U.S. DEPARTMENT OF
ENERGY
Office of Science

Systematic Assessment and Benchmarking of Earth System Models

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

Resource Competition, Environmental Security and Stability (RECESS) Meeting

June 18, 2024

Forrest M. Hoffman, Computational Earth System Scientist

- Group Leader for the ORNL Computational Earth Sciences Group
- 35 years at ORNL in Environmental Sciences Division, then Computer Science and Mathematics Division, and now Computational Sciences and Engineering Division
- Develop and apply Earth system models to study global biogeochemical cycles, including terrestrial & marine carbon cycle
- Investigate methods for reconciling uncertainties in carbon–climate feedbacks through comparison with observations
- Apply artificial intelligence methods (machine learning and data mining) to environmental characterization, simulation, & analysis
- Joint Faculty, University of Tennessee, Knoxville, Department of Civil & Environmental Engineering

Oak Ridge National Laboratory

The University of Tennessee Knoxville

US Dept. of Energy's RUBISCO Scientific Focus Area (SFA)

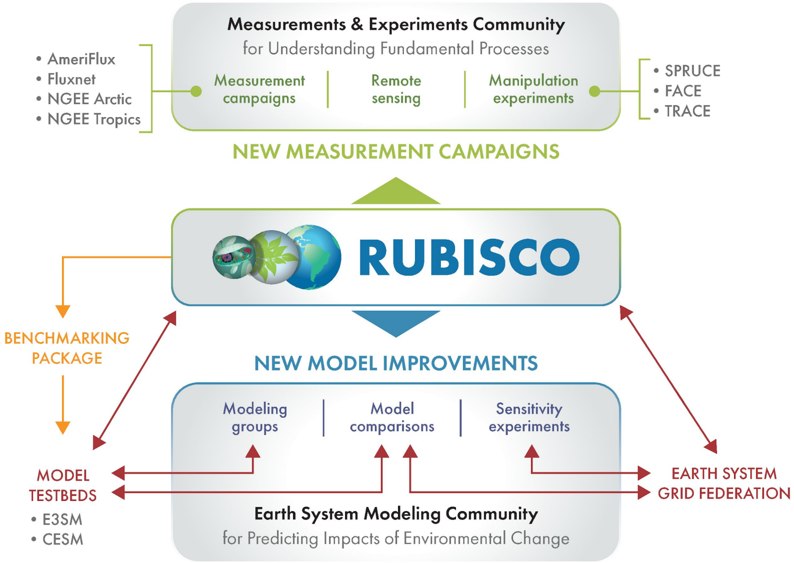
Forrest M. Hoffman (Laboratory Research Manager), William J. Riley (Senior Science Co-Lead), and James T. Randerson (Chief Scientist)

Research Goals

- Identify and quantify feedbacks between biogeochemical cycles and the Earth system
- Quantify and reconcile 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) tools for systematic evaluation of model fidelity
- Conduct and evaluate CMIP6 experiments with ESMs



Measurements & Experiments Community
for Understanding Fundamental Processes

Measurement campaigns | Remote sensing | Manipulation experiments

- AmeriFlux
- Fluxnet
- NGEE Arctic
- NGEE Tropics

- SPRUCE
- FACE
- TRACE

NEW MEASUREMENT CAMPAIGNS

RUBISCO

BENCHMARKING PACKAGE

NEW MODEL IMPROVEMENTS

Modeling groups | Model comparisons | Sensitivity experiments

MODEL TESTBEDS

- E3SM
- CESM

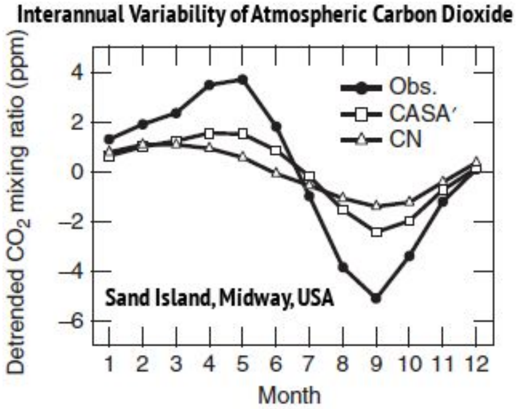
EARTH SYSTEM GRID FEDERATION

Earth System Modeling Community
for Predicting Impacts of Environmental Change

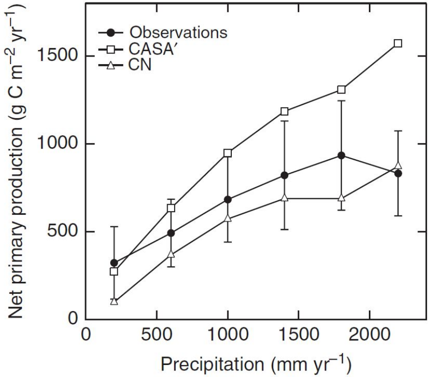
The RUBISCO SFA works with the measurements and the modeling communities to use best-available data to evaluate the fidelity of ESMs. RUBISCO identifies model gaps and weaknesses, informs new model development efforts, and suggests new measurements and field campaigns.

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 relationship 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?

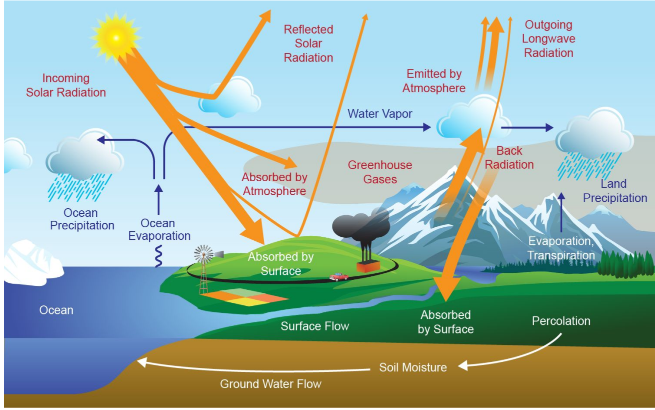
- To **quantify and reduce uncertainties** in carbon cycle feedbacks to improve projections of future climate change (Eyring et al., 2019; Collier et al., 2018)
- To **diagnose impacts of process-based or machine learning model development** on process representations and their interactions
- To **guide synthesis efforts**, such as the Intergovernmental Panel on Climate Change (IPCC), by determining which models are broadly consistent with 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

RUBISCO

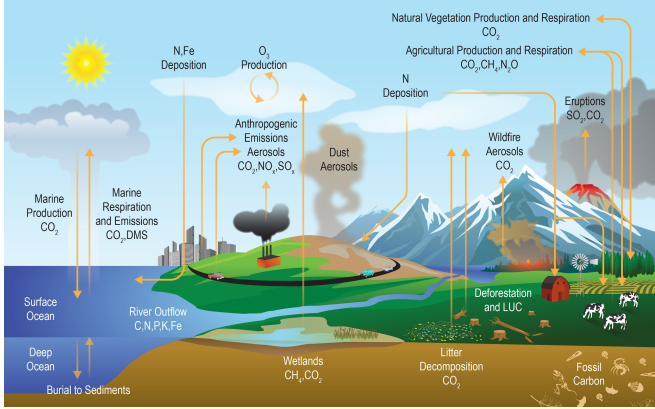
What is ILAMB?

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**



Energy and Water Cycles



Carbon and Biogeochemical Cycles

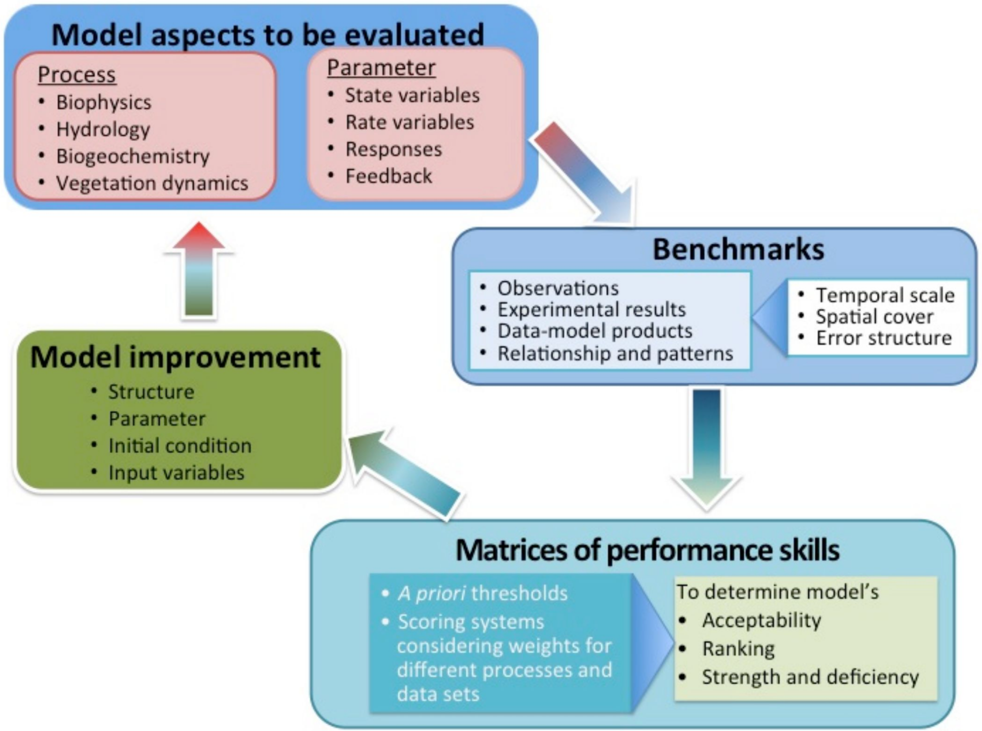
Argonne NATIONAL LABORATORY BROOKHAVEN NATIONAL LABORATORY BERKELEY LAB Los Alamos NATIONAL LABORATORY NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH OAK RIDGE National Laboratory UCI MICHIGAN

International Land Model Benchmarking (ILAMB) Meeting
The Beckman Center, Irvine, CA, USA January 24-26, 2011

- **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



Model aspects to be evaluated

Process
- Biophysics
- Hydrology
- Biogeochemistry
- Vegetation dynamics

Parameter
- State variables
- Rate variables
- Responses
- Feedback

Benchmarks
- Observations
- Experimental results
- Data-model products
- Relationship and patterns

- Temporal scale
- Spatial cover
- Error structure

Matrices of performance skills
- *A priori* thresholds
- Scoring systems considering weights for different processes and data sets

To determine model's
- Acceptability
- Ranking
- Strength and deficiency

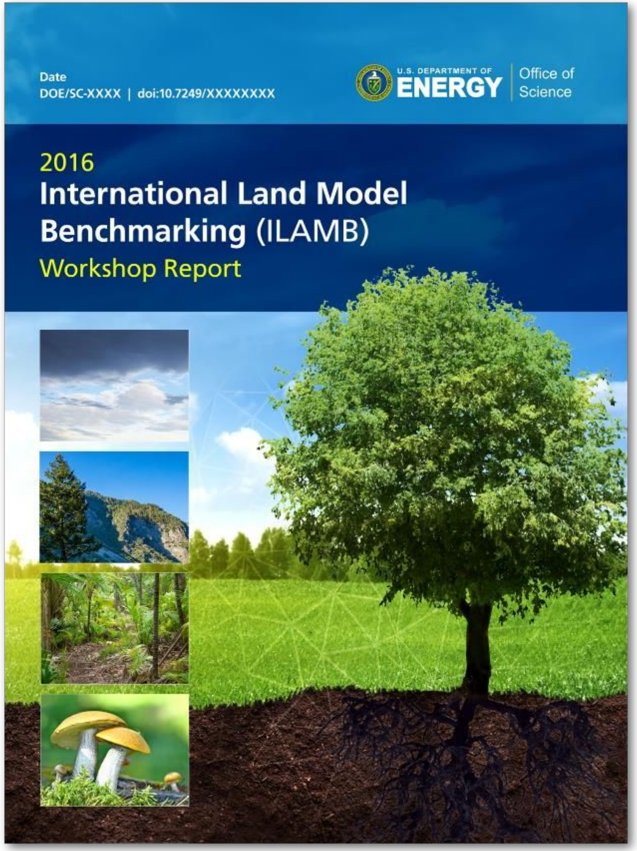
Model improvement
- Structure
- Parameter
- Initial condition
- Input variables

(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 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



Date
DOE/SC-XXXX | doi:10.7249/XXXXXXXX

U.S. DEPARTMENT OF ENERGY | Office of Science

2016
International Land Model Benchmarking (ILAMB)
Workshop Report

(Hoffman et al., 2017)

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

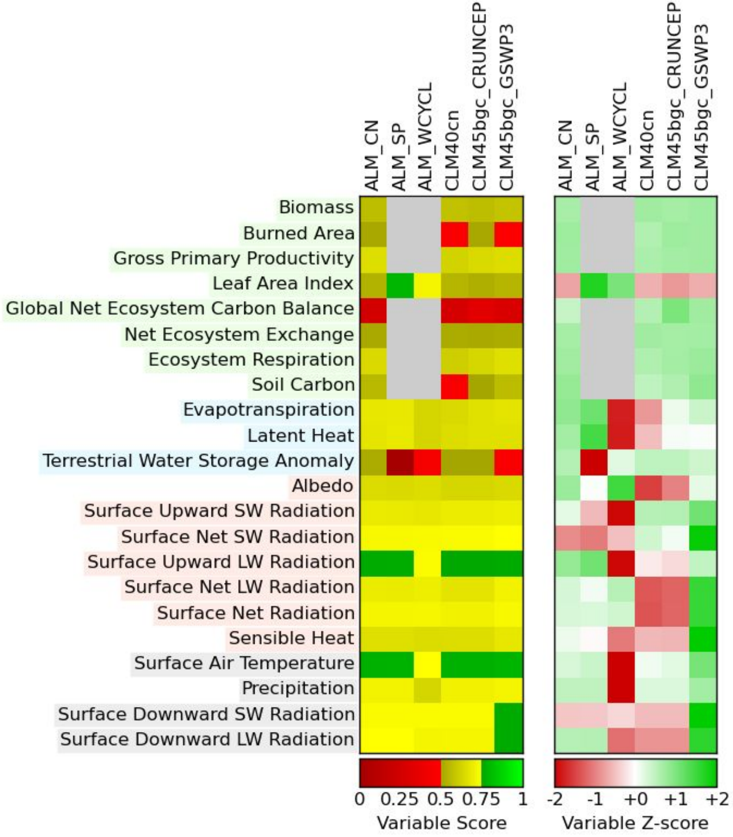
- **ILAMBv1** released at 2015 AGU Fall Meeting Town Hall, doi:10.18139/ILAMB.v001.00/1251597

- **ILAMBv2** released at 2016 ILAMB Workshop, doi:10.18139/ILAMB.v002.00/1251621

- **Open Source software** written in Python; **runs in parallel** on laptops, clusters, and supercomputers

- Routinely used for land model evaluation during development of ESMs, including the **E3SM Land Model** (Zhu et al., 2019) and the **CESM Community Land Model** (Lawrence et al., 2019)

- **Models are scored** based on statistical comparisons 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}) →

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

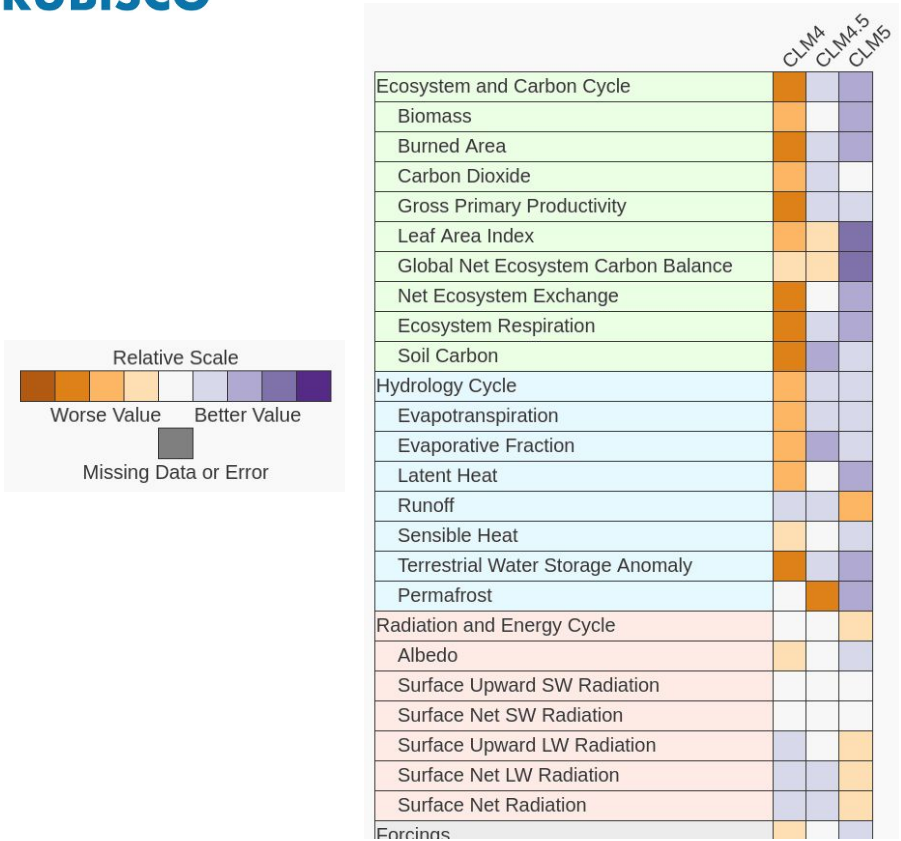
  - **Graphical diagnostics**
    - Spatial contour maps
    - Time series line plots
    - Spatial Taylor diagrams (Taylor, 2001)
- Similar **tables** and **graphical diagnostics** for functional relationships

RUBISCO

ILAMBv2.6 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

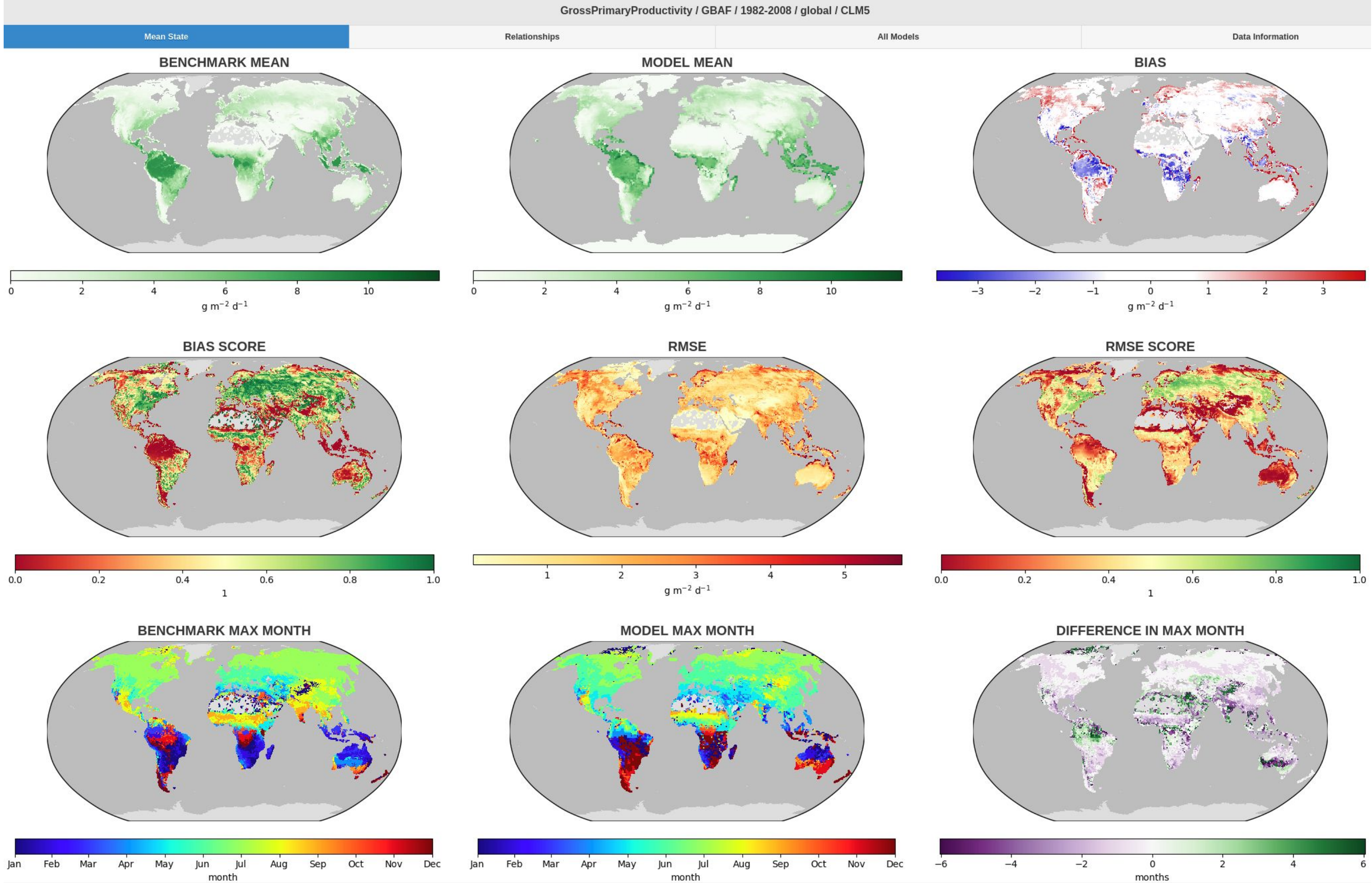


- Improvements in mechanistic treatment of hydrology, ecology, and land use with much more complexity in Community Land Model version 5 (CLM5)
- Simulations improved even with enhanced complexity
- Observational datasets not always self-consistent
- Forcing uncertainty confounds assessment of model development

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

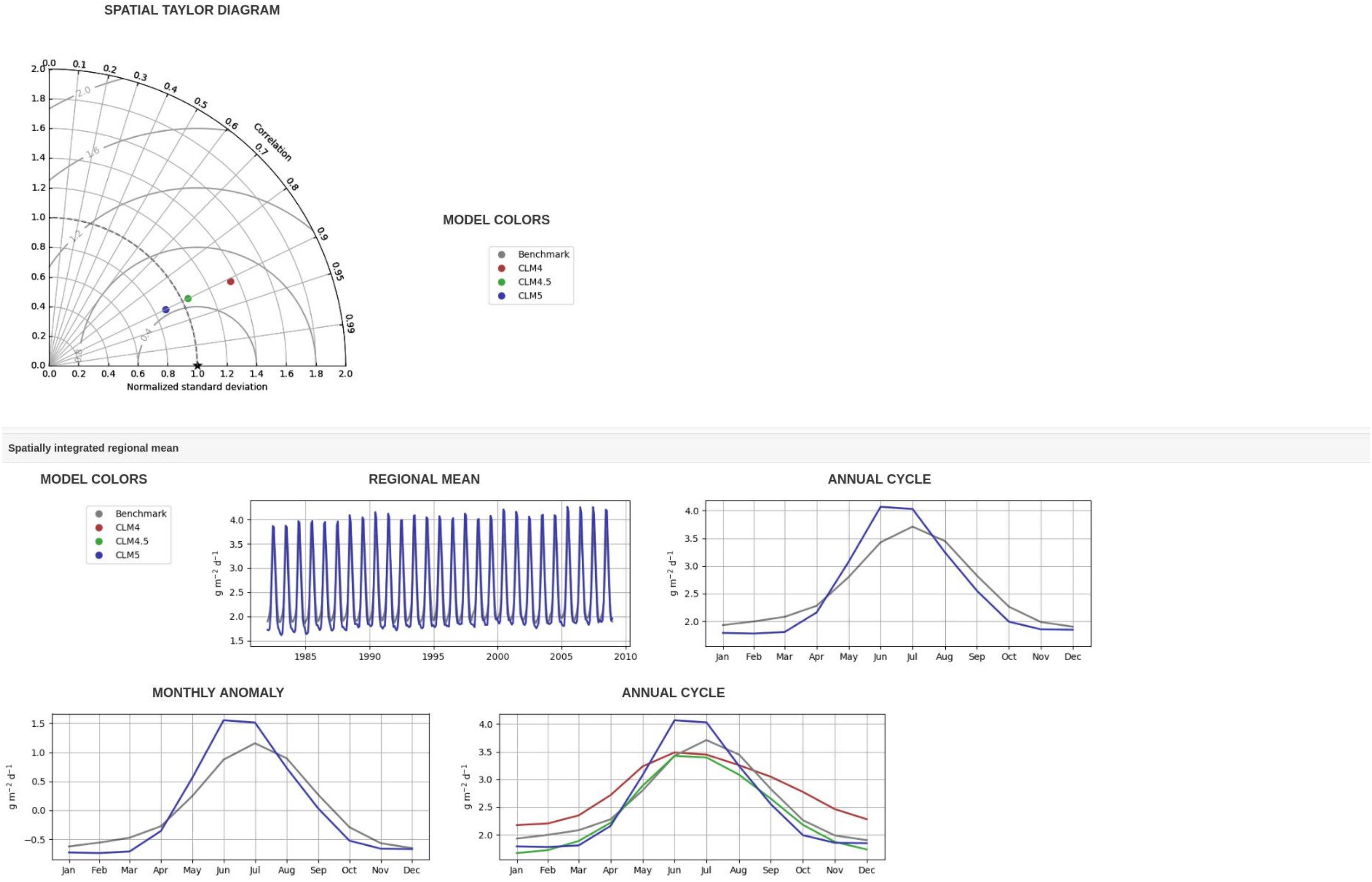
(Lawrence et al., 2019)

ILAMB Graphical Diagnostics



GrossPrimaryProductivity / GBAF / 1982-2008 / global / CLM5

ILAMB Graphical Diagnostics



SPATIAL TAYLOR DIAGRAM

MODEL COLORS

- ● Benchmark
- ● CLM4
- ● CLM4.5
- ● CLM5

Spatially integrated regional mean

MODEL COLORS

- ● Benchmark
- ● CLM4
- ● CLM4.5
- ● CLM5

REGIONAL MEAN

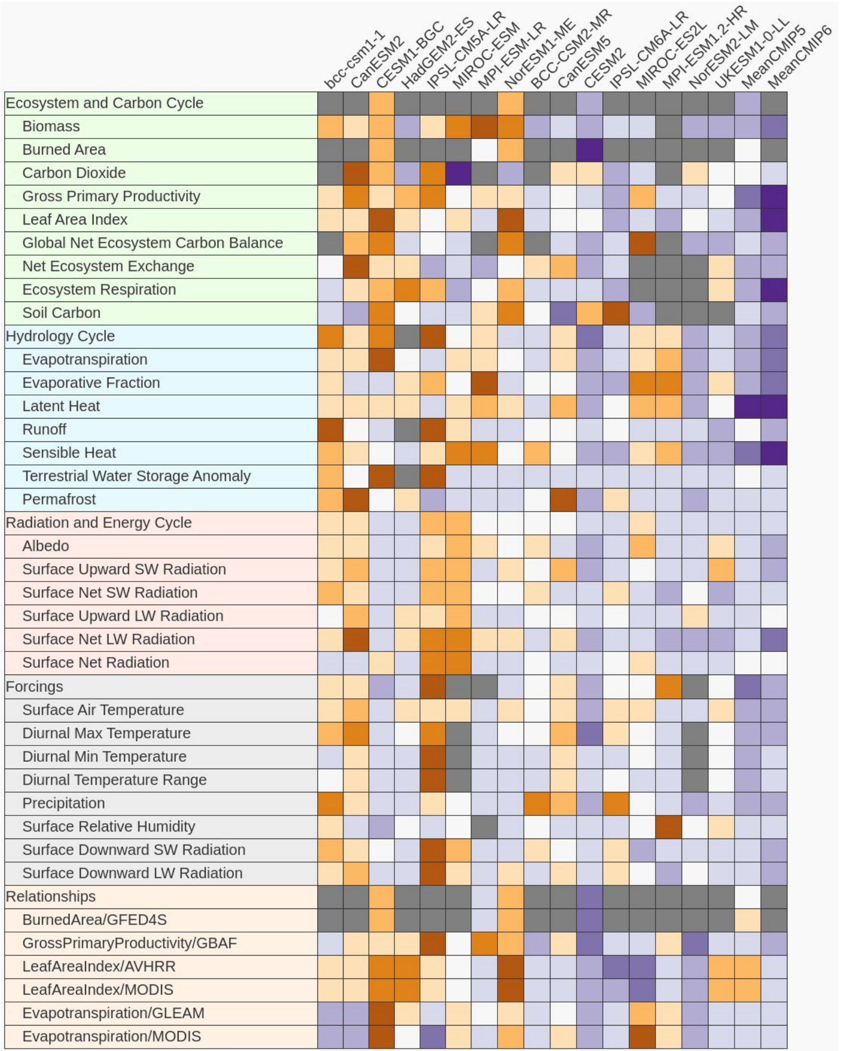
ANNUAL CYCLE

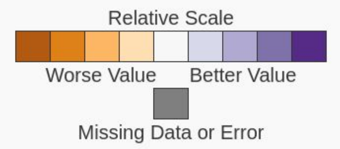
MONTHLY ANOMALY

ANNUAL CYCLE

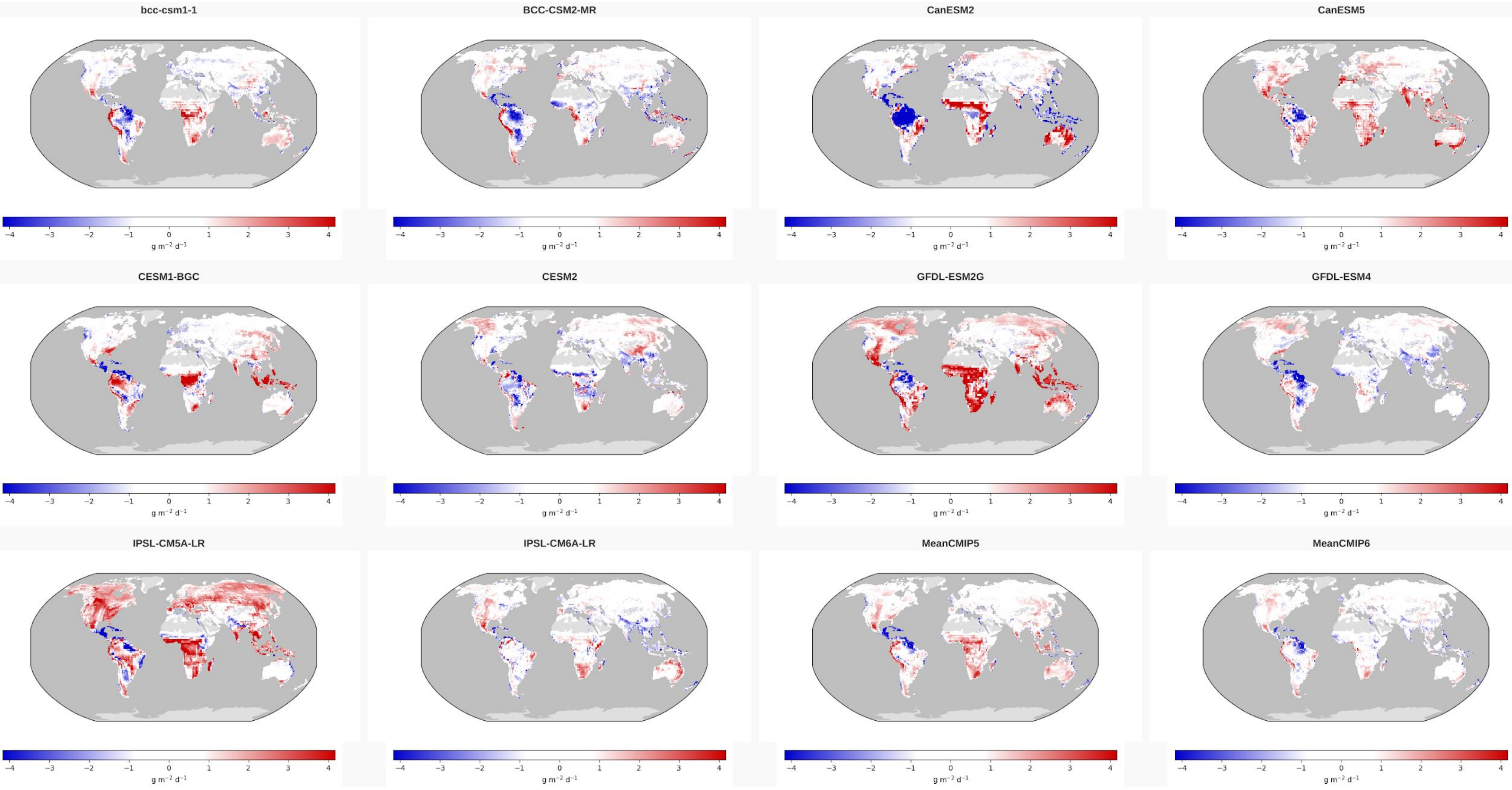
CMIP5 vs. CMIP6 Models

- 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
- Why did CMIP6 land models improve?

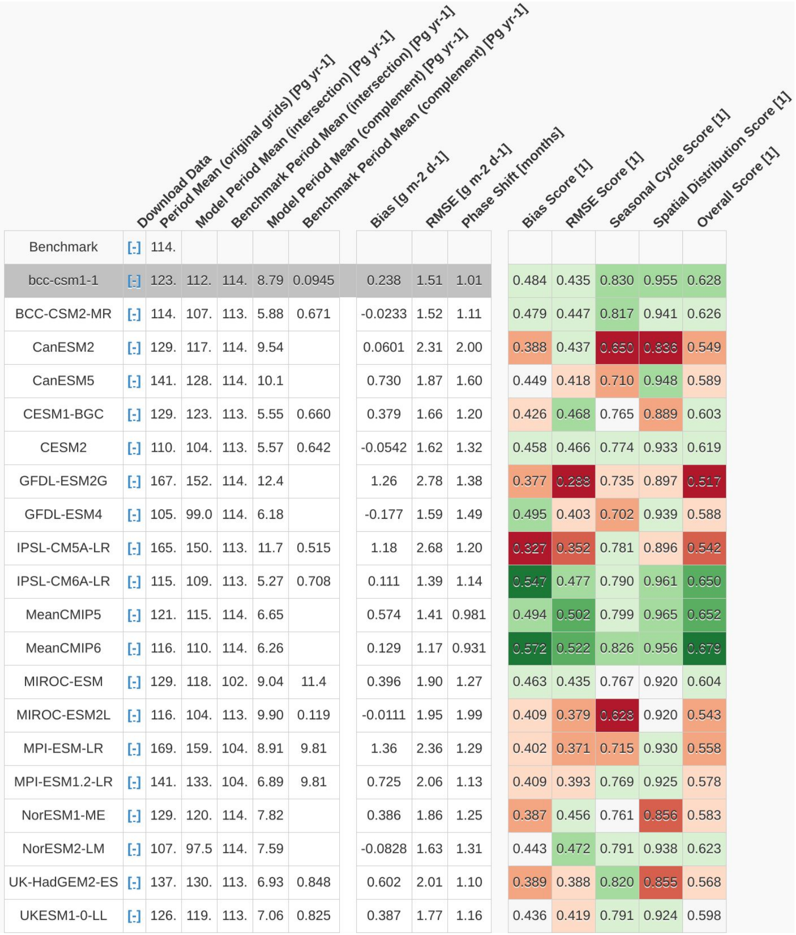




(Hoffman et al., in prep)


bcc-csm1-1
−4 −3 −2 −1 0 1 2 3 4
g m^{-2} d^{-1}
BCC-CSM2-MR
−4 −3 −2 −1 0 1 2 3 4
g m^{-2} d^{-1}
CanESM2
−4 −3 −2 −1 0 1 2 3 4
g m^{-2} d^{-1}
CanESM5
−4 −3 −2 −1 0 1 2 3 4
g m^{-2} d^{-1}
CESM1-BGC
−4 −3 −2 −1 0 1 2 3 4
g m^{-2} d^{-1}
CESM2
−4 −3 −2 −1 0 1 2 3 4
g m^{-2} d^{-1}
GFDL-ESM2G
−4 −3 −2 −1 0 1 2 3 4
g m^{-2} d^{-1}
GFDL-ESM4
−4 −3 −2 −1 0 1 2 3 4
g m^{-2} d^{-1}
IPSL-CM5A-LR
−4 −3 −2 −1 0 1 2 3 4
g m^{-2} d^{-1}
IPSL-CM6A-LR
−4 −3 −2 −1 0 1 2 3 4
g m^{-2} d^{-1}
MeanCMIP5
−4 −3 −2 −1 0 1 2 3 4
g m^{-2} d^{-1}
MeanCMIP6
−4 −3 −2 −1 0 1 2 3 4
g m^{-2} d^{-1}

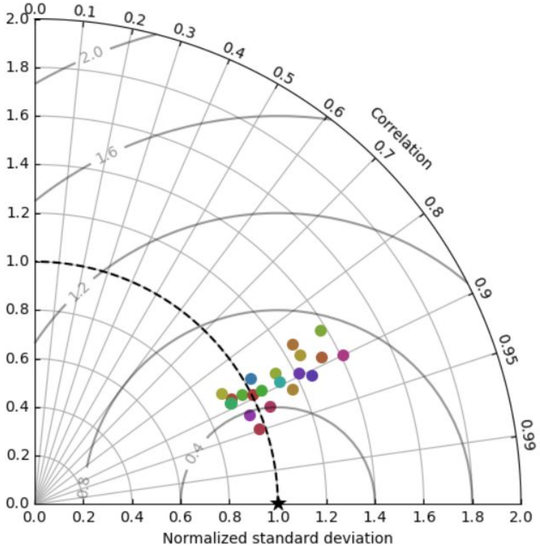
Gross Primary Productivity



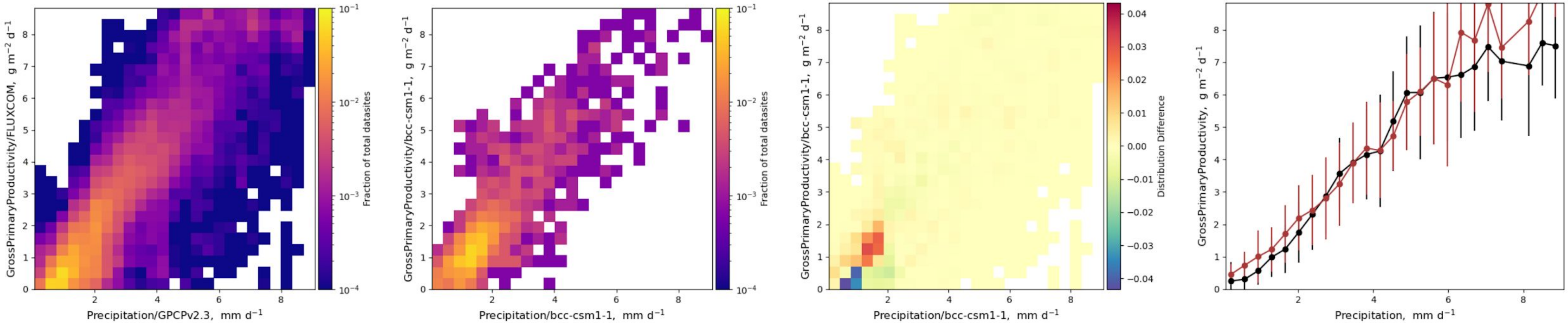
	Download Data	Period Mean (original grids) [Pg yr-1]	Model Period Mean (intersection) [Pg yr-1]	Benchmark Period Mean (intersection) [Pg yr-1]	Model Period Mean (complement) [Pg yr-1]	Benchmark Period Mean (complement) [Pg yr-1]	Bias [g m-2 d-1]	RMSE [g m-2 d-1]	Phase Shift [months]	Bias Score [1]	RMSE 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

- 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



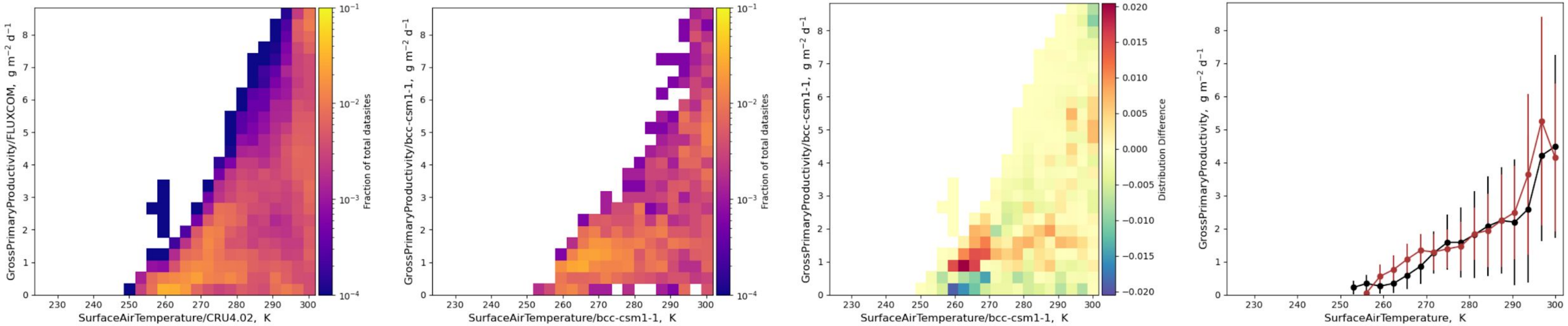
⊖ Precipitation/GPCPv2.3



⊕ SurfaceDownwardSWRadiation/CERESed4.1

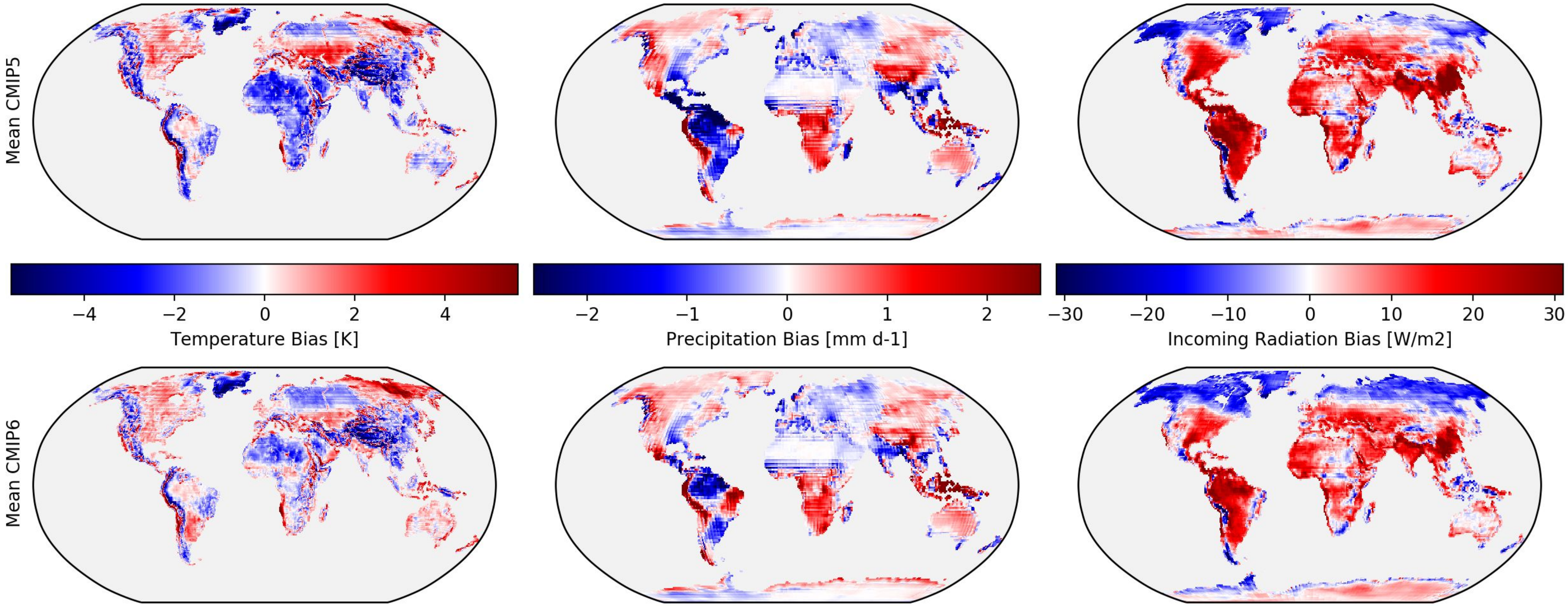
⊕ SurfaceNetSWRadiation/CERESed4.1

⊖ SurfaceAirTemperature/CRU4.02



Reasons for Land Model Improvements

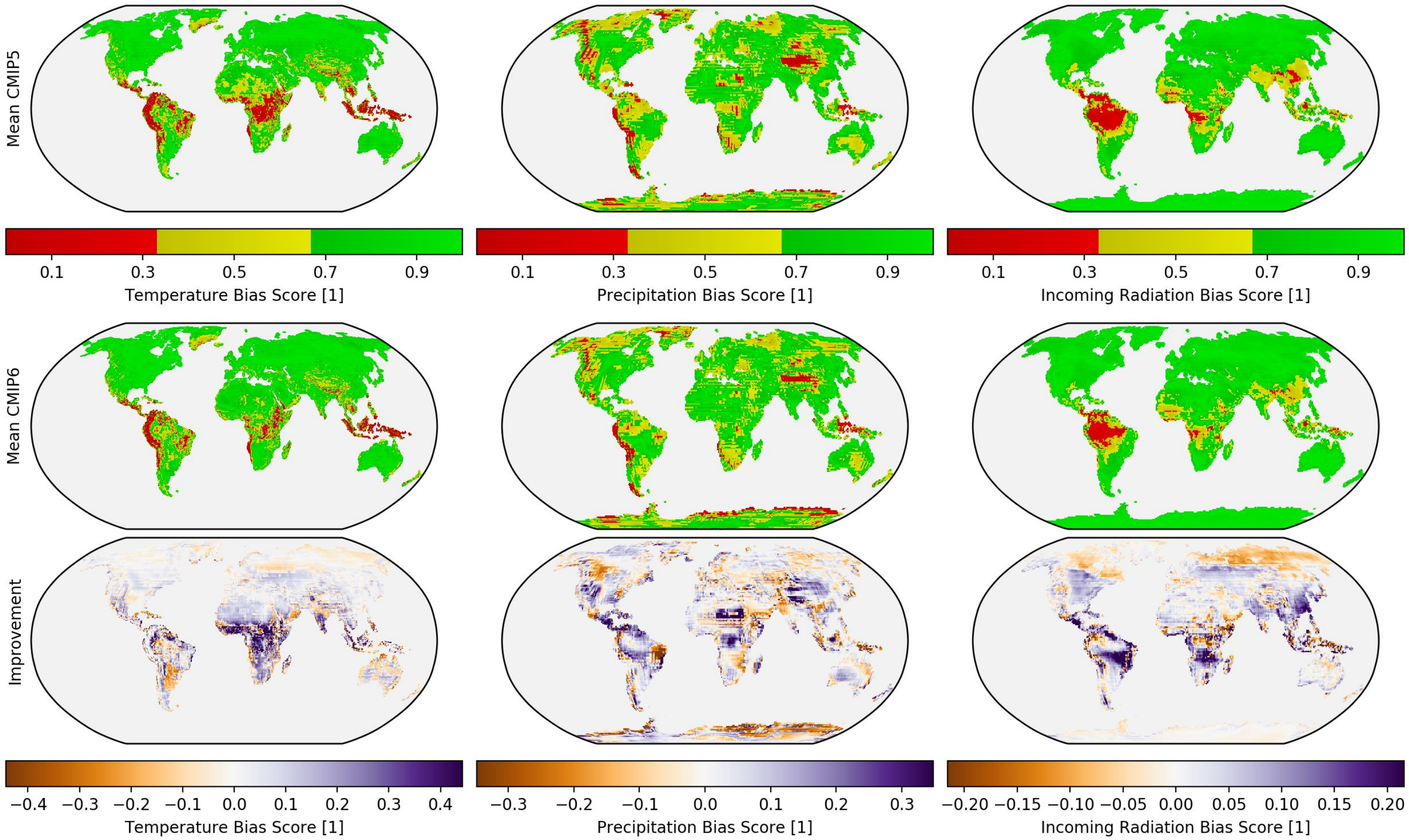
ESM improvements in climate forcings (temperature, precipitation, radiation) likely partially drove improvements exhibited by land carbon cycle models



(Hoffman et al., in prep)

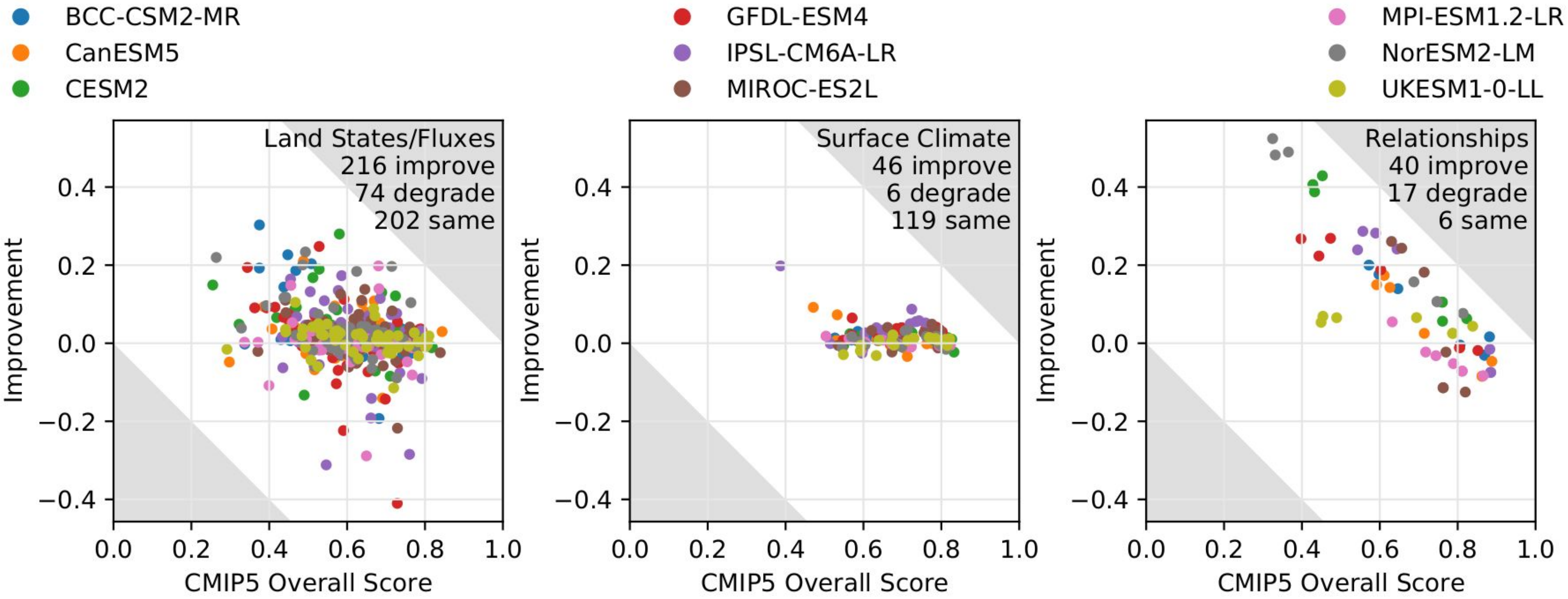
Reasons for Land Model Improvements

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



Mean CMIP5
Mean CMIP6
Improvement
0.1 0.3 0.5 0.7 0.9
Temperature Bias Score [1]
Precipitation Bias Score [1]
Incoming Radiation Bias Score [1]
−0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4
Temperature Bias Score [1]
−0.3 −0.2 −0.1 0.0 0.1 0.2 0.3
Precipitation Bias Score [1]
−0.20 −0.15 −0.10 −0.05 0.00 0.05 0.10 0.15 0.20
Incoming Radiation Bias Score [1]

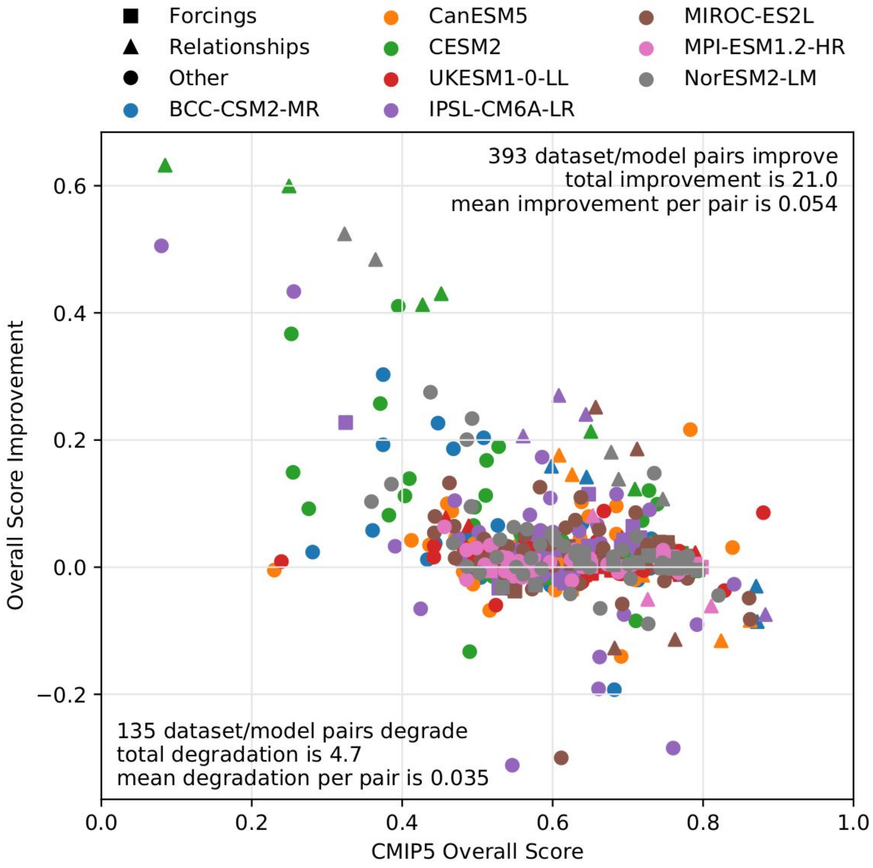
(Hoffman et al., in prep)



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.

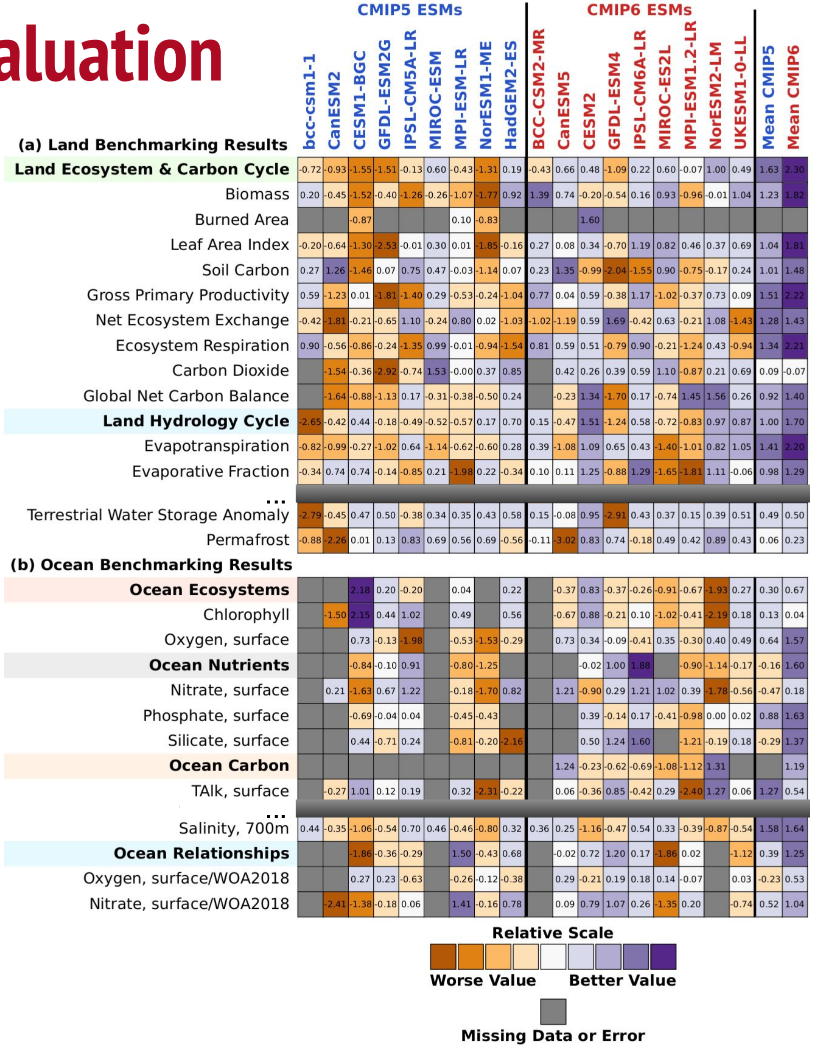
Reasons for Land Model Improvements

While forcings got better, the largest improvements were in **variable-to-variable relationships**, suggesting that increased land model complexity was also partially responsible for higher CMIP6 model scores



ILAMB & IOMB CMIP5 vs 6 Evaluation

- (a) ILAMB and (b) IOMB have been used to evaluate how land and ocean model performance has changed from CMIP5 to CMIP6
- Model fidelity is assessed through comparison of historical simulations with a wide variety of contemporary observational datasets
- The UN's Intergovernmental Panel on Climate Change (IPCC) Sixth Assessment Report (AR6) from Working Group 1 (WG1) Chapter 5 contains the full ILAMB/IOMB evaluation as Figure 5.22

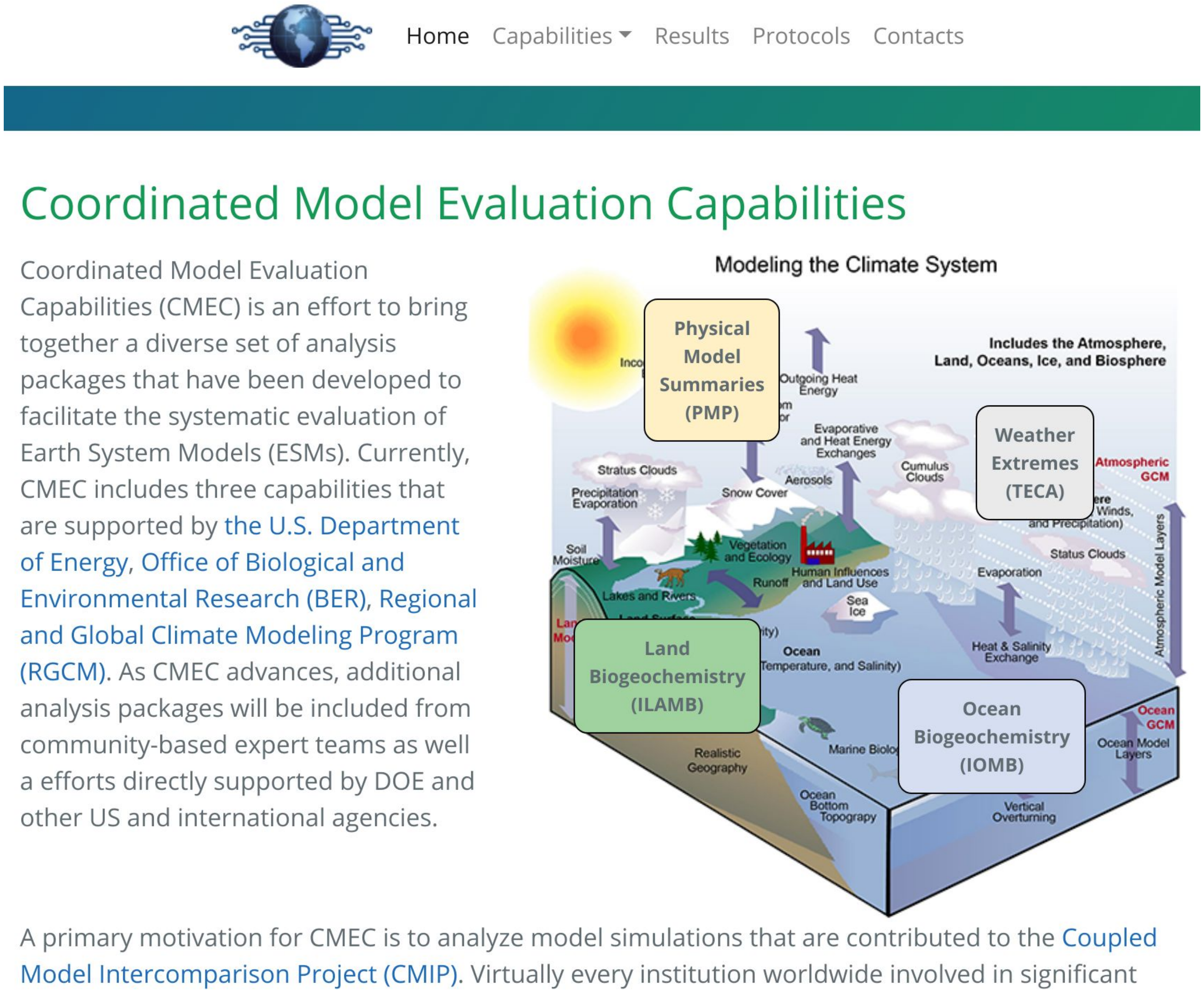


	CMIP5 ESMs									CMIP6 ESMs										
(a) Land Benchmarking Results	bcc-csm1-1	CanESM2	CESM1-BGC	GFDL-ESM2G	IPSL-CM5A-LR	MIROC-ESM	MPI-ESM-LR	NorESM1-ME	HadGEM2-ES	BCC-CSM2-MR	CanESM5	CESM2	GFDL-ESM4	IPSL-CM6A-LR	MIROC-ES2L	MPI-ESM1.2-LR	NorESM2-LM	UKESM1-0-LL	Mean CMIP5	Mean CMIP6
Land Ecosystem & Carbon Cycle	-0.72	-0.93	-1.55	-1.51	-0.13	0.60	-0.43	-1.31	0.19	-0.43	0.66	0.48	-1.09	0.22	0.60	-0.07	1.00	0.49	1.63	2.30
Biomass	0.20	-0.45	-1.52	-0.40	-1.26	-0.26	-1.07	-1.77	0.92	1.39	0.74	-0.20	-0.54	0.16	0.93	-0.96	-0.01	1.04	1.23	1.82
Burned Area			-0.87				0.10	-0.83				1.60								
Leaf Area Index	-0.20	-0.64	-1.30	-2.53	-0.01	0.30	0.01	-1.85	-0.16	0.27	0.08	0.34	-0.70	1.19	0.82	0.46	0.37	0.69	1.04	1.81
Soil Carbon	0.27	1.26	-1.46	0.07	0.75	0.47	-0.03	-1.14	0.07	0.23	1.35	-0.99	-2.04	-1.55	0.90	-0.75	-0.17	0.24	1.01	1.48
Gross Primary Productivity	0.59	-1.23	0.01	-1.81	-1.40	0.29	-0.53	-0.24	-1.04	0.77	0.04	0.59	-0.38	1.17	-1.02	-0.37	0.73	0.09	1.51	2.22
Net Ecosystem Exchange	-0.42	-1.81	-0.21	-0.65	1.10	-0.24	0.80	0.02	-1.03	-1.02	-1.19	0.59	1.69	-0.42	0.63	-0.21	1.08	-1.43	1.28	1.43
Ecosystem Respiration	0.90	-0.56	-0.86	-0.24	-1.35	0.99	-0.01	-0.94	-1.54	0.81	0.59	0.51	-0.79	0.90	-0.21	-1.24	0.43	-0.94	1.34	2.21
Carbon Dioxide		-1.54	-0.36	-2.92	-0.74	1.53	-0.00	0.37	0.85		0.42	0.26	0.39	0.59	1.10	-0.87	0.21	0.69	0.09	-0.07
Global Net Carbon Balance		-1.64	-0.88	-1.13	0.17	-0.31	-0.38	-0.50	0.24		-0.23	1.34	-1.70	0.17	-0.74	1.45	1.56	0.26	0.92	1.40
Land Hydrology Cycle	-2.65	-0.42	0.44	-0.18	-0.49	-0.52	-0.57	0.17	0.70	0.15	-0.47	1.51	-1.24	0.58	-0.72	-0.83	0.97	0.87	1.00	1.70
Evapotranspiration	-0.82	-0.99	-0.27	-1.02	0.64	-1.14	-0.62	-0.60	0.28	0.39	-1.08	1.09	0.65	0.43	-1.40	-1.01	0.82	1.05	1.41	2.20
Evaporative Fraction	-0.34	0.74	0.74	-0.14	-0.85	0.21	-1.98	0.22	-0.34	0.10	0.11	1.25	-0.88	1.29	-1.65	-1.81	1.11	-0.06	0.98	1.29
...																				
Terrestrial Water Storage Anomaly	-2.79	-0.45	0.47	0.50	-0.38	0.34	0.35	0.43	0.58	0.15	-0.08	0.95	-2.91	0.43	0.37	0.15	0.39	0.51	0.49	0.50
Permafrost	-0.88	-2.26	0.01	0.13	0.83	0.69	0.56	0.69	-0.56	-0.11	-3.02	0.83	0.74	-0.18	0.49	0.42	0.89	0.43	0.06	0.23
(b) Ocean Benchmarking Results																				
Ocean Ecosystems			2.18	0.20	-0.20		0.04		0.22		-0.37	0.83	-0.37	-0.26	-0.91	-0.67	-1.93	0.27	0.30	0.67
Chlorophyll		-1.50	2.15	0.44	1.02		0.49		0.56		-0.67	0.88	-0.21	0.10	-1.02	-0.41	-2.19	0.18	0.13	0.04
Oxygen, surface			0.73	-0.13	-1.98		-0.53	-1.53	-0.29		0.73	0.34	-0.09	-0.41	0.35	-0.30	0.40	0.49	0.64	1.57
Ocean Nutrients			-0.84	-0.10	0.91		-0.80	-1.25				-0.02	1.00	1.88		-0.90	-1.14	-0.17	-0.16	1.60
Nitrate, surface		0.21	-1.63	0.67	1.22		-0.18	-1.70	0.82		1.21	-0.90	0.29	1.21	1.02	0.39	-1.78	-0.56	-0.47	0.18
Phosphate, surface			-0.69	-0.04	0.04		-0.45	-0.43				0.39	-0.14	0.17	-0.41	-0.98	0.00	0.02	0.88	1.63
Silicate, surface			0.44	-0.71	0.24		-0.81	-0.20	-2.16			0.50	1.24	1.60		-1.21	-0.19	0.18	-0.29	1.37
Ocean Carbon											1.24	-0.23	-0.62	-0.69	-1.08	-1.12	1.31			1.19
TAlk, surface		-0.27	1.01	0.12	0.19		0.32	-2.31	-0.22		0.06	-0.36	0.85	-0.42	0.29	-2.40	1.27	0.06	1.27	0.54
...																				
Salinity, 700m	0.44	-0.35	-1.06	-0.54	0.70	0.46	-0.46	-0.80	0.32	0.36	0.25	-1.16	-0.47	0.54	0.33	-0.39	-0.87	-0.54	1.58	1.64
Ocean Relationships			-1.86	-0.36	-0.29		1.50	-0.43	0.68		-0.02	0.72	1.20	0.17	-1.86	0.02		-1.12	0.39	1.25
Oxygen, surface/WOA2018			0.27	0.23	-0.63		-0.26	-0.12	-0.38		0.29	-0.21	0.19	0.18	0.14	-0.07		0.03	-0.23	0.53
Nitrate, surface/WOA2018		-2.41	-1.38	-0.18	0.06		1.41	-0.16	0.78		0.09	0.79	1.07	0.26	-1.35	0.20		-0.74	0.52	1.04

Relative Scale

Worse Value **Better Value**

Missing Data or Error



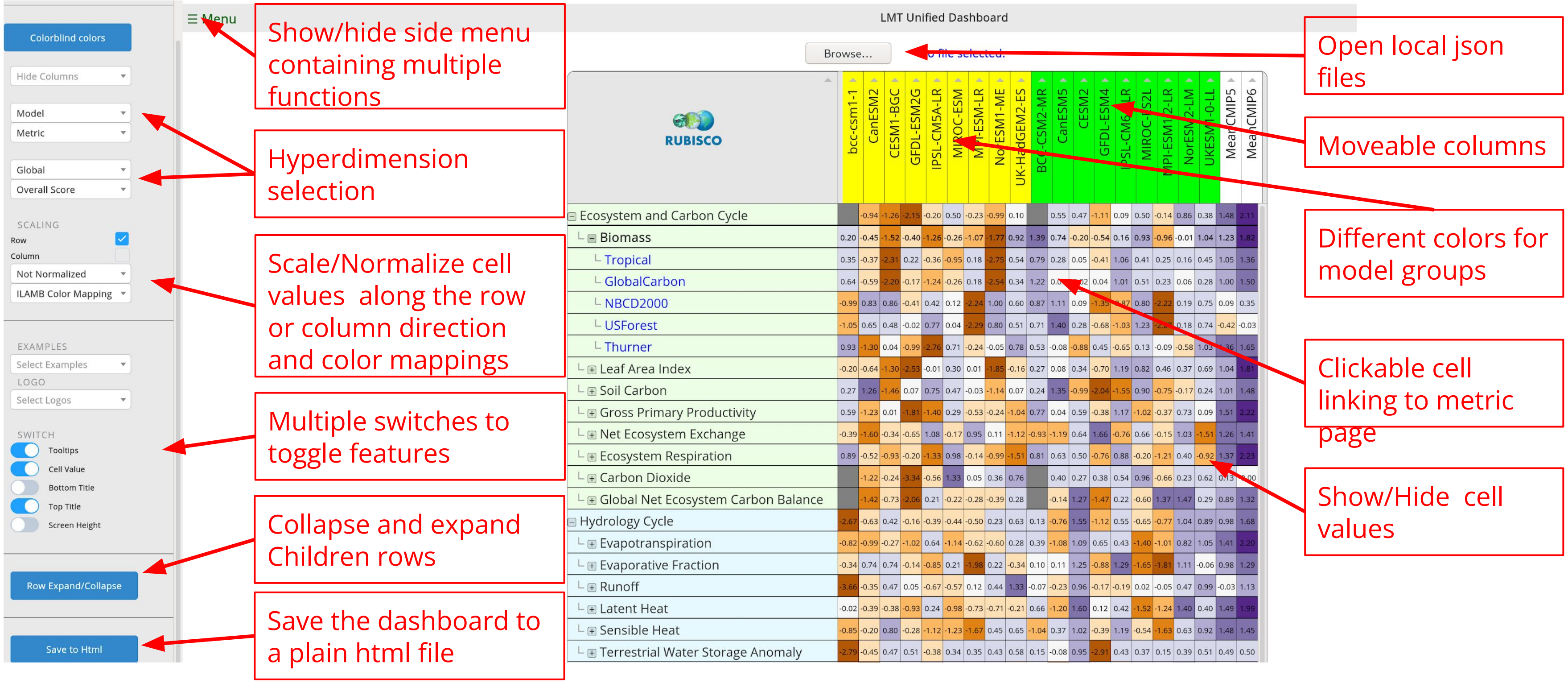
Coordinated Model Evaluation Capabilities

Coordinated Model Evaluation Capabilities (CMEC) is an effort to bring together a diverse set of analysis packages that have been developed to facilitate the systematic evaluation of Earth System Models (ESMs). Currently, CMEC includes three capabilities that are supported by the U.S. Department of Energy, Office of Biological and Environmental Research (BER), Regional and Global Climate Modeling Program (RGCM). As CMEC advances, additional analysis packages will be included from community-based expert teams as well a efforts directly supported by DOE and other US and international agencies.

A primary motivation for CMEC is to analyze model simulations that are contributed to the Coupled Model Intercomparison Project (CMIP). Virtually every institution worldwide involved in significant

https://cmec.llnl.gov/

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



Colorblind colors

Hide Columns

Model

Metric

Global

Overall Score

SCALING

Row ☑

Column □

Not Normalized

ILAMB Color Mapping

EXAMPLES

Select Examples

LOGO

Select Logos

SWITCH

● Tooltips

● Cell Value

○ Bottom Title

● Top Title

○ Screen Height

Row Expand/Collapse

Save to Html

☰ Menu

LMT Unified Dashboard

Browse... No file selected.

RUBISCO

	bcc-csm1-1	CanESM2	CESM1-BGC	GFDL-ESM2G	IPSL-CM5A-LR	MIROC-ESM	MPI-ESM-LR	NorESM1-ME	UK-HadGEM2-ES	BCC-CSM2-MR	CanESM5	CESM2	GFDL-ESM4	IPSL-CM6A-LR	MIROC-ES2L	MPI-ESM1.2-LR	NorESM2-LM	UKESM1-0-LL	MeanCMIP5	MeanCMIP6
⊟ Ecosystem and Carbon Cycle		-0.94	-1.26	-2.15	-0.20	0.50	-0.23	-0.99	0.10		0.55	0.47	-1.11	0.09	0.50	-0.14	0.86	0.38	1.48	2.11
└ ⊟ Biomass	0.20	-0.45	-1.52	-0.40	-1.26	-0.26	-1.07	-1.77	0.92	1.39	0.74	-0.20	-0.54	0.16	0.93	-0.96	-0.01	1.04	1.23	1.82
└ Tropical	0.35	-0.37	-2.31	0.22	-0.36	-0.95	0.18	-2.75	0.54	0.79	0.28	0.05	-0.41	1.06	0.41	0.25	0.16	0.45	1.05	1.36
└ GlobalCarbon	0.64	-0.59	-2.20	-0.17	-1.24	-0.26	0.18	-2.54	0.34	1.22	[illegible]	[illegible]	0.04	1.01	0.51	0.23	0.06	0.28	1.00	1.50
└ NBCD2000	-0.99	0.83	0.86	-0.41	0.42	0.12	-2.24	1.00	0.60	0.87	1.11	0.09	-1.35	-0.87	0.80	-2.22	0.19	0.75	0.09	0.35
└ USForest	-1.05	0.65	0.48	-0.02	0.77	0.04	-2.29	0.80	0.51	0.71	1.40	0.28	-0.68	-1.03	1.23	[illegible]	0.18	0.74	-0.42	-0.03
└ Thurner	0.93	-1.30	0.04	-0.99	-2.76	0.71	-0.24	-0.05	0.78	0.53	-0.08	-0.88	0.45	-0.65	0.13	-0.09	-0.58	1.03	1.36	1.65
└ ⊞ Leaf Area Index	-0.20	-0.64	-1.30	-2.53	-0.01	0.30	0.01	-1.85	-0.16	0.27	0.08	0.34	-0.70	1.19	0.82	0.46	0.37	0.69	1.04	1.81
└ ⊞ Soil Carbon	0.27	1.26	-1.46	0.07	0.75	0.47	-0.03	-1.14	0.07	0.24	1.35	-0.99	-2.04	-1.55	0.90	-0.75	-0.17	0.24	1.01	1.48
└ ⊞ Gross Primary Productivity	0.59	-1.23	0.01	-1.81	-1.40	0.29	-0.53	-0.24	-1.04	0.77	0.04	0.59	-0.38	1.17	-1.02	-0.37	0.73	0.09	1.51	2.22
└ ⊞ Net Ecosystem Exchange	-0.39	-1.60	-0.34	-0.65	1.08	-0.17	0.95	0.11	-1.12	-0.93	-1.19	0.64	1.66	-0.76	0.66	-0.15	1.03	-1.51	1.26	1.41
└ ⊞ Ecosystem Respiration	0.89	-0.52	-0.93	-0.20	-1.33	0.98	-0.14	-0.99	-1.51	0.81	0.63	0.50	-0.76	0.88	-0.20	-1.21	0.40	-0.92	1.37	2.23
└ ⊞ Carbon Dioxide		-1.22	-0.24	-3.34	-0.56	1.33	0.05	0.36	0.76		0.40	0.27	0.38	0.54	0.96	-0.66	0.23	0.62	0.13	0.00
└ ⊞ Global Net Ecosystem Carbon Balance		-1.42	-0.73	-2.06	0.21	-0.22	-0.28	-0.39	0.28		-0.14	1.27	-1.47	0.22	-0.60	1.37	1.47	0.29	0.89	1.32
⊟ Hydrology Cycle	-2.67	-0.63	0.42	-0.16	-0.39	-0.44	-0.50	0.23	0.63	0.13	-0.76	1.55	-1.12	0.55	-0.65	-0.77	1.04	0.89	0.98	1.68
└ ⊞ Evapotranspiration	-0.82	-0.99	-0.27	-1.02	0.64	-1.14	-0.62	-0.60	0.28	0.39	-1.08	1.09	0.65	0.43	-1.40	-1.01	0.82	1.05	1.41	2.20
└ ⊞ Evaporative Fraction	-0.34	0.74	0.74	-0.14	-0.85	0.21	-1.98	0.22	-0.34	0.10	0.11	1.25	-0.88	1.29	-1.65	-1.81	1.11	-0.06	0.98	1.29
└ ⊞ Runoff	-3.66	-0.35	0.47	0.05	-0.67	-0.57	0.12	0.44	1.33	-0.07	-0.23	0.96	-0.17	-0.19	0.02	-0.05	0.47	0.99	-0.03	1.13
└ ⊞ Latent Heat	-0.02	-0.39	-0.38	-0.93	0.24	-0.98	-0.73	-0.71	-0.21	0.66	-1.20	1.60	0.12	0.42	-1.52	-1.24	1.40	0.40	1.49	1.99
└ ⊞ Sensible Heat	-0.85	-0.20	0.80	-0.28	-1.12	-1.23	-1.67	0.45	0.65	-1.04	0.37	1.02	-0.39	1.19	-0.54	-1.63	0.63	0.92	1.48	1.45
└ ⊞ Terrestrial Water Storage Anomaly	-2.79	-0.45	0.47	0.51	-0.38	0.34	0.35	0.43	0.58	0.15	-0.08	0.95	-2.91	0.43	0.37	0.15	0.39	0.51	0.49	0.50

Show/hide side menu containing multiple functions

Hyperdimension selection

Scale/Normalize cell values along the row or column direction and color mappings

Multiple switches to toggle features

Collapse and expand Children rows

Save the dashboard to a plain html file

Open local json files

Moveable columns

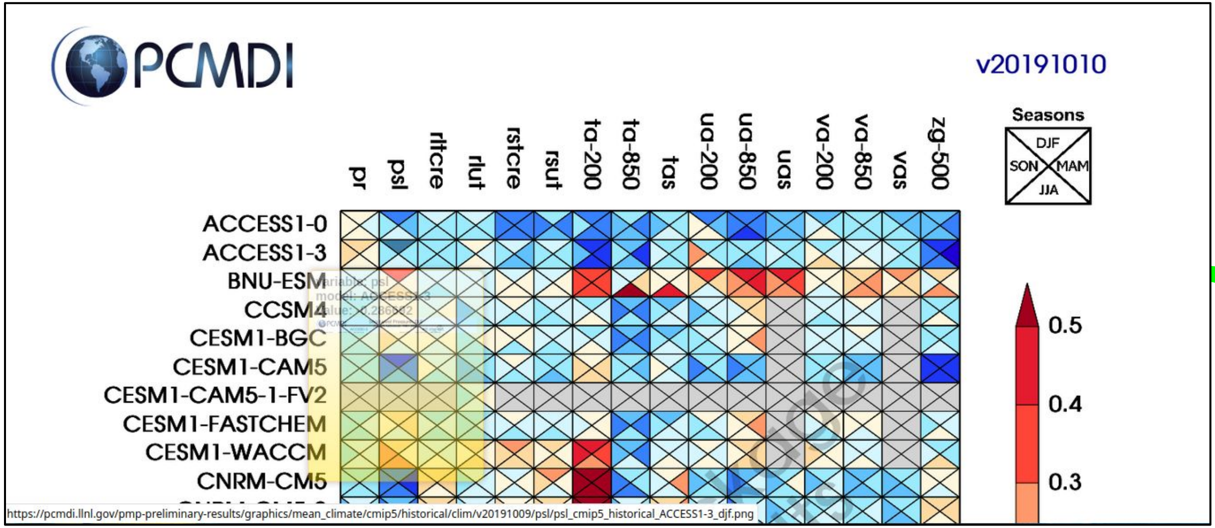
Different colors for model groups

Clickable cell linking to metric page

Show/Hide cell values

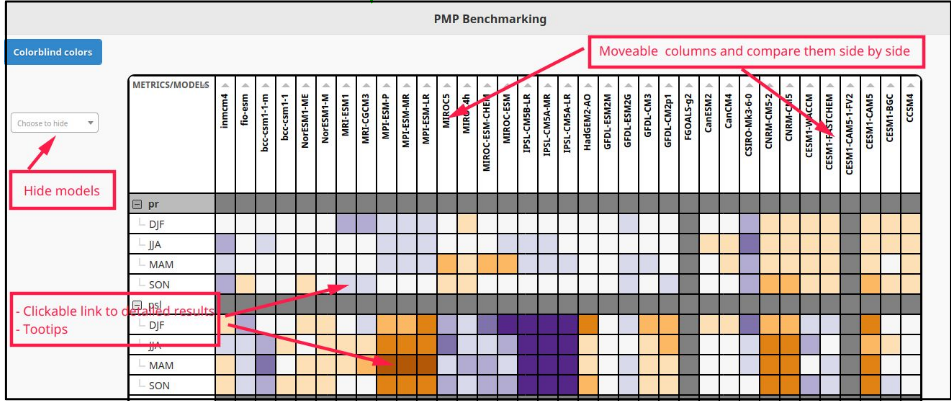
- **Tooltips:** show scores when mouse hovers the cells.
- **Column Hiding:** hide some models (columns) to focus into models of interest.
- **Column sorting:** sort the scores along the columns/models to see the best metric for the model.

Convert other diagnostic results for use in LMT dashboard



https://pcmdi.llnl.gov/pmp-preliminary-results/graphics/mean_climate/cmip5/historical/clim/v20191009/psl/psl_cmip5_historical_ACCESS1-3_djf.png

https://lmt.ornl.gov/tab_pmp



Moveable columns and compare them side by side

Hide models

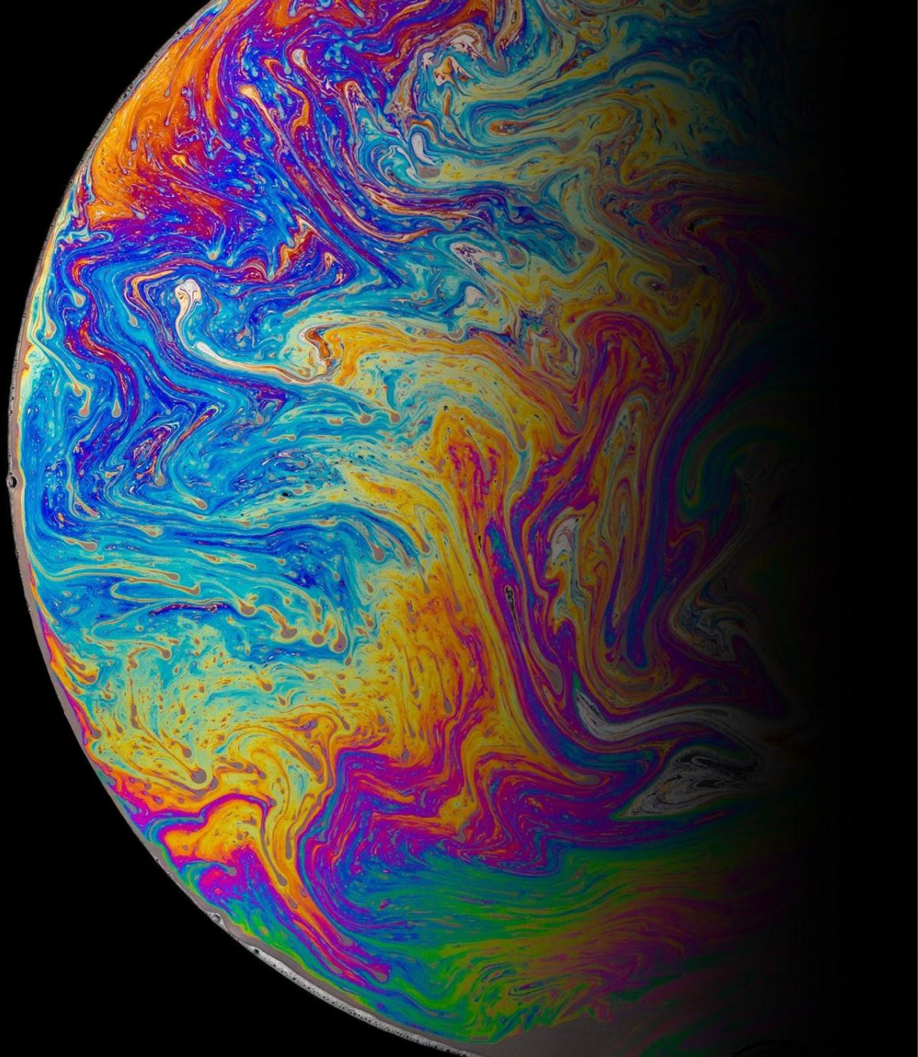
- Clickable link to detailed results
- Tootips

PMP: The Program for Climate Model Diagnostics and Intercomparison (PCMDI) Metrics Package (PMP)

- Clicking cell will go to maps of geographic distributions generated by PMP
- Our LMT dashboard can be used to study science questions like ENSO-BGC feedbacks

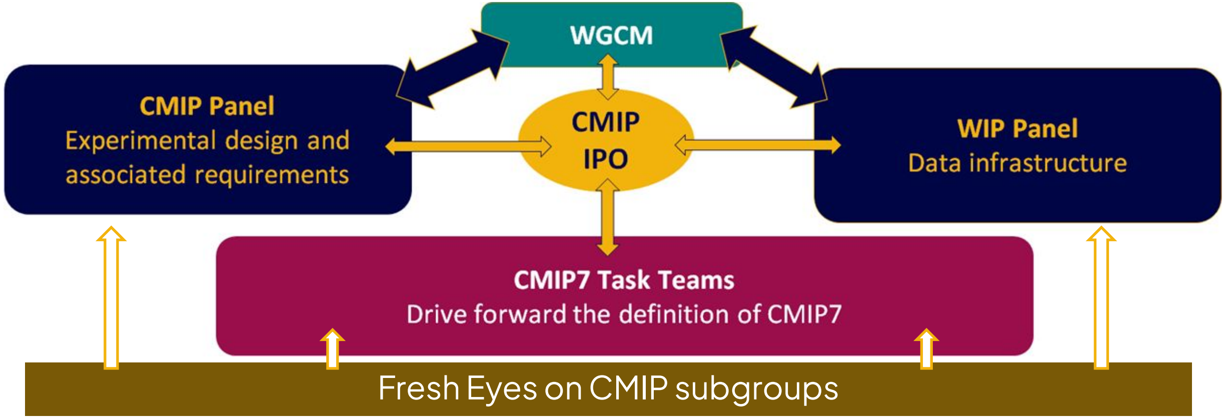
CMIP Coupled Model Intercomparison Project
Model Benchmarking

Climate Model Benchmarking for CMIP7



Forrest M. Hoffman and Birgit Hassler
CMIP Climate Model Benchmarking Task Team Co-Leads

CMIP and Preparations for CMIP7



WGCM
CMIP Panel
Experimental design and associated requirements
CMIP IPO
WIP Panel
Data infrastructure
CMIP7 Task Teams
Drive forward the definition of CMIP7
Fresh Eyes on CMIP subgroups

- CMIP is part of the WCRP's Earth System Modelling and Observation (ESMO) realm, and the Working Group on Coupled Modelling (WGCM)
- CMIP activities are coordinated by the CMIP Panel, the WGCM Infrastructure Panel (WIP), the CMIP IPO and the CMIP7 Task Teams

- Two new CMIP co-chairs were appointed (2022/2023):
  - Helene Hewitt (Met Office, UK)
  - John Dunne (NOAA GFDL, USA)
- The CMIP IPO, newly established in March 2022 and hosted at the European Space Agency's (ESA's) ECSAT site in Harwell, UK, is staffed by five people and is tasked with helping coordinate and support CMIP activities and responsibilities

CMIP Task Teams

- The CMIP Panel and WIP have established a number of Task Teams to support the design, scope, and definition of the next phase of CMIP and evolution of CMIP infrastructure and future operationalization
- Individual Task Teams aim to address specific topics (shown at right) and interact with each other to develop recommendations for the CMIP Panel and WIP

- Data Access
(Robert Pincus & Atef Ben-Nasser)
- Data Citation
(Martina Stockhause & Sasha Ames)
- Data Request
(Martin Juckes & Chloe Mackallah)
- Forcings
(Paul Durack & Vaishali Naik)
- Model Benchmarking
(Birgit Hassler & Forrest Hoffman)
- Model Documentation
(David Hassell & Guillaume Levavasseur)
- Strategic Ensemble Design
(Ben Sanderson & Isla Simpson)

The Model Benchmarking TT

- Rebecca Beadling, *USA*
- Ed Blockley, *UK*
- Jiwoo Lee, *USA*
- Valerio Lembo, *Italy*
- Jared Lewis, *Australia*
- Jianhua Lu, *China*
- Luke Madaus, *USA*
- Elizaveta Malinina, *Canada*
- Brian Medeiros, *USA*
- Wilfried Pokam Mba, *Cameroon*
- Enrico Scoccimarro, *Italy*
- Ranjini Swaminathan, *UK*

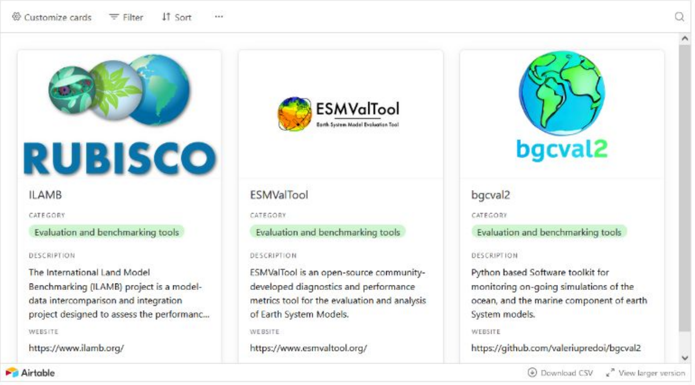
- Diversity in expertise (realms and methods), user group representation, gender, location, career stage
- Overarching goals:
  - Systematic and rapid performance assessment of the expected models participating in CMIP7 (including the model output and documentation)
  - Enhancing existing community evaluation tools that facilitate performance assessment of models
  - Integration of evaluation tools into CMIP publication workflows and fostering publication of their diagnostic outputs alongside the model output on the ESGF
- Collaboration with two Fresh Eyes on CMIP Subgroups
  - Model Evaluation
  - Data Analysis

#CMIP

Model Benchmarking Tools – Info “Cards” & Videos

- Main characteristics of (open source) benchmarking and evaluation tools available for analyses of CMIP-style data summarized in an overview “card” or an information video
- Collected information presented centrally on the CMIP website for easy access
- Cards can be filled out for all available open source benchmarking and evaluation tools if they can be used for CMIP data analysis; pre-defined questionnaire available on the CMIP website

https://wcrp-cmip.org/tools/model-benchmarking-and-evaluation-tools/



ILAMB

CATEGORY

Evaluation and benchmarking tools

DESCRIPTION

The International Land Model Benchmarking (ILAMB) project is a model-data intercomparison and integration project designed to assess the performanc...

WEBSITE

https://www.ilamb.org/

ESMValTool

CATEGORY

Evaluation and benchmarking tools

DESCRIPTION

ESMValTool is an open-source community-developed diagnostics and performance metrics tool for the evaluation and analysis of Earth System Models.

WEBSITE

https://www.esmvaltool.org/

bgcval2

CATEGORY

Evaluation and benchmarking tools

DESCRIPTION

Python based Software toolkit for monitoring on-going simulations of the ocean, and the marine component of earth System models.

WEBSITE

https://github.com/valeriupredoi/bgcval2

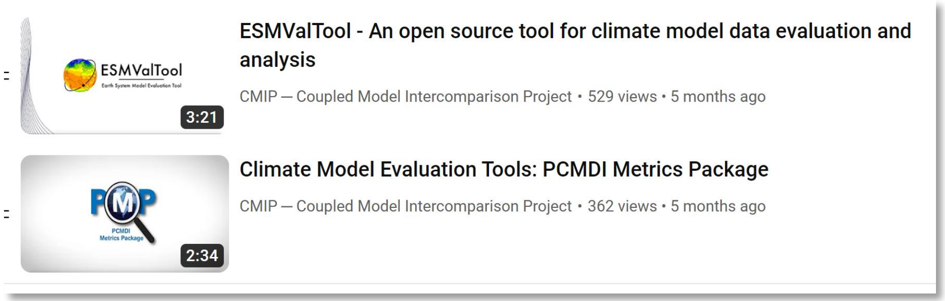
Status: first cards available

Started: October 2023

#CMIP

Model Benchmarking Tools – Information Videos

- Videos with descriptions of different benchmarking and evaluation tools
- Contain also the main characteristics of the different tools, just presented in a different way than the "cards"
- Videos can also be of different style
- All videos are presented in one central location linked to CMIP



ESMValTool - An open source tool for climate model data evaluation and analysis
CMIP — Coupled Model Intercomparison Project • 529 views • 5 months ago

Climate Model Evaluation Tools: PCMDI Metrics Package
CMIP — Coupled Model Intercomparison Project • 362 views • 5 months ago

□ More videos of tools welcome!

□ More info cards about evaluation/benchmarking tools welcome!

https://wcrp-cmip.org/tools/model-benchmarking-and-evaluation-tools/

Retrospective paper

- Definitions of "evaluation", "validation", and "benchmarking"
- Retrospective look at evolution of evaluation & benchmarking metrics
- What tools were available for CMIP6 (methods, philosophies, tools)?
- What approaches were used for CMIP6?
- Which of them worked well for CMIP6 and what did not work for CMIP6?
- Extensive information about different benchmarking and evaluation tools

Status: Currently being finalized

Planned submission: August 2024

What is the way forward?

- Based on the findings of the extensive information collected about different tools, and the retrospective paper – What do we think should be the benchmarking/evaluation focus for CMIP7?
- What framework would ideally be available for instantaneous benchmarking and evaluation at the time of data submission? Is such a framework even possible?
- How to avoid the bottlenecks encountered in CMIP6 benchmarking/ evaluation?
- Comprehensive community evaluation in near-real time possible?

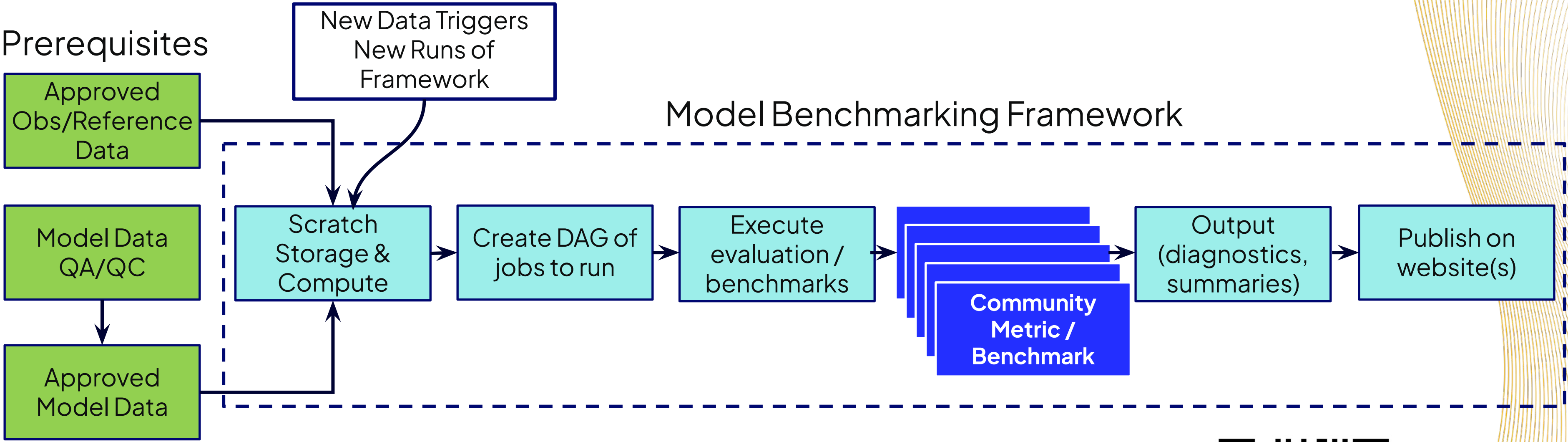
Status: Under development

Planned submission: Summer 2024

Other Planned TT Activities

- In collaboration with Fresh Eyes on CMIP groups
  - Scope out a Rapid Evaluation Framework for automated benchmarking capabilities at the time of AR7 Fast Track data publication
  - Develop scope for better quality assurance / quality control (QA/QC) for CMIP model output
  - Develop a white paper on observational data needs for model benchmarking, including uncertainties

Rapid Evaluation Framework Overview


Prerequisites
Approved Obs/Reference Data
Model Data QA/QC
Approved Model Data
New Data Triggers New Runs of Framework
Model Benchmarking Framework
Scratch Storage & Compute
Create DAG of jobs to run
Execute evaluation / benchmarks
Community Metric / Benchmark
Output (diagnostics, summaries)
Publish on website(s)

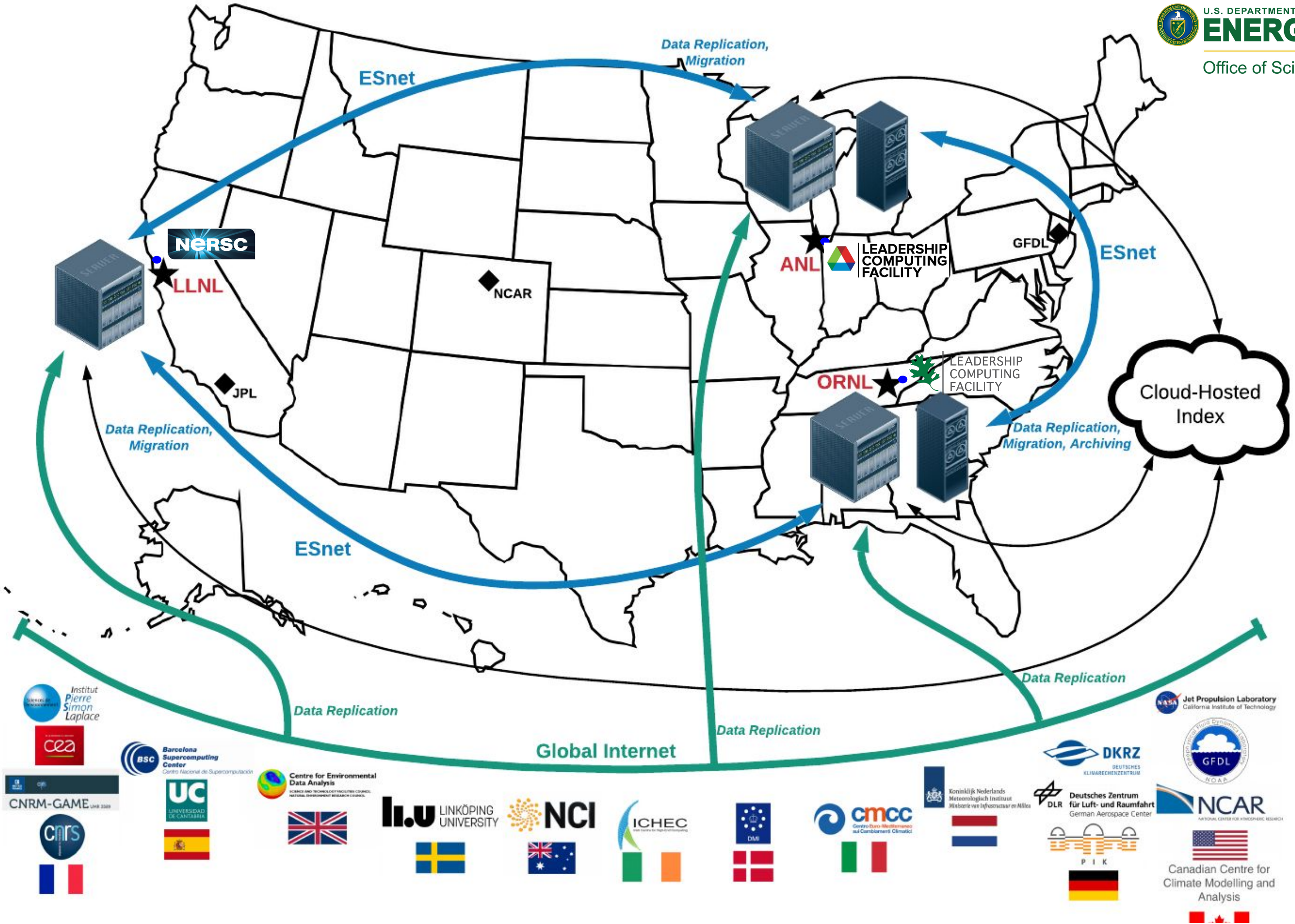
ESGF2 US

U.S. DEPARTMENT OF ENERGY

Office of Science

DOE's Next Generation Earth System Grid Federation

- As many as three nodes co-located at DOE's major computing facilities
- Replicating data from the global Federation
- Providing cloud indexing and tape archiving


Data Replication, Migration
ESnet
NERSC
LLNL
NCAR
JPL
ANL
LEADERSHIP COMPUTING FACILITY
GFDL
ESnet
ORNL
LEADERSHIP COMPUTING FACILITY
Data Replication, Migration
Data Replication, Migration, Archiving
Cloud-Hosted Index
ESnet
Data Replication
Data Replication
Data Replication
Global Internet
Institut Pierre Simon Laplace
cea
CNRM-GAME
cnrs
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
UC Universidad de Cantabria
Centre for Environmental Data Analysis
LINKÖPING UNIVERSITY
NCI
ICHEC
DMI
cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici
Koninklijk Nederlands Meteorologisch Instituut
DKRZ Deutsches Klimarechenzentrum
DLR Deutsches Zentrum für Luft- und Raumfahrt German Aerospace Center
PIK
Jet Propulsion Laboratory California Institute of Technology
GFDL NOAA
NCAR
Canadian Centre for Climate Modelling and Analysis

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

Questions?

References (1/2)

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 (2/2)

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.

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.