



U.S. DEPARTMENT
of **ENERGY**

Systematic Evaluation of Earth System Models

*Forrest M. Hoffman¹, Nathan Collier¹, Jitendra Kumar¹, Min Xu¹, Elias Massoud¹,
Mingquan Mu², David M. Lawrence³, Charles D. Koven⁴, Gretchen Keppel-Aleks⁵,
Weiwei Fu², and James T. Randerson²*

¹Oak Ridge National Laboratory, Oak Ridge, TN, USA

²University of California, Irvine, CA, USA

³National Center for Atmospheric Research, Boulder, CO, USA

⁴Lawrence Berkeley National Laboratory, Berkeley, CA, USA

⁵University of Michigan, Ann Arbor, MI, USA

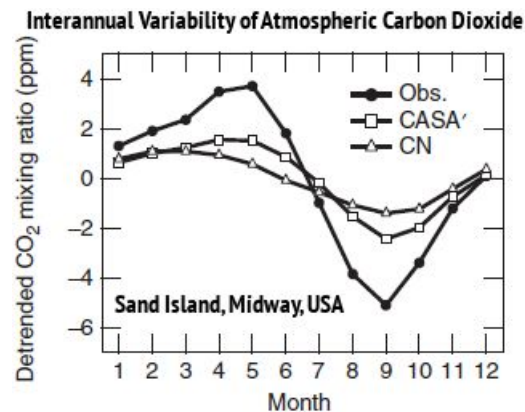
NGEE Arctic MODEX Workshop – Santa Fe, New Mexico

January 15, 2026

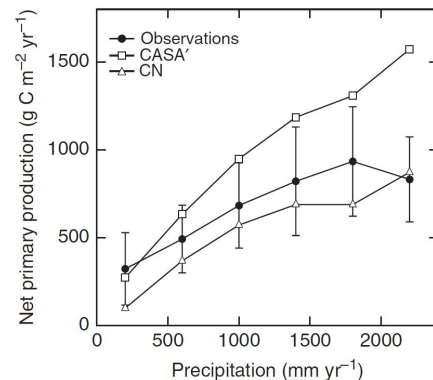


What is a Benchmark?

- 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 composition

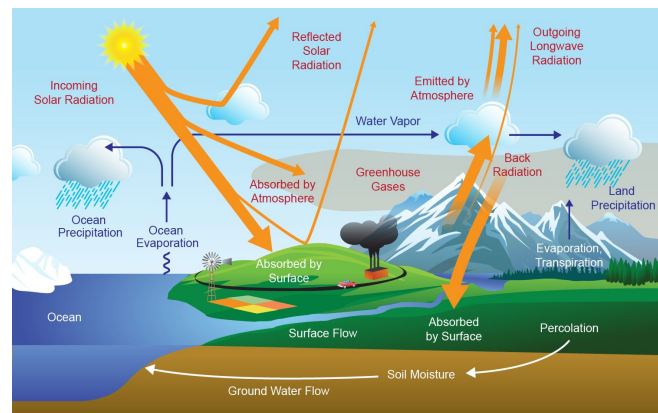


Models may reproduce correct responses over only a limited range of forcing variables

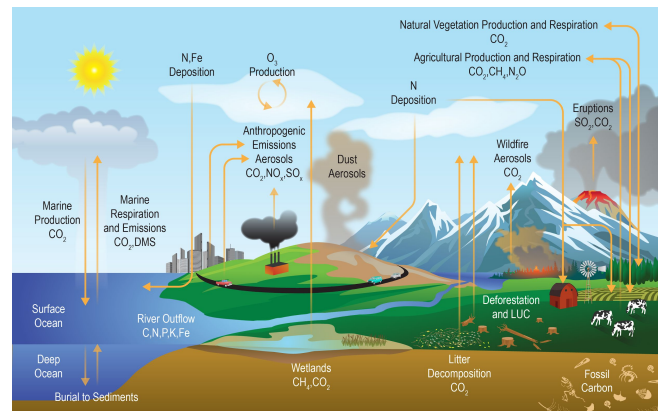
What is ILAMB?

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



Biogeochemical Cycles

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}) $\longrightarrow S_{\text{overall}} = \frac{S_{\text{bias}} + 2S_{\text{rmse}} + S_{\text{phase}} + S_{\text{iav}} + S_{\text{dist}}}{1 + 2 + 1 + 1 + 1}$
 - Graphical diagnostics**
 - Spatial contour maps
 - Time series line plots
 - Spatial Taylor diagrams (Taylor, 2001)
- Similar **tables** and **graphical diagnostics** for functional relationships

ILAMBv2.7 Package Current Variables

- **Biogeochemistry:** Biomass (Contiguous US, Pan Tropical Forest), Burned area (GFED3), CO₂ (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 & IOMB CMIP5 vs 6 Evaluation

Evaluation of CMIP5 vs CMIP6 with ILAMB and IOMB

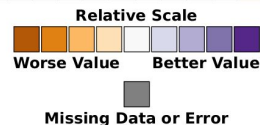
- (a) ILAMB and (b) IOMB have been used to evaluate how land and ocean model performance has changed from CMIP5 to CMIP6
- Model fidelity is assessed through comparison of historical simulations with a wide variety of contemporary observational datasets
- The UN's Intergovernmental Panel on Climate Change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

(a) Land Benchmarking Results

	bcc-csm1-1-1	CanESM2	CanESM1-BGC	GFGL-ESM2G	IPSL-CM5A-LR	MIROC-ESM	MPI-ESM-LR	NorESM1-ME	HadGEM2-ES	BCC-CSM2-MR	CanESM5	ESM2	GFGL-ESM4	IPSL-CM6A-LR	MIROC-ES2L	MPI-ESM1.2-LR	NorESM2-LM	UKESM1-0-LL	Mean CMIP5	Mean CMIP6
Land Ecosystem & Carbon Cycle																				
Biomass	-0.72	-0.93	-1.95	-1.51	-0.13	0.60	-0.43	-1.11	0.19	-0.43	0.66	0.48	-1.09	0.22	0.60	-0.07	1.00	0.49	1.63	2.30
Burned Area	0.20	-0.45	-1.52	-0.40	-1.26	-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	
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	-2.02	-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																				
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

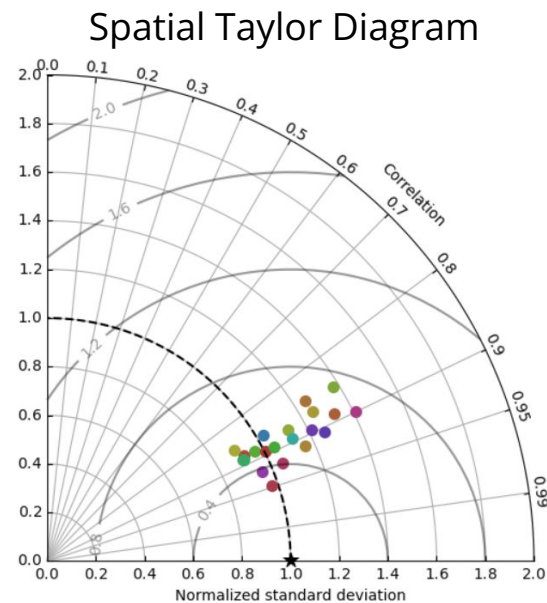
(b) Ocean Benchmarking Results

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	-3.50	2.19	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	1.98		-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	-2.21	-0.22		0.06	-0.36	0.85	-0.42	0.29	2.48	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



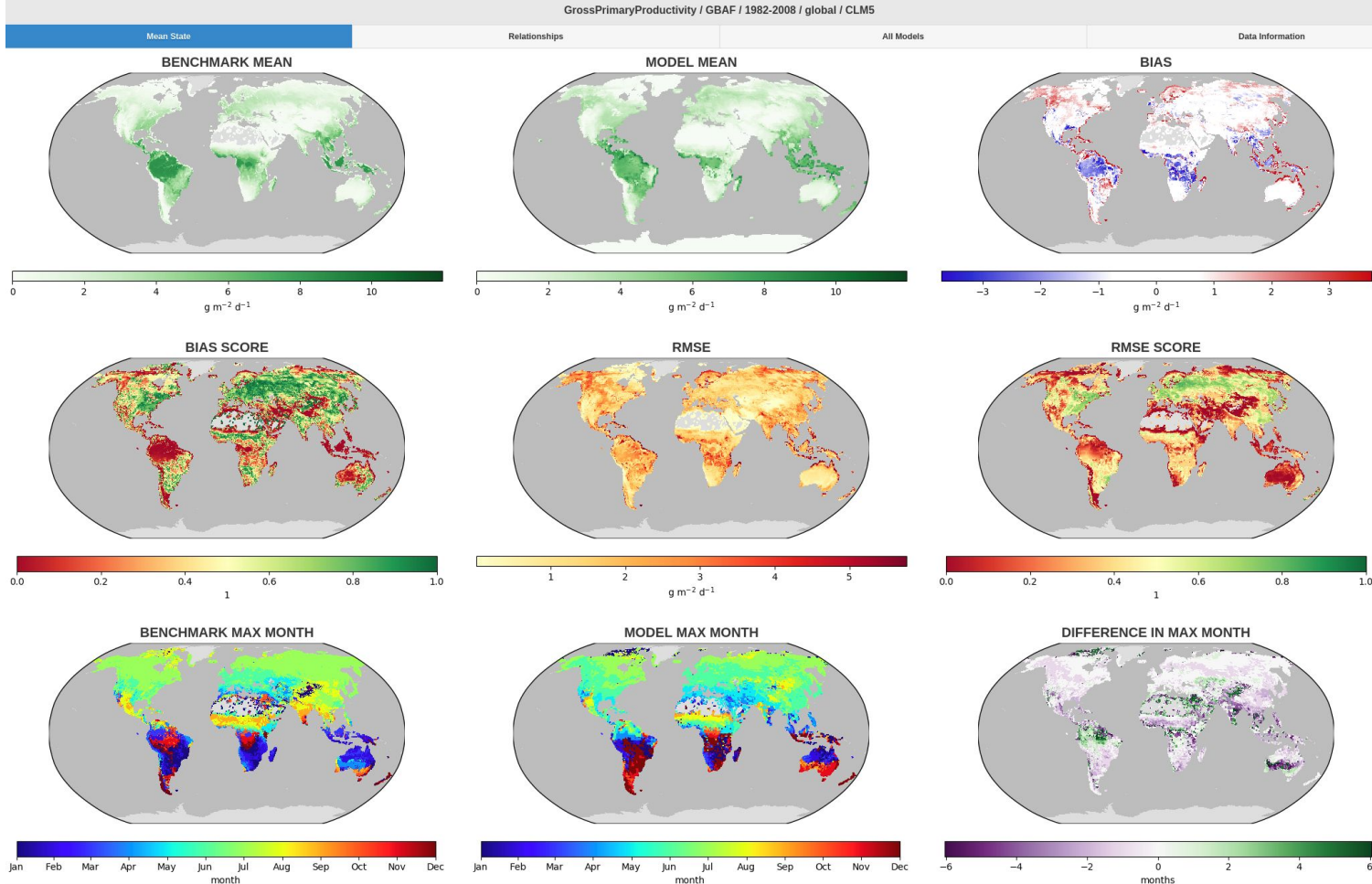
Gross Primary Productivity

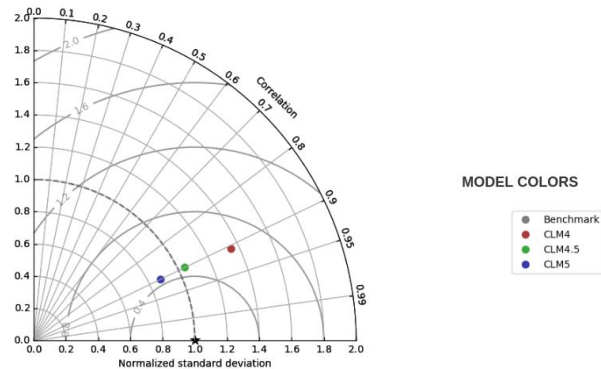
- Multimodel GPP is compared with global seasonal GBAF estimates
- We can see Improvements across generations of models (e.g., CESM1 vs. CESM2, IPSL-CM5A vs. 6A)
- The mean CMIP6 and CMIP5 models perform best



Benchmark	[1]	Download Data													
		Period Mean (original grids) [Pg yr ⁻¹]						Period Mean (intersection) [Pg yr ⁻¹]							
		Model Period Mean			Benchmark Period Mean (intersection)			Model Period Mean (complement)			Benchmark Period Mean (complement)				
		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]															
bcc-csm1-1	[1]	114.	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	[1]	114.	107.	113.	5.88	0.671	-0.0233	1.52	1.11	0.479	0.447	0.817	0.941	0.626	
CanESM2	[1]	129.	117.	114.	9.54		0.0601	2.31	2.00	0.388	0.437	0.850	0.838	0.549	
CanESM5	[1]	141.	128.	114.	10.1		0.730	1.87	1.60	0.449	0.418	0.710	0.948	0.589	
CESM1-BGC	[1]	129.	123.	113.	5.55	0.660	0.379	1.66	1.20	0.426	0.468	0.765	0.889	0.603	
CESM2	[1]	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	[1]	167.	152.	114.	12.4		1.26	2.78	1.38	0.377	0.288	0.735	0.897	0.517	
GFDL-ESM4	[1]	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	[1]	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	[1]	115.	109.	113.	5.27	0.708	0.111	1.39	1.14	0.547	0.477	0.790	0.961	0.650	
MeanCMIP5	[1]	121.	115.	114.	6.65		0.574	1.41	0.981	0.494	0.502	0.799	0.965	0.652	
MeanCMIP6	[1]	116.	110.	114.	6.26		0.129	1.17	0.931	0.572	0.522	0.826	0.956	0.576	
MIROC-ESM	[1]	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	[1]	116.	104.	113.	9.90	0.119	-0.0111	1.95	1.99	0.409	0.379	0.828	0.920	0.543	
MPI-ESM-LR	[1]	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	[1]	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	[1]	129.	120.	114.	7.82		0.386	1.86	1.25	0.387	0.456	0.761	0.856	0.583	
NorESM2-LM	[1]	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	[1]	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	[1]	126.	119.	113.	7.06	0.825	0.387	1.77	1.16	0.436	0.419	0.791	0.924	0.598	

ILAMB Graphical Diagnostics



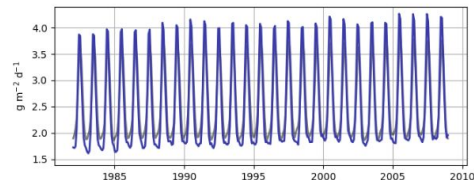


Spatially integrated regional mean

MODEL COLORS



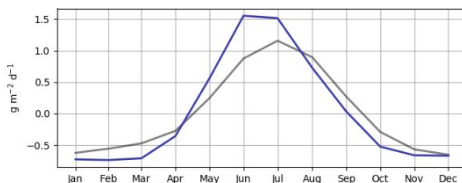
REGIONAL MEAN



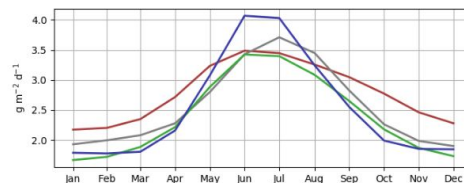
ANNUAL CYCLE

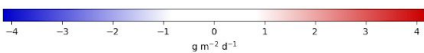
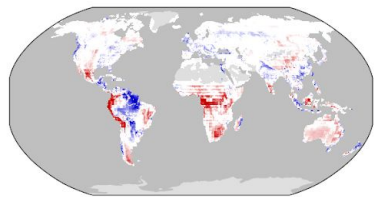


MONTHLY ANOMALY

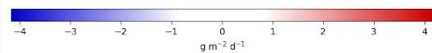
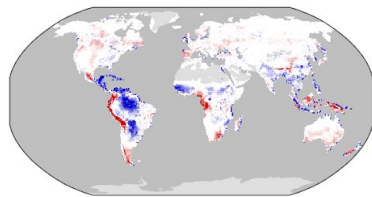


ANNUAL CYCLE

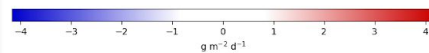
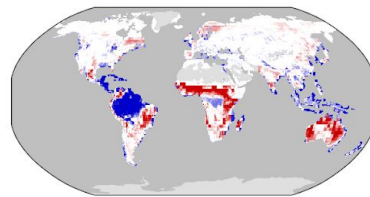


bcc-csm1-1

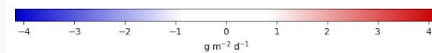
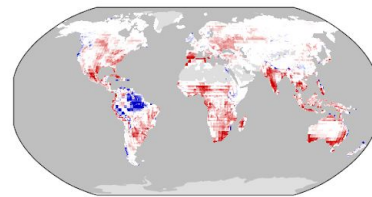
BCC-CSM2-MR



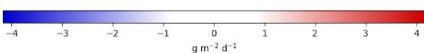
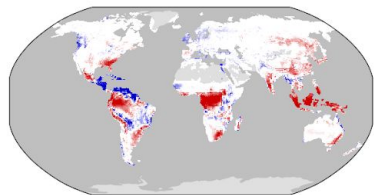
CanESM2



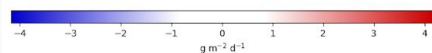
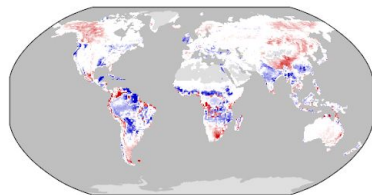
CanESM5



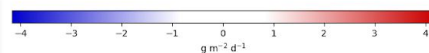
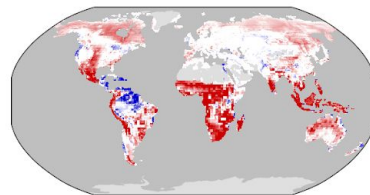
CESM1-BGC



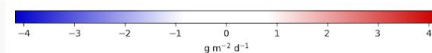
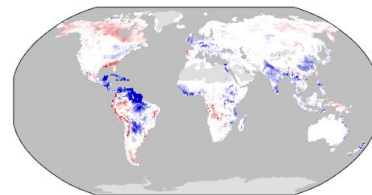
CESM2



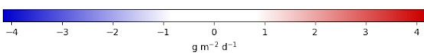
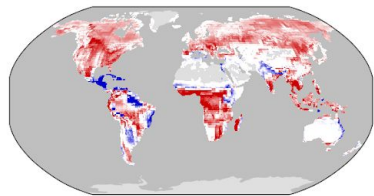
GFDL-ESM2G



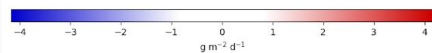
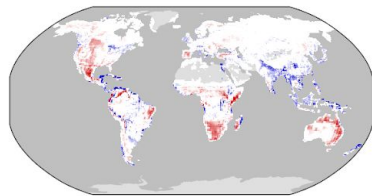
GFDL-ESM4



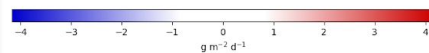
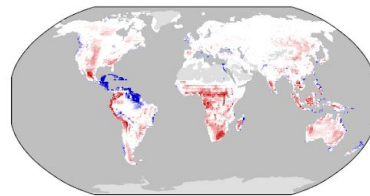
IPSL-CM5A-LR



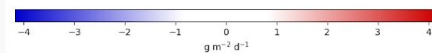
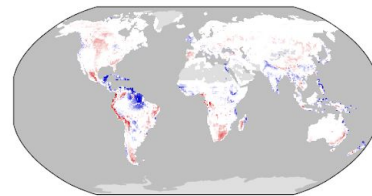
IPSL-CM6A-LR



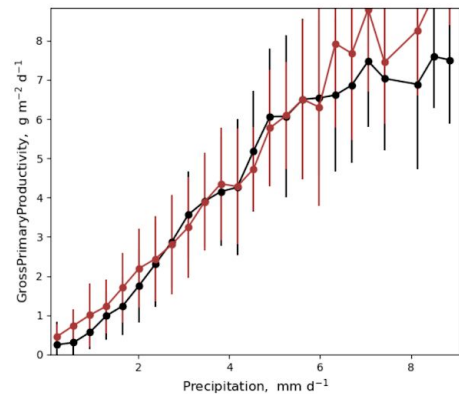
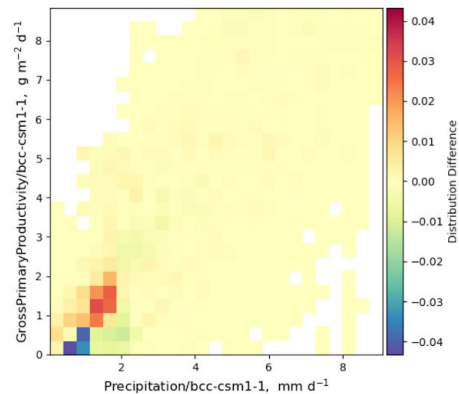
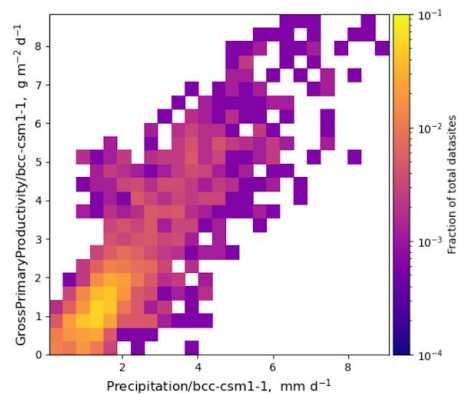
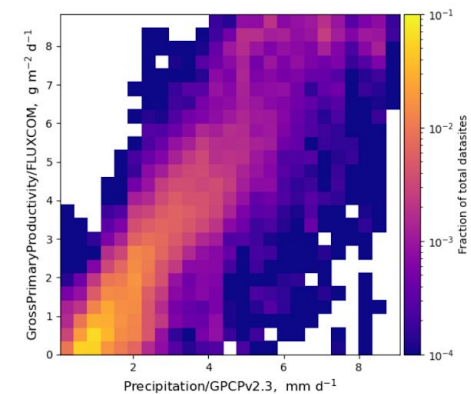
MeanCMIP5



MeanCMIP6



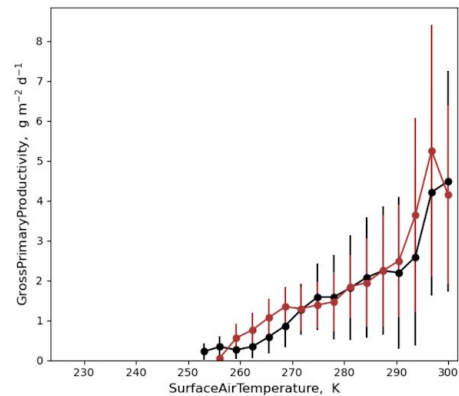
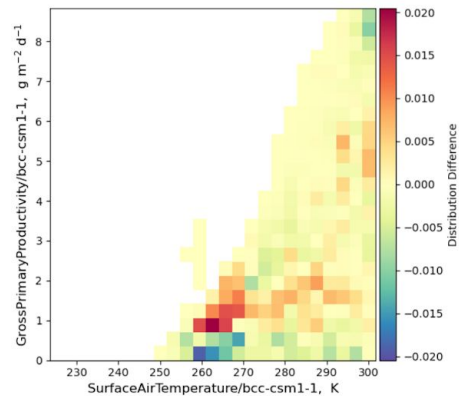
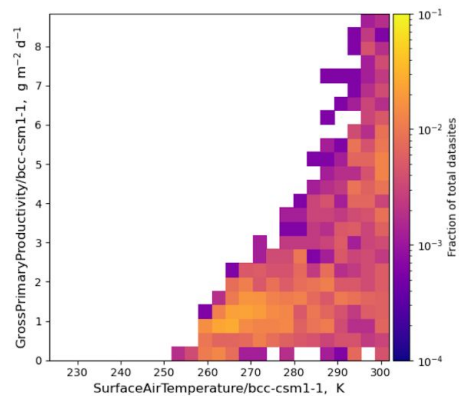
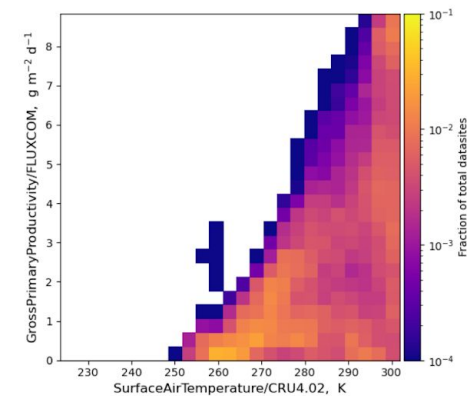
⊖ Precipitation/GPCv2.3



⊕ SurfaceDownwardSWRadiation/CERESed4.1

⊕ SurfaceNetSWRadiation/CERESed4.1

⊖ SurfaceAirTemperature/CRU4.02



ILAMB Data Collections

- ILAMB hosts thematic data collections to support model benchmarking that can be downloaded using “**ilamb-fetch**” tool.
- Currently three collections are supported:
 - ILAMB-Data, ABoVE-Data, NGEEA-Data

```
ilamb-fetch --local_root=ILAMB/DATA --collection=ABoVE-Data --no-check-certificate
```

- Data inventory: <https://ilamb.org/datasets.html>