suite of land models (left)
- The multi-model mean outperforms any single model for each suite of models
- The multi-model mean CMIP6 land model is the “best” model overall

(Hoffman et al., in prep)

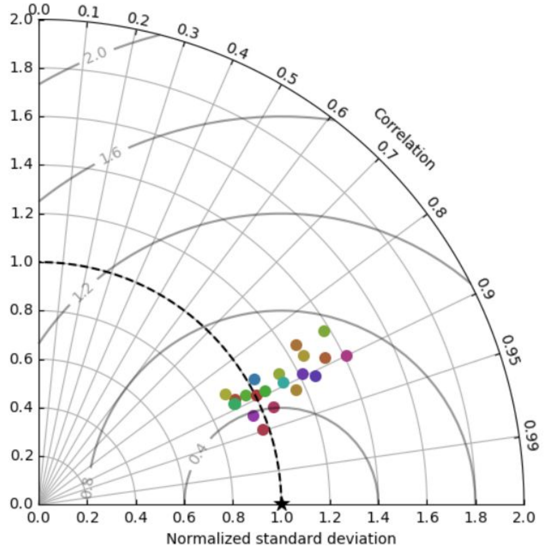


Gross Primary Productivity

	Download Data	Period Mean (original grids) [Pg yr-1]	Model Period Mean (intersection) [Pg yr-1]	Model Period Mean (complement) [Pg yr-1]	Benchmark Period Mean (intersection) [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	[-]	118.												
bcc-csm1-1	[-]	123.	114.	6.80	118.	0.0600	0.203	1.94	1.27	0.424	0.267	0.809	0.946	0.543
bcc-csm1-1-m	[-]	112.	108.	4.10	118.	0.501	-0.116	1.94	1.38	0.413	0.265	0.794	0.934	0.534
BCC-CSM2-MR	[-]	123.	115.	8.31	118.	0.501	-0.0721	1.68	1.28	0.433	0.326	0.796	0.941	0.564
BCC-ESM1	[-]	157.	133.	21.4	118.	0.0640	0.325	1.84	1.23	0.429	0.302	0.808	0.945	0.557
CanESM5	[-]	141.	131.	8.05	118.		0.675	1.85	1.70	0.427	0.330	0.701	0.934	0.544
CESM1-BGC	[-]	129.	124.	4.32	118.	0.501	0.309	1.74	1.38	0.392	0.350	0.761	0.873	0.545
CESM2	[-]	110.	105.	4.21	118.	0.473	-0.0938	1.72	1.52	0.411	0.364	0.786	0.935	0.572
CESM2-WACCM	[-]	110.	106.	4.28	118.	0.473	-0.0889	1.73	1.50	0.410	0.364	0.788	0.936	0.572
EC-Earth3-Veg	[-]	136.	134.	2.52	118.		0.330	1.99	1.49	0.417	0.312	0.755	0.931	0.545
GFDL-ESM2G	[-]	167.	155.	9.78	118.		1.19	3.18	1.45	0.360	0.185	0.726	0.880	0.467
GISS-E2-1-G	[-]	133.	118.	12.6	117.	1.29	0.0302	1.55	1.23	0.411	0.355	0.741	0.905	0.553
GISS-E2-1-H	[-]	131.	116.	13.8	118.	0.654	-0.0269	1.57	1.19	0.400	0.353	0.760	0.913	0.556
inmcm4	[-]	136.	128.	8.25	113.	5.44	0.351	1.78	1.41	0.451	0.308	0.766	0.935	0.554
IPSL-CM5A-LR	[-]	165.	153.	9.00	118.	0.347	1.10	2.73	1.30	0.318	0.241	0.770	0.889	0.492
IPSL-CM6A-LR	[-]	116.	111.	4.25	118.	0.486	0.0566	1.45	1.32	0.498	0.364	0.751	0.960	0.587
MeanCMIP5	[-]	138.	131.	6.75	118.		0.561	1.44	1.13	0.462	0.408	0.794	0.959	0.606
MeanCMIP6	[-]	121.	116.	5.10	118.		0.159	1.10	1.12	0.522	0.470	0.796	0.973	0.646
MIROC-ESM	[-]	129.	121.	6.01	108.	10.1	0.308	2.06	1.40	0.425	0.322	0.749	0.918	0.547
MPI-ESM-LR	[-]	170.	162.	6.90	110.	8.62	1.22	2.37	1.43	0.378	0.291	0.699	0.926	0.517
NorESM1-ME	[-]	129.	121.	6.29	118.		0.331	1.92	1.46	0.354	0.350	0.759	0.838	0.530
SAM0-UNICON	[-]	131.	126.	4.95	118.	0.501	0.371	1.75	1.39	0.398	0.338	0.764	0.845	0.537

- Multimodel GPP is compared with global seasonal GBAF estimates
- We can see Improvements across generations of models (e.g., CESM1 vs. CESM2, IPSL-CM5A vs. 6A)
- The mean CMIP6 and CMIP5 models perform best

Spatial Taylor Diagram



Argonne NATIONAL LABORATORY · BROOKHAVEN NATIONAL LABORATORY · BERKELEY LAB · Los Alamos NATIONAL LABORATORY · NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH · OAK RIDGE National Laboratory · UCI · MICHIGAN

RUBISCO

Summary

- **Model benchmarking** is increasingly important as model complexity increases
- Systematic model benchmarking is useful for
  - **Verification** – during model development to confirm that new model code improves performance in a targeted area without degrading performance in another area
  - **Validation** – when comparing performance of one model or model version to observations and to other models or other model versions
- The **ILAMB package** employs a suite of in situ, remote sensing, and reanalysis datasets to comprehensively evaluate and score land model performance, *irrespective of any model structure or set of process representations*
- ILAMB is **Open Source**, is written in **Python**, **runs in parallel** on laptops to supercomputers, and has been **adopted in most modeling centers**
- *Usefulness* of ILAMB depends on the quality of incorporated observational data, characterization of uncertainty, and selection of relevant metrics

Argonne NATIONAL LABORATORY · BROOKHAVEN NATIONAL LABORATORY · BERKELEY LAB · Los Alamos NATIONAL LABORATORY · NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH · OAK RIDGE National Laboratory · UCI · MICHIGAN

For more information...

- **International Land Model Benchmarking (ILAMB) Package**
  https://www.ilamb.org/
- **Reducing Uncertainties in Biogeochemical Interactions through Synthesis and Computation (RUBISCO) Scientific Focus Area**
  https://www.bgc-feedbacks.org/
- **Forrest M. Hoffman**
  Oak Ridge National Laboratory
  forrest@climatemodeling.org

References (1/3)

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

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

Eyring, V., P. M. Cox, G. M. Flato, P. J. Gleckler, G. Abramowitz, P. Caldwell, W. D. Collins, B. K. Gier, A. D. Hall, F. M. Hoffman, G. C. Hurtt, A. Jahn, C. D. Jones, S. A. Klein, J. Krasting, L. Kwiatkowski, R. Lorenz, E. Maloney, G. A. Meehl, A. Pendergrass, R. Pincus, A. C. Ruane, J. L. Russell, B. M. Sanderson, B. D. Santer, S. C. Sherwood, I. R. Simpson, R. J. Stouffer, and M. S. Williamson (2019), Taking climate model evaluation to the next level, *Nat. Clim. Change*, 9(2):102–110, doi:10.1038/s41558-018-0355-y.

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

References (2/3)

Lawrence, D. M., R. A. Fisher, C. D. Koven, K. W. Oleson, S. C. Swenson, G. B. Bonan, N. Collier, B. Ghimire, L. van Kampenhout, D. Kennedy, E. Kluzek, P. J. Lawrence, F. Li, H. Li, D. Lombardozzi, W. J. Riley, W. J. Sacks, M. Shi, M. Vertenstein, W. R. Wieder, C. Xu, A. A. Ali, A. M. Badger, G. Bisht, M. van den Broeke, M. A. Brunke, S. P. Burns, J. Buzan, M. Clark, A. Craig, K. Dahlin, B. Drewniak, J. B. Fisher, M. Flanner, A. M. Fox, P. Gentine, F. M. Hoffman, G. Keppel-Aleks, R. Knox, S. Kumar, J. Lenaerts, L. R. Leung, W. H. Lipscomb, Y. Lu, A. Pandey, J. D. Pelletier, J. Perket, J. T. Randerson, D. M. Ricciuto, B. M. Sanderson, A. Slater, Z. M. Subin, J. Tang, R. Q. Thomas, M. V. Martin, and X. Zeng (2019), The Community Land Model Version 5: Description of new features, benchmarking, and impact of forcing uncertainty, *J. Adv. Model. Earth Syst.*, 11(12):4245–4287, doi:10.1029/2018MS001583.

Le Quéré, C., R. M. Andrew, P. Friedlingstein, S. Sitch, J. Hauck, J. Pongratz, P. A. Pickers, J. I. Korsbakken, G. P. Peters, J. G. Canadell, A. Arneth, V. K. Arora, L. Barbero, A. Bastos, L. Bopp, F. Chevallier, L. P. Chini, P. Ciais, S. C. Doney, T. Gkritzalis, D. S. Goll, I. Harris, V. Haverd, F. M. Hoffman, M. Hoppema, R. A. Houghton, G. Hurtt, T. Ilyina, A. K. Jain, T. Johannessen, C. D. Jones, E. Kato, R. F. Keeling, K. K. Goldewijk, P. Landschützer, N. Lefèvre, S. Lienert, Z. Liu, D. Lombardozzi, N. Metzl, D. R. Munro, J. E. M. S. Nabel, S.-I. Nakaoka, C. Neill, A. Olsen, T. Ono, P. Patra, A. Peregon, W. Peters, P. Peylin, B. Pfeil, D. Pierrot, B. Poulter, G. Rehder, L. Resplandy, E. Robertson, M. Rocher, C. Rödenbeck, U. Schuster, J. Schwinger, R. Séférian, I. Skjelvan, T. Steinhoff, A. Sutton, P. P. Tans, H. Tian, B. Tilbrook, F. N. Tubiello, I. T. van der Laan-Luijkx, G. R. van der Werf, N. Viovy, A. P. Walker, A. J. Wiltshire, R. Wright, S. Zaehle, and B. Zheng (2018), Global Carbon Budget 2018, Earth Syst. Sci. Data, 10(4):2141–2194, doi:10.5194/essd-10-2141-2018.

References (3/3)

Luo, Y. Q., J. T. Randerson, G. Abramowitz, C. Bacour, E. Blyth, N. Carvalhais, P. Ciais, D. Dalmonech, J. B. Fisher, R. Fisher, P. Friedlingstein, K. Hibbard, F. Hoffman, D. Huntzinger, C. D. Jones, C. Koven, D. Lawrence, D. J. Li, M. Mahecha, S. L. Niu, R. Norby, S. L. Piao, X. Qi, P. Peylin, I. C. Prentice, W. Riley, M. Reichstein, C. Schwalm, Y. P. Wang, J. Y. Xia, S. Zaehle, and X. H. Zhou (2012), A framework for benchmarking land models, *Biogeosci.*, 9(10):3857–3874, doi:10.5194/bg-9-3857-2012.

Randerson, J. T., F. M. Hoffman, P. E. Thornton, N. M. Mahowald, K. Lindsay, Y.-H. Lee, C. D. Nevison, S. C. Doney, G. Bonan, R. Stöckli, C. Covey, S. W. Running, and I. Y. Fung (2009), Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models, *Glob. Change Biol.*, 15(10):2462–2484, doi:10.1111/j.1365-2486.2009.01912.x.

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

Zhu, Q., W. J. Riley, J. Tang, N. Collier, F. M. Hoffman, X. Yang, and G. Bisht (2019), Representing nitrogen, phosphorus, and carbon interactions in the E3SM Land Model: Development and global benchmarking, *J. Adv. Model. Earth Syst.*, 11(7):2238–2258, doi:10.1029/2018MS001571.