U.S. DEPARTMENT of ENERGY

International Land Model Benchmarking (ILAMB)

Forrest M. Hoffman[1,2], Nathan Collier[1], Mingquan Mu[3], Min Xu[1], Gretchen Keppel-Aleks[4], David M. Lawrence[5], Charles D. Koven[6], Weiwei Fu[3], 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

ILAMB

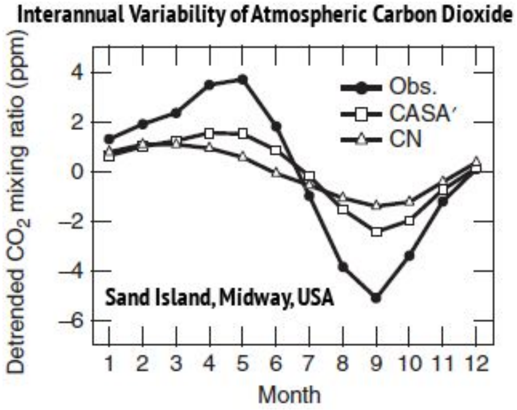
RUBISCO

2025 ILAMB Hybrid Meeting

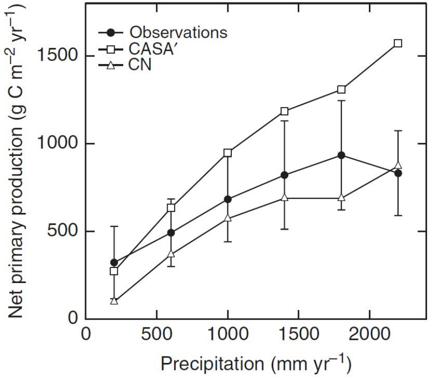
December 11, 2025

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)

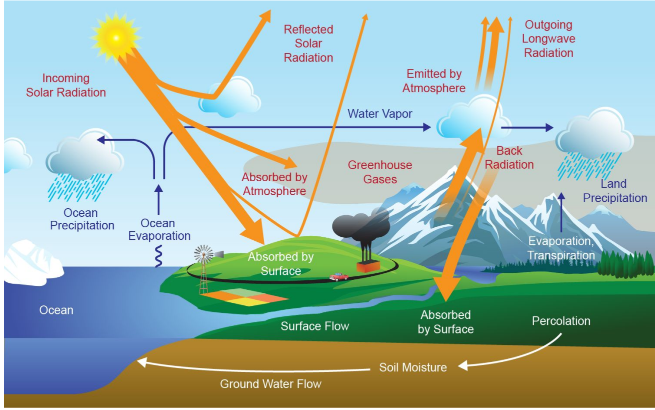
Why Benchmark Models?

- 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

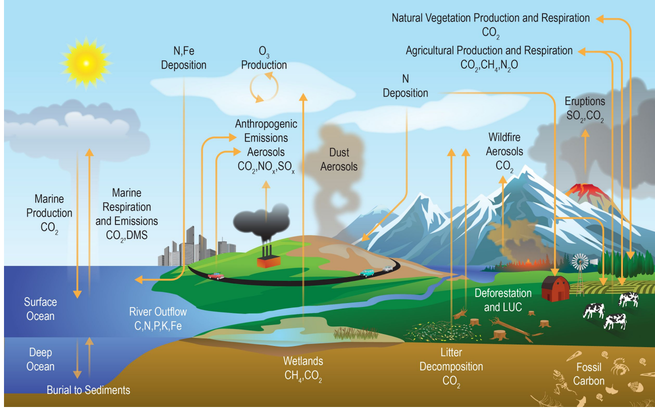
What is ILAMB?

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



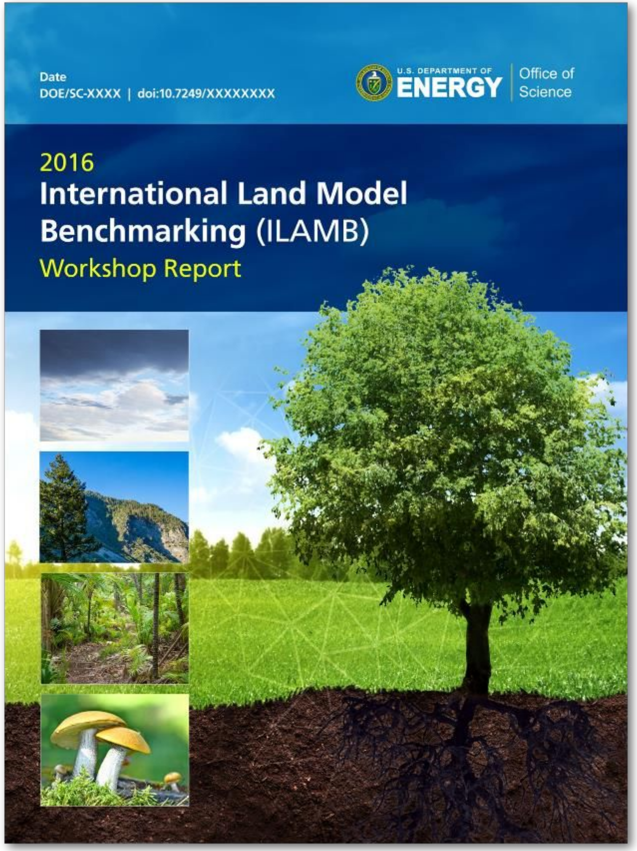
Carbon and Biogeochemical Cycles

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)

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



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

U.S. DEPARTMENT OF ENERGY | Office of Science

2016
International Land Model Benchmarking (ILAMB)
Workshop Report

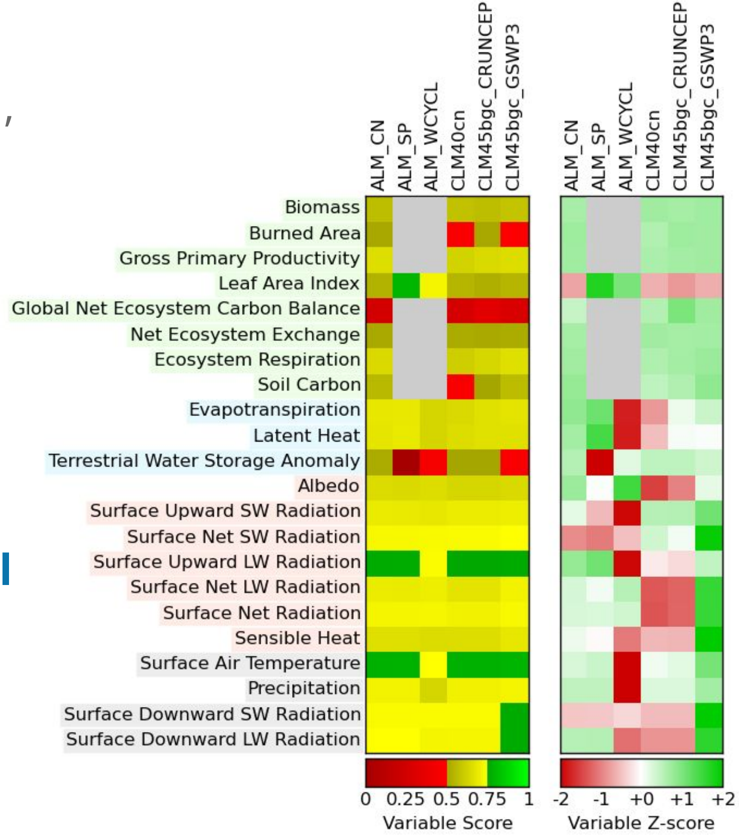
(Hoffman et al., 2017)

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

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
- **ILAMBv3** *Coming Soon!*
- **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}) →

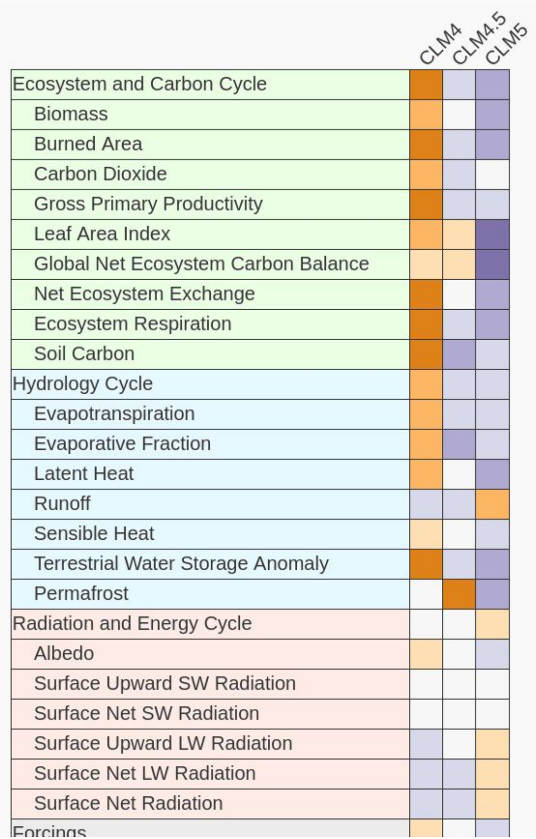
$$S_{\text{overall}} = \frac{S_{\text{bias}} + 2S_{\text{rmse}} + S_{\text{phase}} + S_{\text{iav}} + S_{\text{dist}}}{1 + 2 + 1 + 1 + 1}$$

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

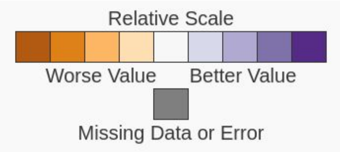
ILAMB Current Variables

- **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 Multiple Generations of CLM



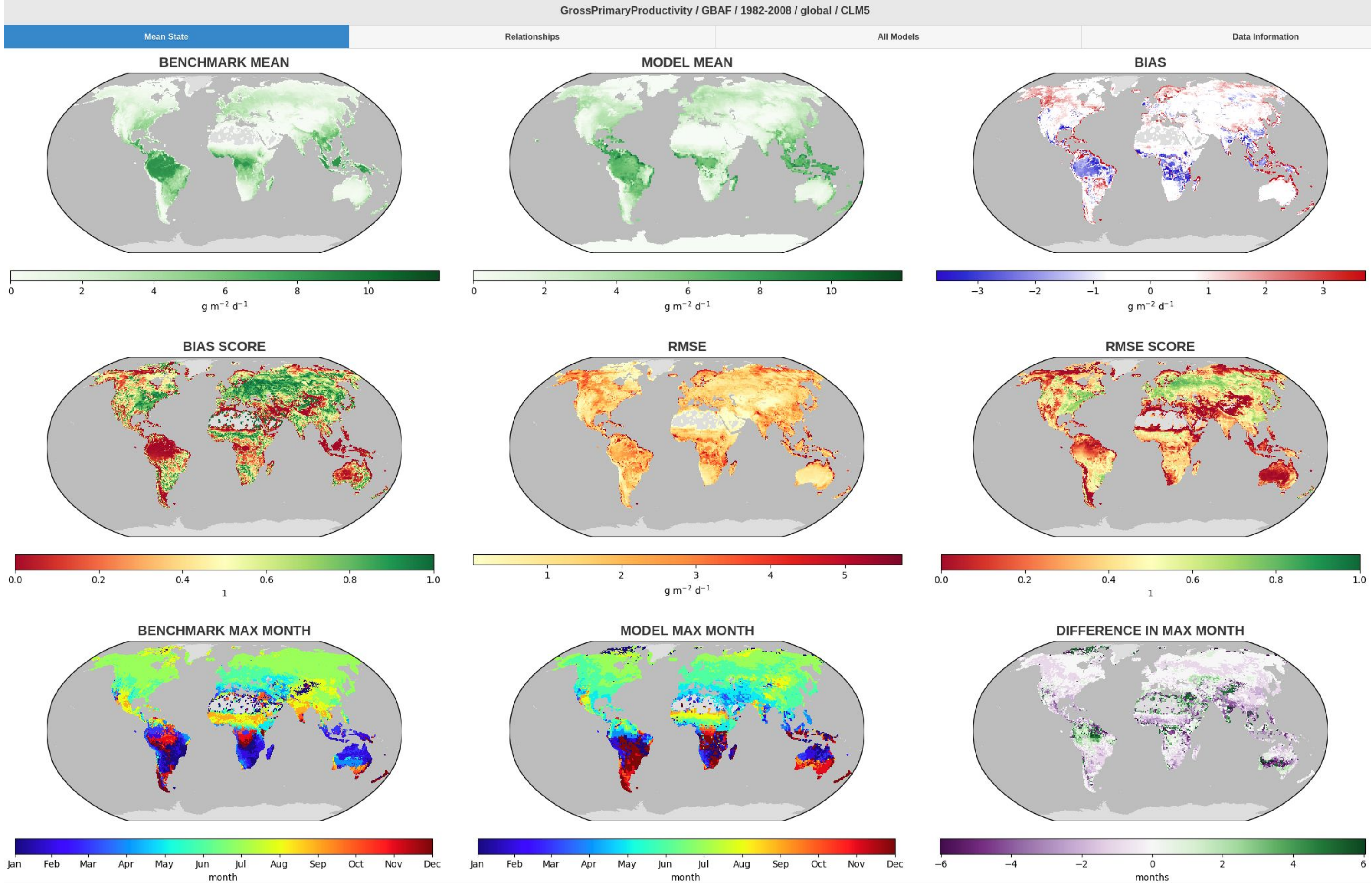
	CLM4	CLM4.5	CLM5
Ecosystem and Carbon Cycle	🟧	🟪	🟪
Biomass	🟧	⬜	🟪
Burned Area	🟧	🟪	🟪
Carbon Dioxide	🟧	🟪	⬜
Gross Primary Productivity	🟧	🟪	🟪
Leaf Area Index	🟧	🟧	🟪
Global Net Ecosystem Carbon Balance	🟧	🟧	🟪
Net Ecosystem Exchange	🟧	⬜	🟪
Ecosystem Respiration	🟧	🟪	🟪
Soil Carbon	🟧	🟪	🟪
Hydrology Cycle	🟧	🟪	🟪
Evapotranspiration	🟧	🟪	🟪
Evaporative Fraction	🟧	🟪	🟪
Latent Heat	🟧	⬜	🟪
Runoff	🟪	🟪	🟧
Sensible Heat	🟧	⬜	🟪
Terrestrial Water Storage Anomaly	🟧	🟪	🟪
Permafrost	⬜	🟧	🟪
Radiation and Energy Cycle	⬜	⬜	🟧
Albedo	🟧	⬜	🟪
Surface Upward SW Radiation	⬜	⬜	⬜
Surface Net SW Radiation	⬜	⬜	⬜
Surface Upward LW Radiation	🟪	⬜	🟧
Surface Net LW Radiation	🟪	🟪	🟧
Surface Net Radiation	🟪	🟪	🟧
Forcings	🟧	⬜	🟪



- 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

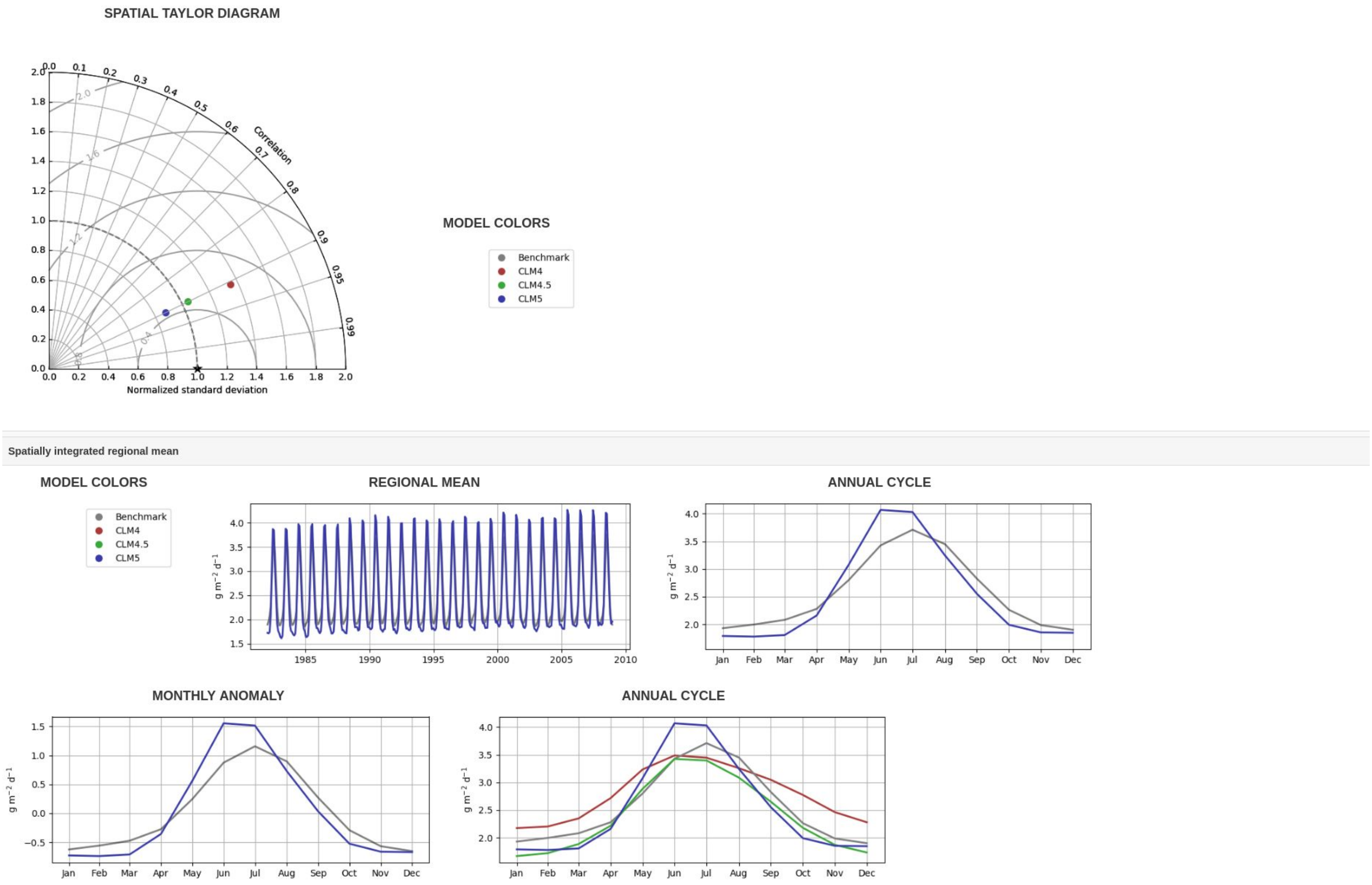
http://webext.cgd.ucar.edu/I20TR/_build_set1F/
(Lawrence et al., 2019)

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

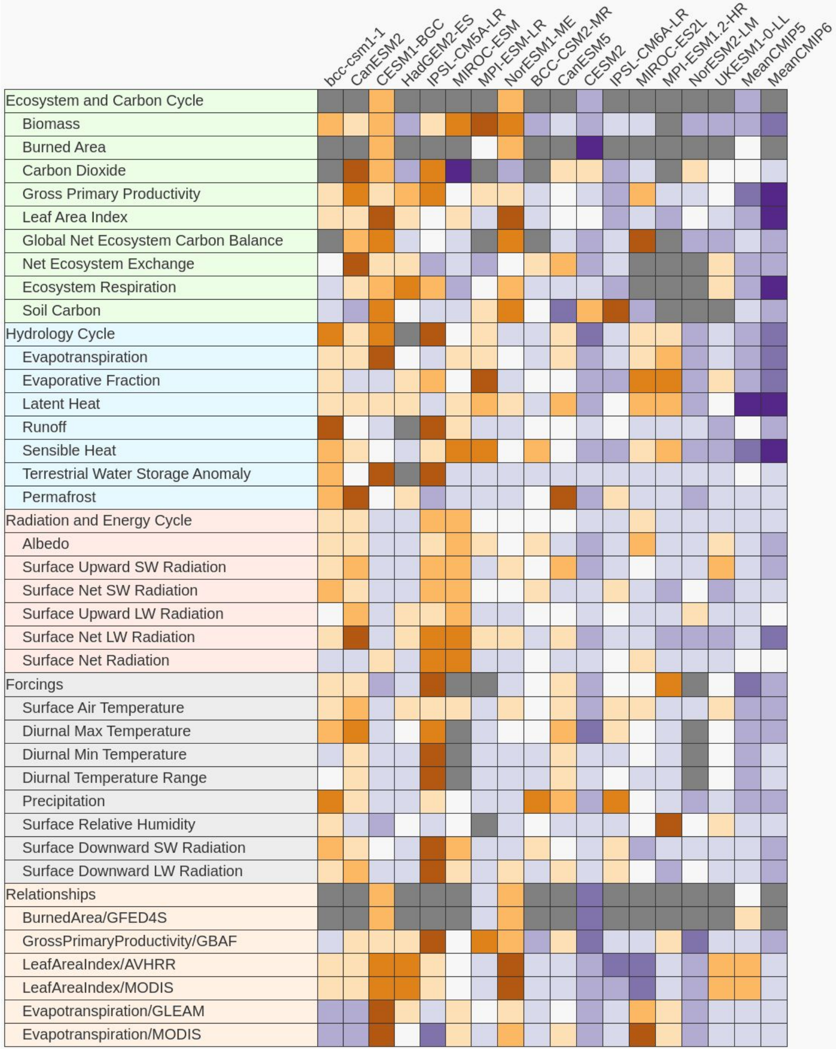
ANNUAL CYCLE

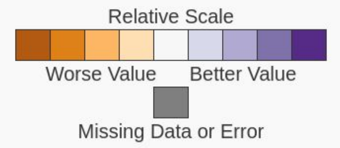
MONTHLY ANOMALY

ANNUAL CYCLE

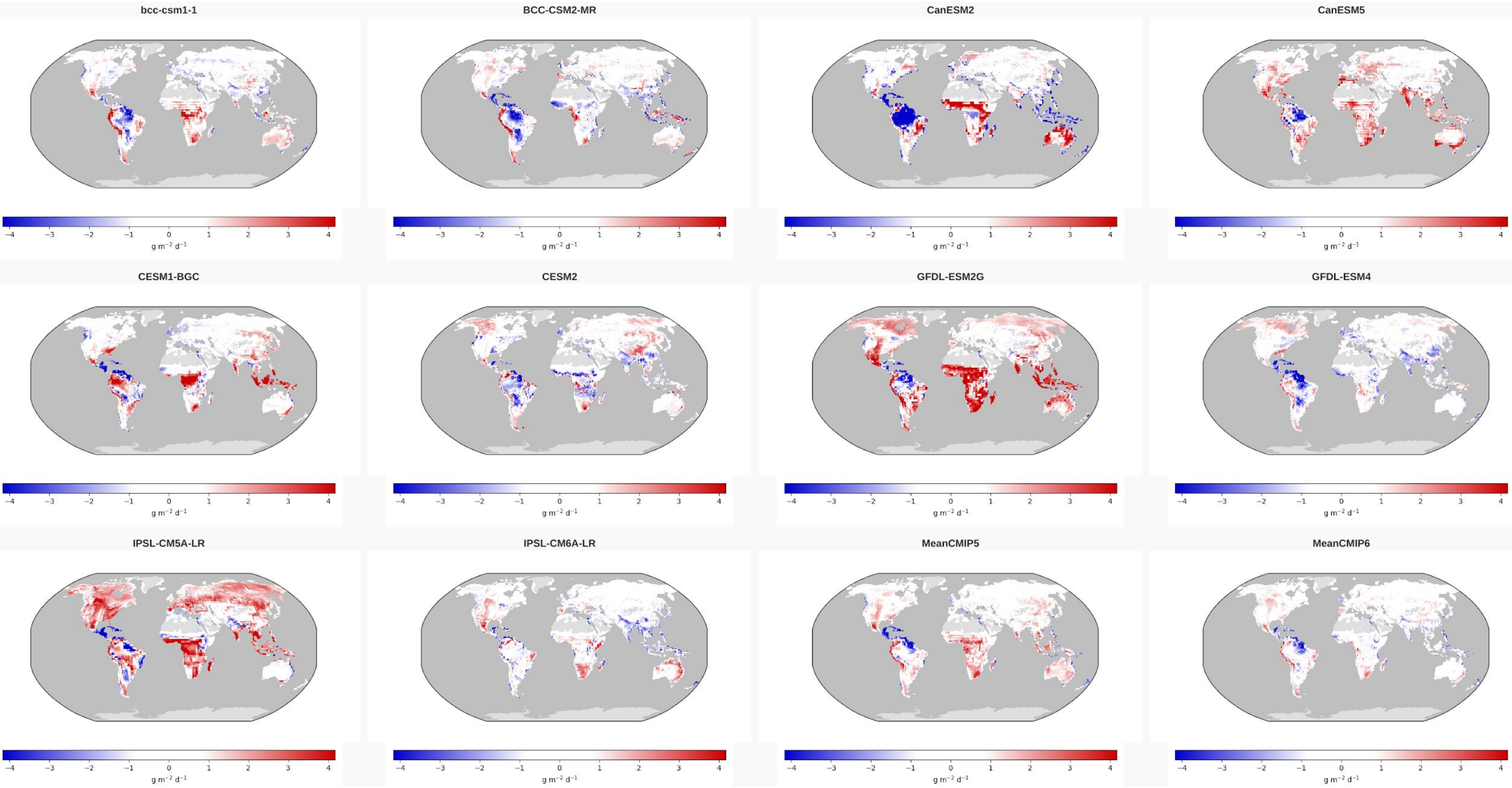
CMIP5 vs. CMIP6 Models

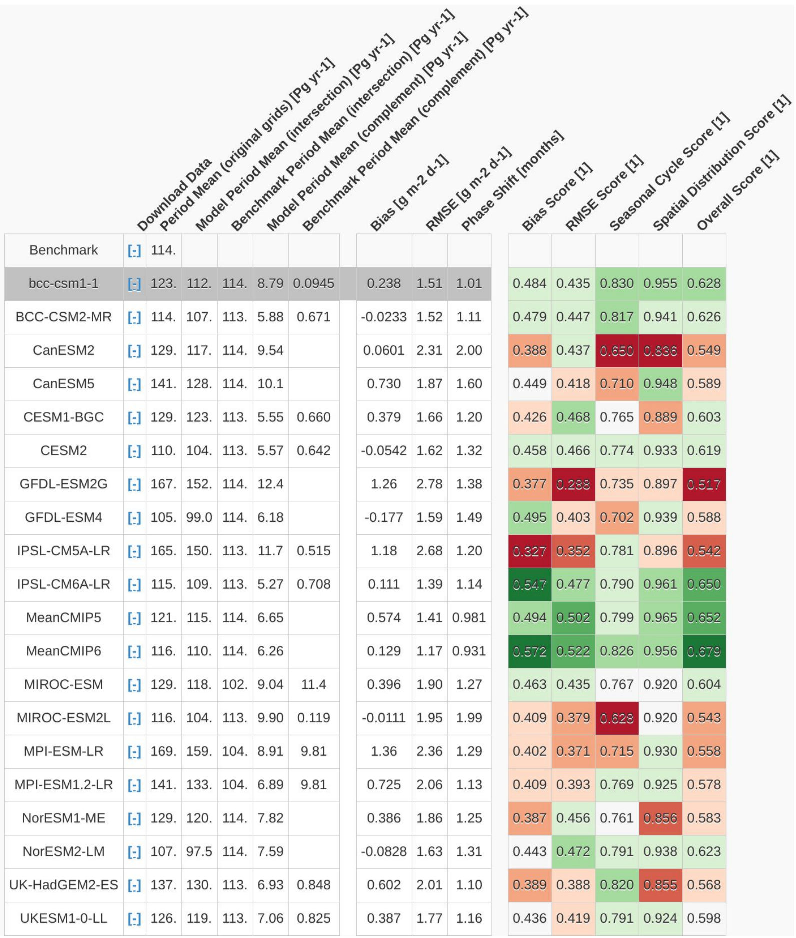
- The CMIP6 suite of land models (right) has improved over the CMIP5 suite of land models (left)
- The multi-model mean outperforms any single model for each suite of models
- The multi-model mean CMIP6 land model is the “best model” overall
- Why did CMIP6 land models improve?





(Hoffman et al., in prep)


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

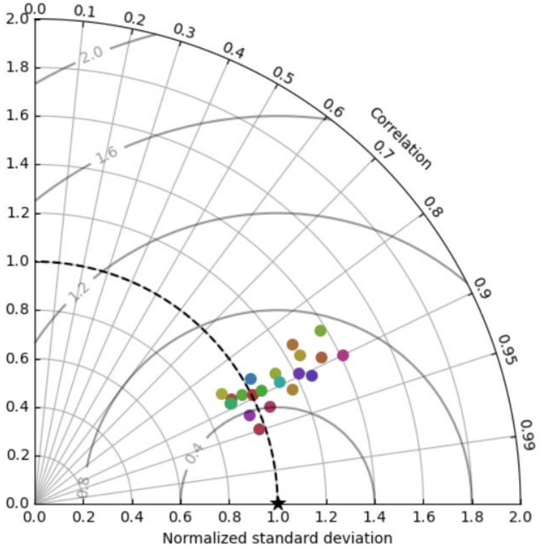


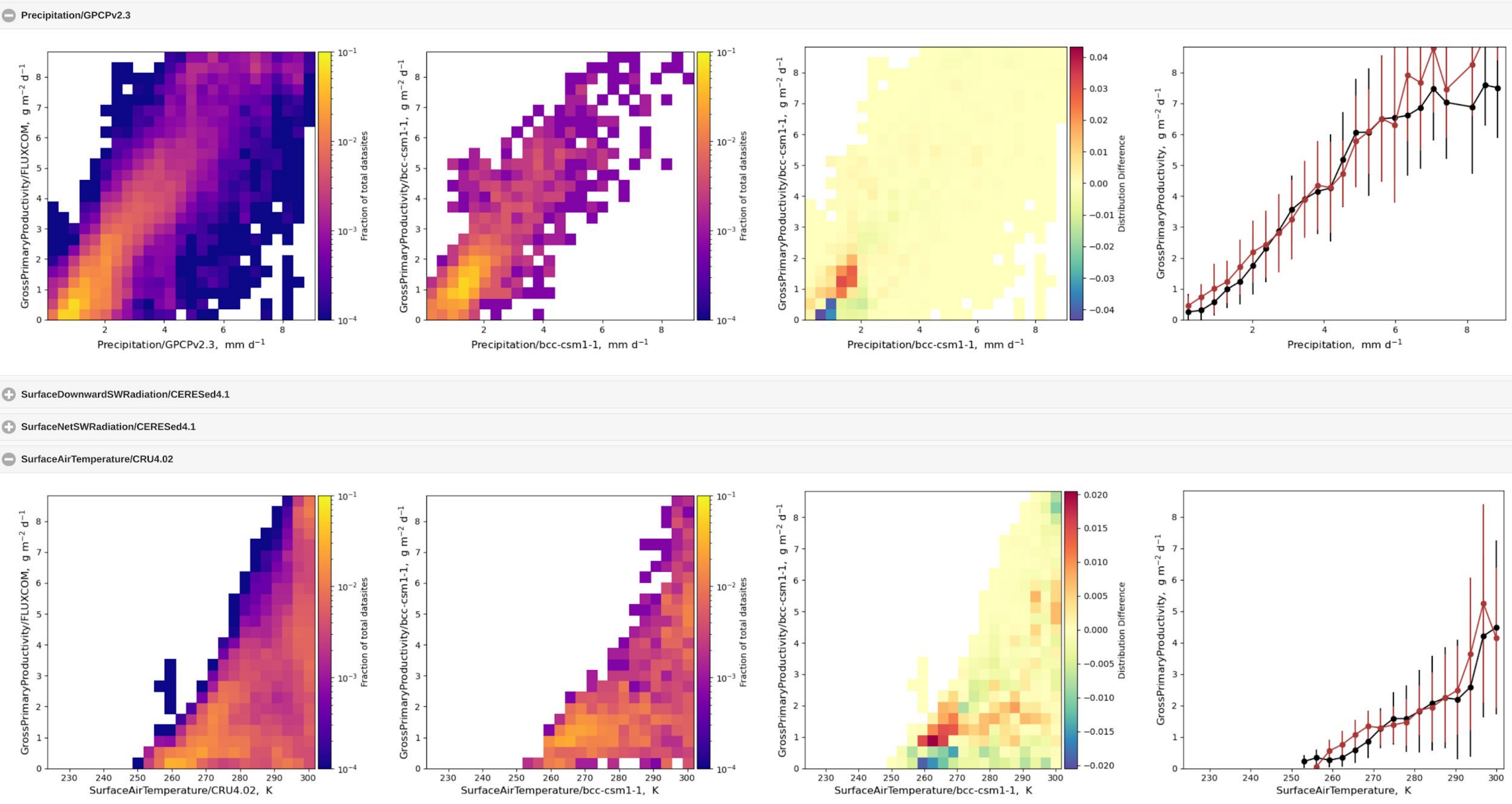
	Download Data	Period Mean (original grids) [Pg yr-1]	Model Period Mean (intersection) [Pg yr-1]	Benchmark Period Mean (intersection) [Pg yr-1]	Model Period Mean (complement) [Pg yr-1]	Benchmark Period Mean (complement) [Pg yr-1]	Bias [g m-2 d-1]	RMSE [g m-2 d-1]	Phase Shift [months]	Bias Score [1]	RMSE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

Gross Primary Productivity

- 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

GrossPrimaryProductivity/FLUXCOM, g m^{-2} d^{-1}
Precipitation/GPCPv2.3, mm d^{-1}
Fraction of total datasites
10^{-1}
10^{-2}
10^{-3}
10^{-4}
GrossPrimaryProductivity/bcc-csm1-1, g m^{-2} d^{-1}
Precipitation/bcc-csm1-1, mm d^{-1}
Distribution Difference
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
GrossPrimaryProductivity, g m^{-2} d^{-1}
Precipitation, mm d^{-1}
0
1
2
3
4
5
6
7
8

SurfaceDownwardSWRadiation/CERESed4.1

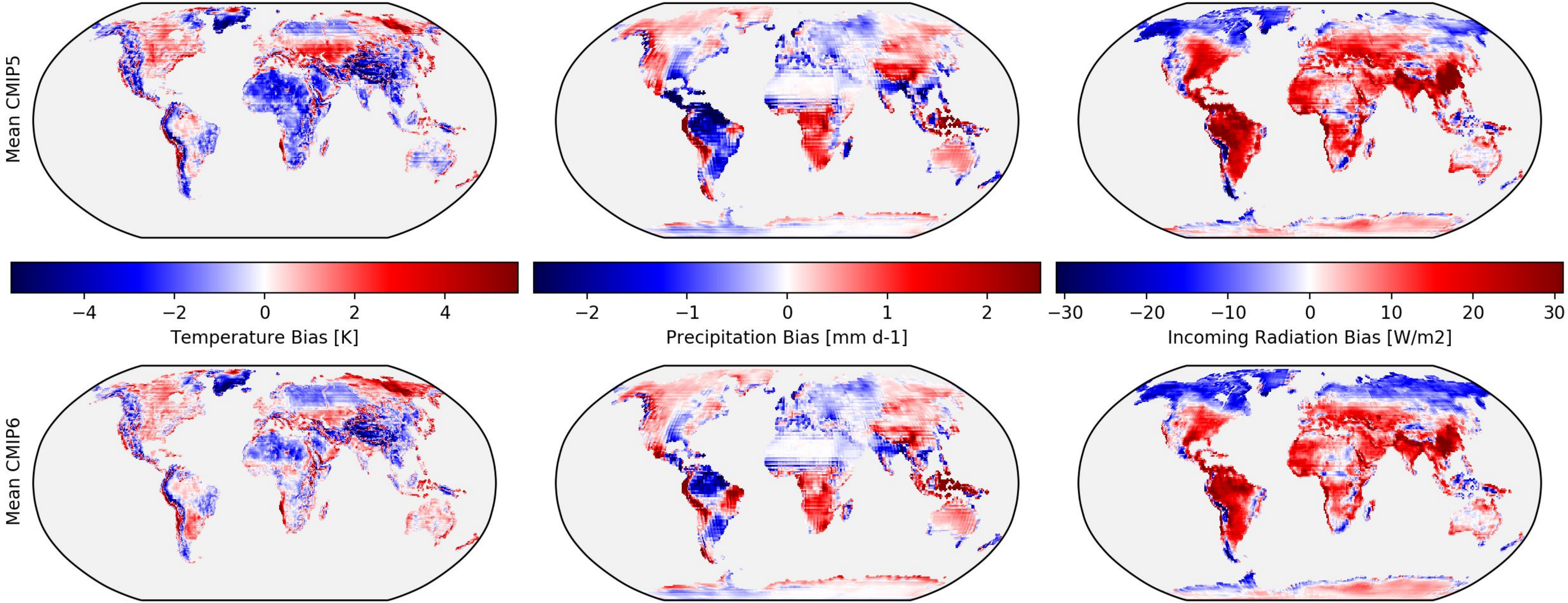
SurfaceNetSWRadiation/CERESed4.1

SurfaceAirTemperature/CRU4.02

GrossPrimaryProductivity/FLUXCOM, g m^{-2} d^{-1}
SurfaceAirTemperature/CRU4.02, K
Fraction of total datasites
GrossPrimaryProductivity/bcc-csm1-1, g m^{-2} d^{-1}
SurfaceAirTemperature/bcc-csm1-1, K
Distribution Difference
0.020
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.020
GrossPrimaryProductivity, g m^{-2} d^{-1}
SurfaceAirTemperature, K
230
240
250
260
270
280
290
300

Reasons for Land Model Improvements

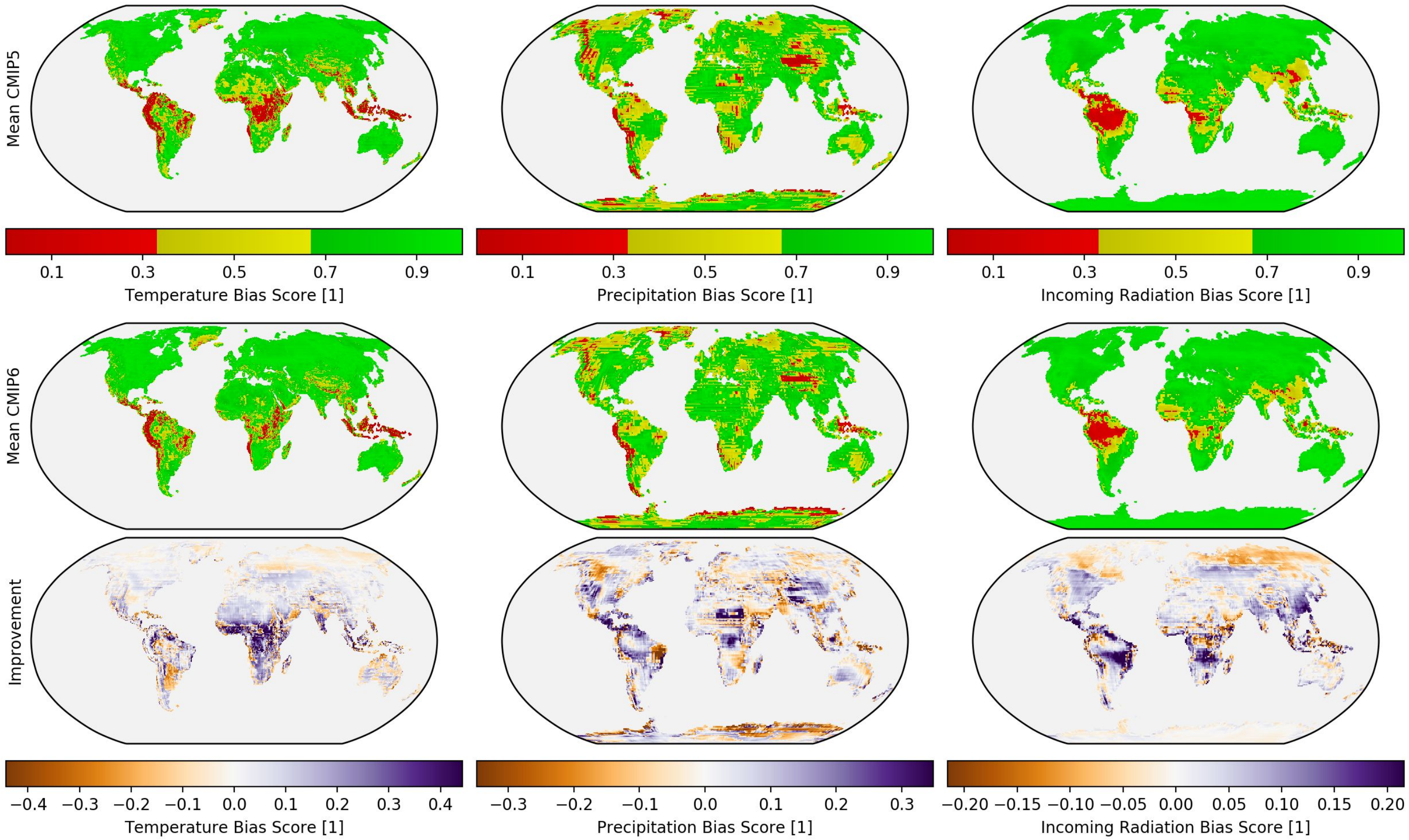
ESM improvements in climate forcings (temperature, precipitation, radiation) likely partially drove improvements exhibited by land carbon cycle models



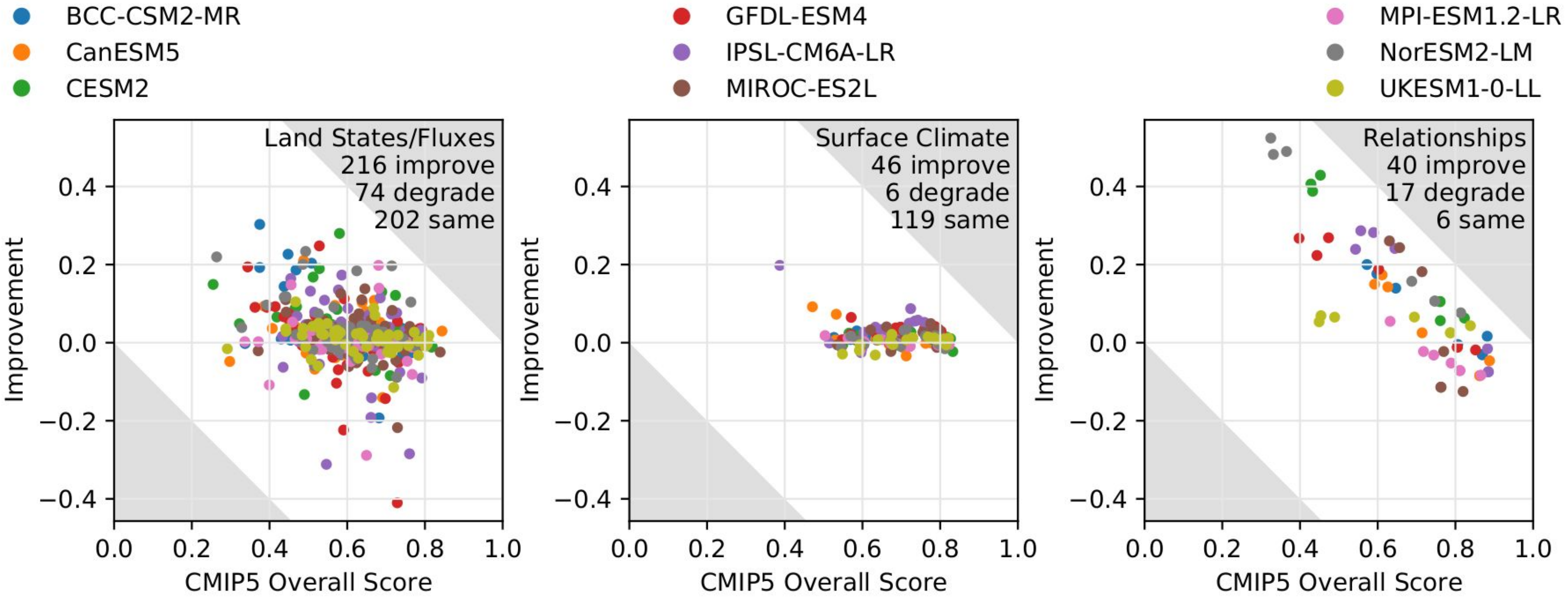
(Hoffman et al., in prep)

Reasons for Land Model Improvements

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



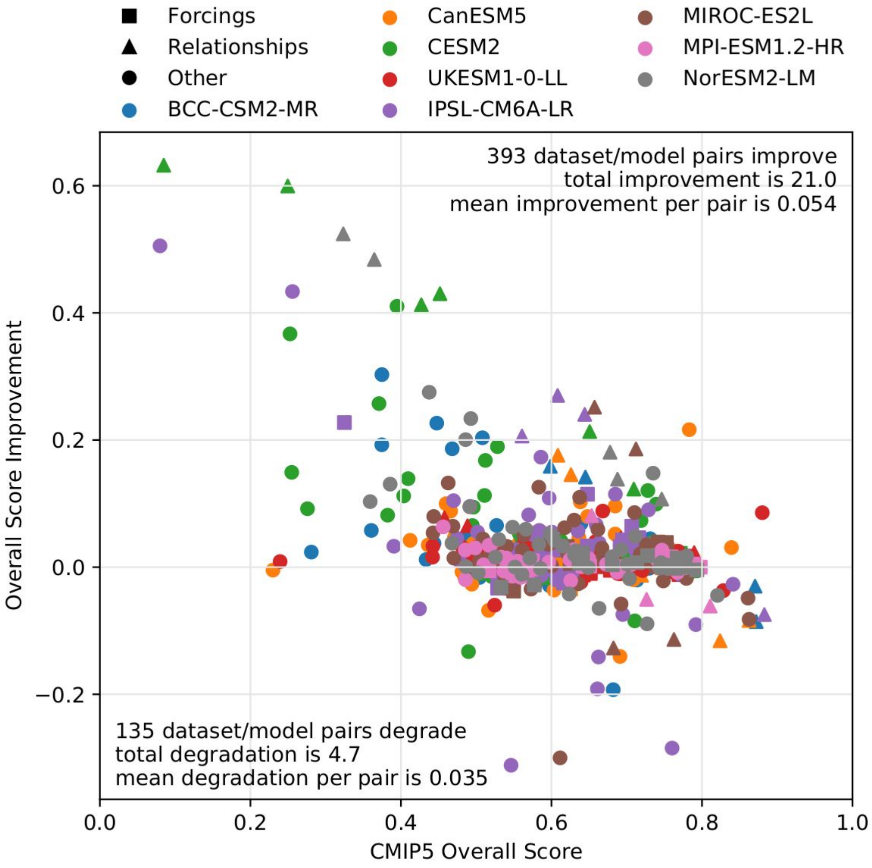
(Hoffman et al., in prep)



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.

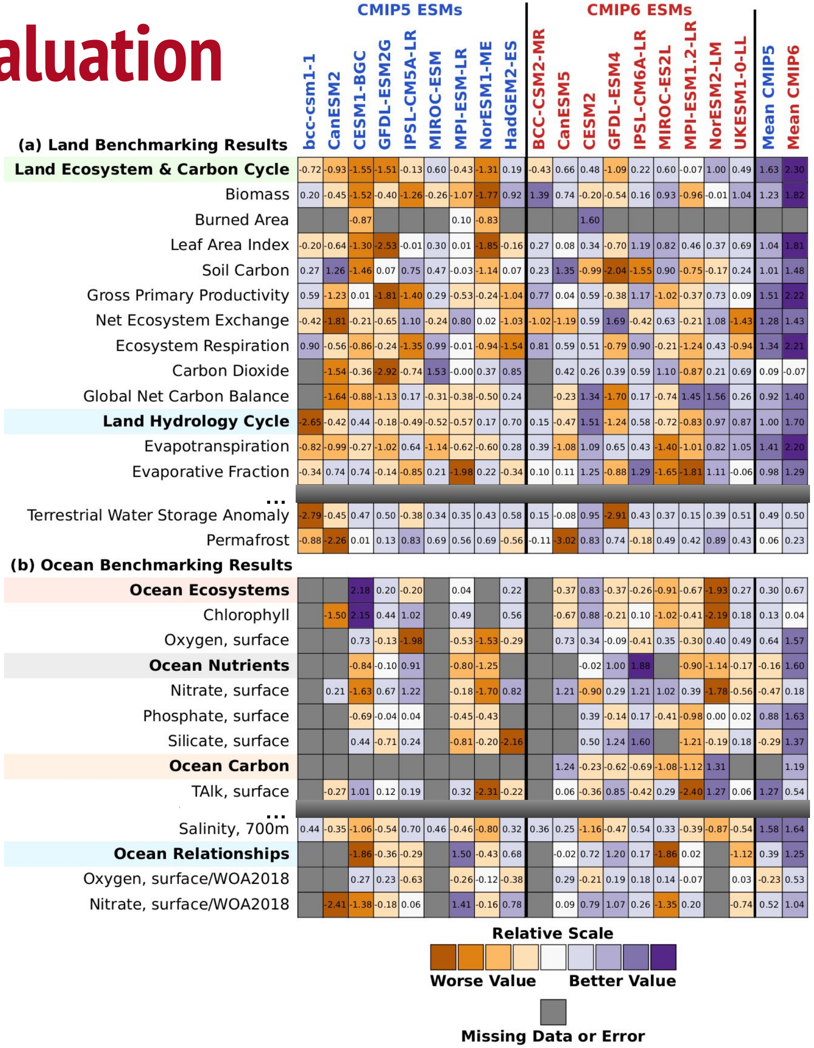
Reasons for Land Model Improvements

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 have 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										
	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
(a) Land Benchmarking Results																				
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

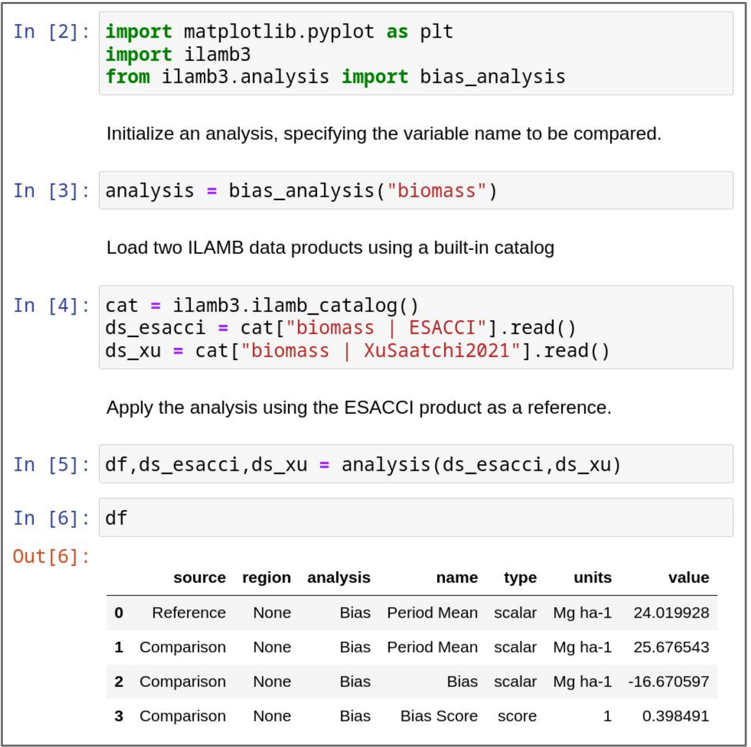
Relative Scale

Worse Value								Better Value
🟫	🟧	🟧	🟨	⬜	🟪	🟪	🟪	🟪

⬛ **Missing Data or Error**

ILAMBv3: Coming Soon!

- Continuously updated documentation as new modules are being developed: https://ilamb3.readthedocs.io/



```
In [2]: import matplotlib.pyplot as plt
        import ilamb3
        from ilamb3.analysis import bias_analysis
```

Initialize an analysis, specifying the variable name to be compared.

```
In [3]: analysis = bias_analysis("biomass")
```

Load two ILAMB data products using a built-in catalog

```
In [4]: cat = ilamb3.ilamb_catalog()
        ds_esacci = cat["biomass | ESACCI"].read()
        ds_xu = cat["biomass | XuSaatchi2021"].read()
```

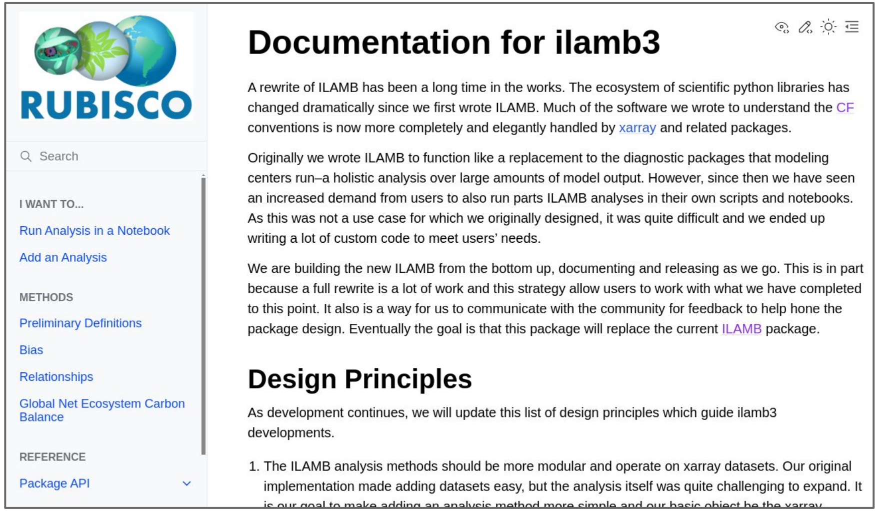
Apply the analysis using the ESACCI product as a reference.

```
In [5]: df,ds_esacci,ds_xu = analysis(ds_esacci,ds_xu)
```

```
In [6]: df
```

Out[6]:

	source	region	analysis	name	type	units	value
0	Reference	None	Bias	Period Mean	scalar	Mg ha-1	24.019928
1	Comparison	None	Bias	Period Mean	scalar	Mg ha-1	25.676543
2	Comparison	None	Bias	Bias	scalar	Mg ha-1	-16.670597
3	Comparison	None	Bias	Bias Score	score	1	0.398491



Documentation for ilamb3

A rewrite of ILAMB has been a long time in the works. The ecosystem of scientific python libraries has changed dramatically since we first wrote ILAMB. Much of the software we wrote to understand the CF conventions is now more completely and elegantly handled by xarray and related packages.

Originally we wrote ILAMB to function like a replacement to the diagnostic packages that modeling centers run–a holistic analysis over large amounts of model output. However, since then we have seen an increased demand from users to also run parts ILAMB analyses in their own scripts and notebooks. As this was not a use case for which we originally designed, it was quite difficult and we ended up writing a lot of custom code to meet users' needs.

We are building the new ILAMB from the bottom up, documenting and releasing as we go. This is in part because a full rewrite is a lot of work and this strategy allow users to work with what we have completed to this point. It also is a way for us to communicate with the community for feedback to help hone the package design. Eventually the goal is that this package will replace the current ILAMB package.

Design Principles

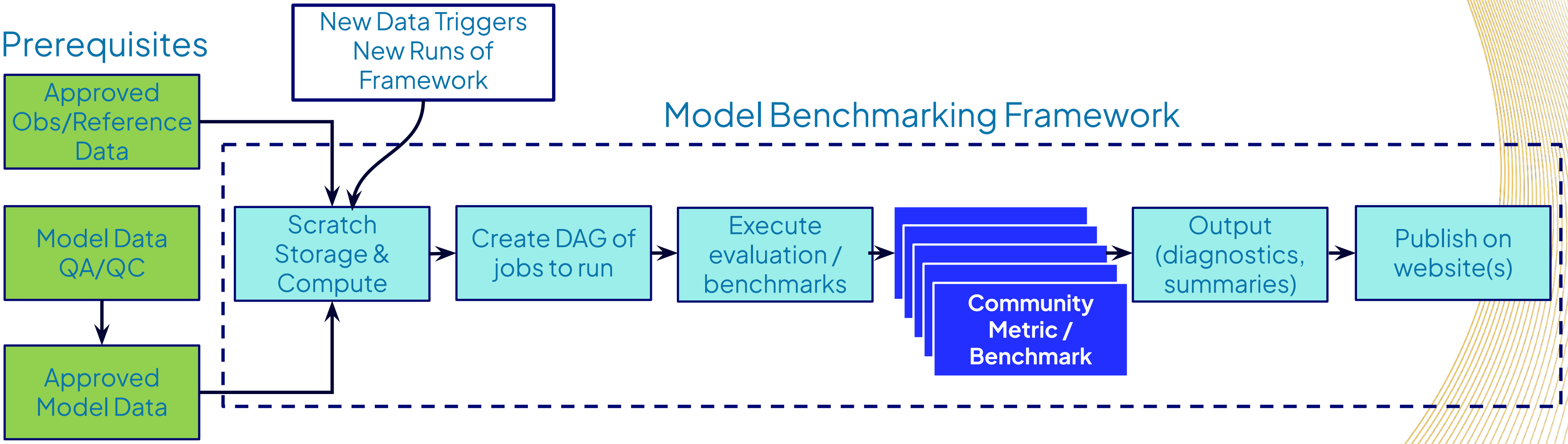
As development continues, we will update this list of design principles which guide ilamb3 developments.

1. The ILAMB analysis methods should be more modular and operate on xarray datasets. Our original implementation made adding datasets easy, but the analysis itself was quite challenging to expand. It is our goal to make adding an analysis method more simple and our basic object be the xarray

- ILAMBv3 allows for analysis methods to be imported into Python scripts and Jupyter notebooks and used to produce the scalars and plots synthesized in the full analysis
- At left, the ILAMB bias analysis is applied to two reference data products and show a table of scalar values

CMIP Rapid Evaluation Framework Overview

The Coupled Model Intercomparison Project (CMIP) Model Benchmarking Task Team developed a system specification for a Rapid Evaluation Framework (REF) that would leverage community benchmarking metrics to evaluate CMIP model output as they are submitted to the Earth System Grid Federation (ESGF)


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)

CMIP Benchmarking Task Team – Birgit Hassler, Forrest Hoffman, & Ranjini Swaminathan, Co-leads

Find out more about the CMIP Rapid Evaluation Framework (REF)

CMIP
Coupled Model
Intercomparison
Project

The CMIP AR7 Fast Track Rapid Evaluation Framework (AR7 FT REF) will be a complete end to end system providing a systematic and rapid performance assessment of the expected models participating in the CMIP AR7 Fast Track, supporting the next IPCC Assessment Report 7 (AR7) cycle.

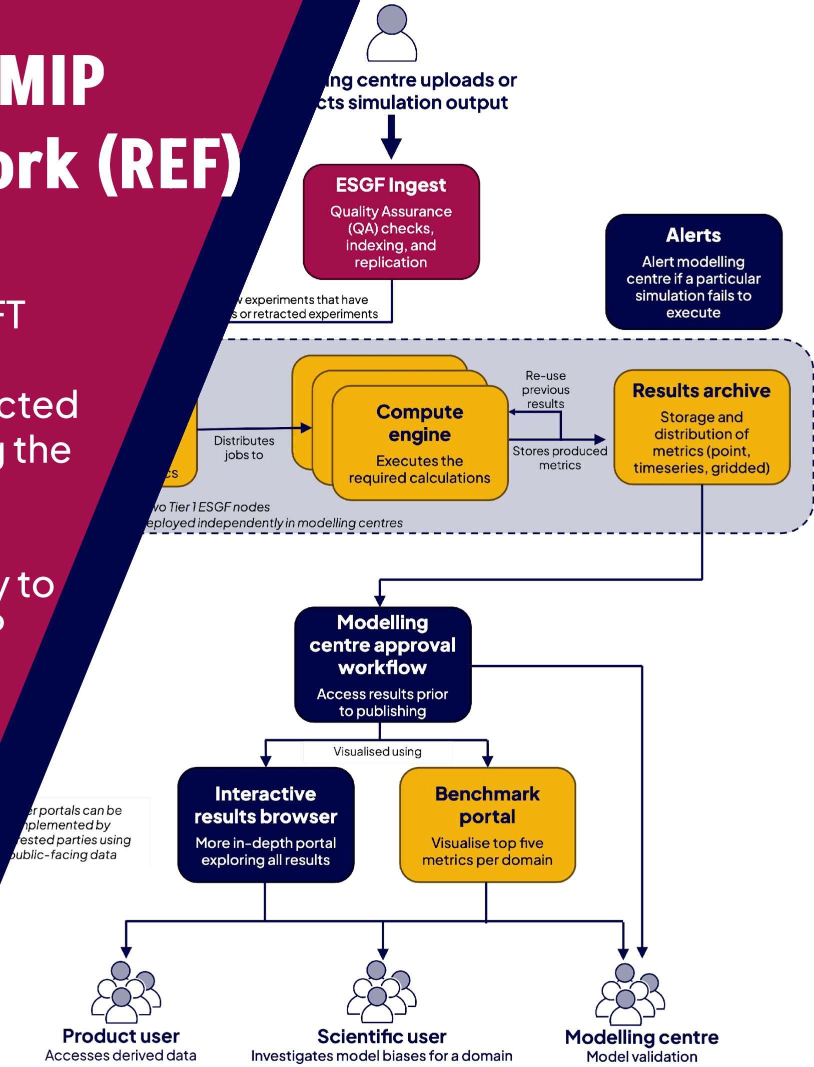
The REF is designed to be a starting point for the community to develop and build upon, with applications across the WCRP and beyond.

Find out more at
wcrp-cmip.org/cmip7/rapid-evaluation-framework/

This project has been made possible by funding from:

U.S. DEPARTMENT of ENERGY

esa



ng centre uploads or
cts simulation output

ESGF Ingest
Quality Assurance (QA) checks, indexing, and replication

v experiments that have
s or retracted experiments

Alerts
Alert modelling centre if a particular simulation fails to execute

cs

Distributes jobs to

Compute engine
Executes the required calculations

Re-use previous results

Stores produced metrics

Results archive
Storage and distribution of metrics (point, timeseries, gridded)

vo Tier 1 ESGF nodes
eployed independently in modelling centres

Modelling centre approval workflow
Access results prior to publishing

Visualised using

Interactive results browser
More in-depth portal exploring all results

Benchmark portal
Visualise top five metrics per domain

r portals can be
nplemented by
ested parties using
ublic-facing data

Product user
Accesses derived data

Scientific user
Investigates model biases for a domain

Modelling centre
Model validation

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