International Land Model Benchmarking (ILAMB) Project

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

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

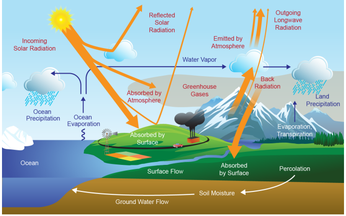
US Global Change Research Program (USGCRP)
Interagency Working Group on Integrated Observations (ObsIWG)

April 9, 2019

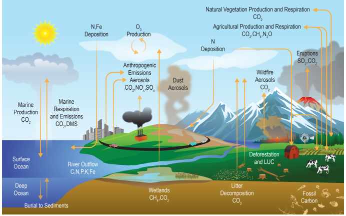
What is ILAMB?

A community coordination activity created to:

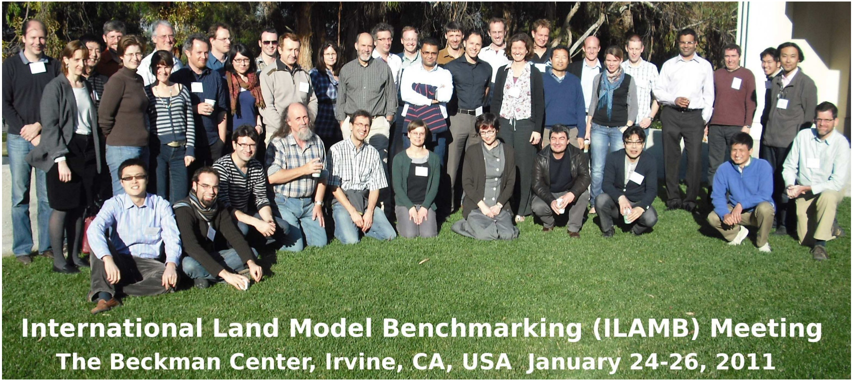
- **Develop internationally accepted benchmarks** for land model performance by drawing upon collaborative expertise
- **Promote the use of these benchmarks** for model intercomparison
- **Strengthen linkages between experimental, remote sensing, and climate modeling communities** in the design of new model tests and new measurement programs
- **Support the design and development of open source benchmarking tools** (Luo et al., 2012)



Energy and Water Cycles



Carbon and Biogeochemical Cycles

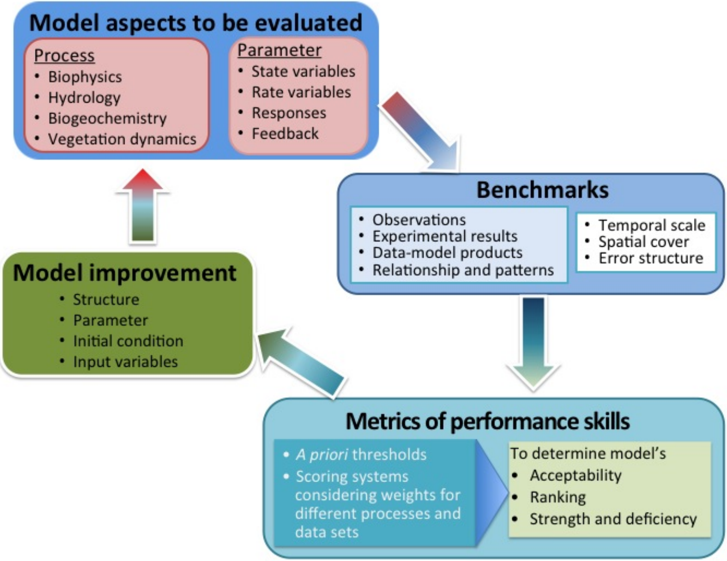


International Land Model Benchmarking (ILAMB) Meeting

The Beckman Center, Irvine, CA, USA January 24-26, 2011

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

General Benchmarking Procedure



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

Metrics 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

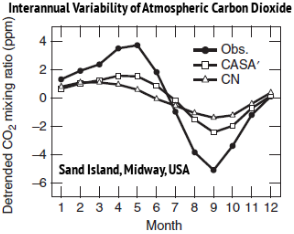
Model improvement

- Structure
- Parameter
- Initial condition
- Input variables

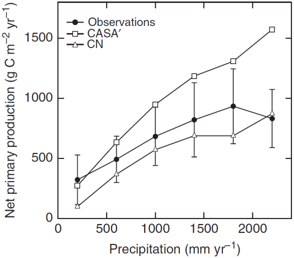
(Luo et al., 2012)

What is a Benchmark?

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



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



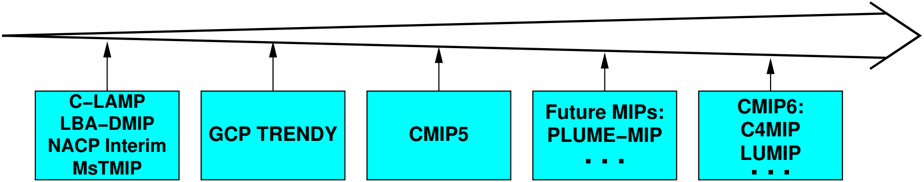
Models may reproduce correct responses over only a limited range of forcing variables.

(Randerson et al., 2009)

Why Benchmark?

- **To demonstrate model improvements** in representation of coupled climate and biogeochemical cycles
- **To quantitatively diagnose impacts of model development** in related fields on carbon cycle processes
- **To guide synthesis efforts**, such as the Intergovernmental Panel on Climate Change (IPCC), in assessing model fidelity
- **To increase scrutiny of key datasets** used for model evaluation
- **To identify gaps in existing observations** needed for model validation
- **To accelerate incorporation of new measurements** for rapid and widespread use in model assessment
- **To offer a quantitative, application-specific set of criteria** for participation in model intercomparison projects (MIPs)
- **To inform a model weighting system** for multi-model estimates of future changes in the carbon cycle

An Open Source Benchmarking Software System



C–LAMP
LBA–DMIP
NACP Interim
MsTMIP

GCP TRENDY

CMIP5

Future MIPs:
PLUME–MIP
. . .

CMIP6:
C4MIP
LUMIP
. . .

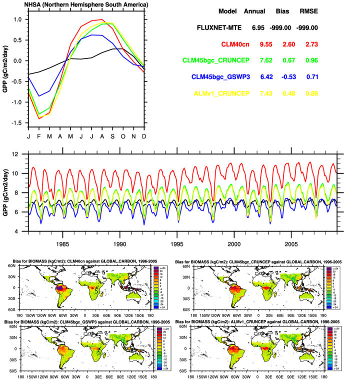
- Human capital costs of making rigorous model–data comparisons is considerable and constrains the scope of individual MIPs.
- Many MIPs spend resources "reinventing the wheel" in terms of variable naming conventions, model simulation protocols, and analysis software.
- **Need for ILAMB:** Each new MIP has access to the model–data comparison modules from past MIPs through ILAMB (*e.g.*, MIPs use one common modular software system). Standardized international naming conventions also increases MIP efficiency.

What is ILAMB Now?

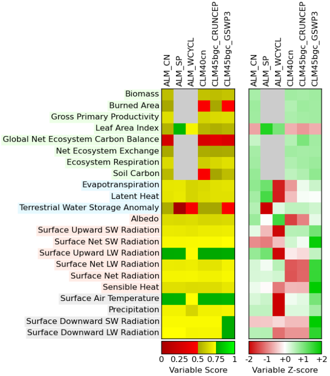
- **Community:** global group of modelers and scientists enthusiastic about benchmarking
- **Datasets:** curated collection of datasets formatted for easy comparison
- **Methods:** innovative assembly of techniques for benchmarking models
- **Software:** open-source python package which you can use or tailor
- **Results:** catalog of comparisons which you can access and peruse

Current Status of the ILAMB Packages

- **ILAMBv1** released at 2015 AGU Town Hall, doi:10.18139/ILAMB.v001.00/1251597
- **ILAMBv2** released at 2016 ILAMB Workshop, doi:10.18139/ILAMB.v002.00/1251621
- Used routinely for E3SM and CESM evaluation during development



Model	Annual	Bias	RMSE
FLUXNET-MTE	6.95	-999.00	-999.00
CLM40cn	9.55	2.60	2.73
CLM45bgc_CRUNCEP	7.62	0.67	0.96
CLM45bgc_GSWP3	6.42	-0.53	0.71
ALMv1_CRUNCEP	7.43	0.48	0.89



ILAMBv2 Diagnostics Package

- Current variables:

  Aboveground live biomass (Contiguous US, Pan Tropical Forest), Burned area (GFED3), CO_2 (NOAA GMD, Mauna Loa), Gross primary production (Fluxnet, MTE), Leaf area index (AVHRR, MODIS), Global net land flux (GCP, Khatiwala/Hoffman), Net ecosystem exchange (Fluxnet, GBA), Ecosystem Respiration (Fluxnet, GBA), Soil C (HWSD, NCSCDv2), Evapotranspiration (GLEAM, MODIS), Latent heat (Fluxnet, MTE), Soil moisture (ESA), Terrestrial water storage anomaly (GRACE), Albedo (CERES, GEWEX, MODIS), Surface up SW/LW radiation (CERES, GEWEX.SRB, WRMC.BSRN), Sensible heat (Fluxnet, GBA), Surface air temperature (CRU, Fluxnet), Precipitation (Fluxnet, GPCC, GPCP2), Surface down SW/LW radiation (Fluxnet, CERES, GEWEX.SRB, WRMC.BSRN),

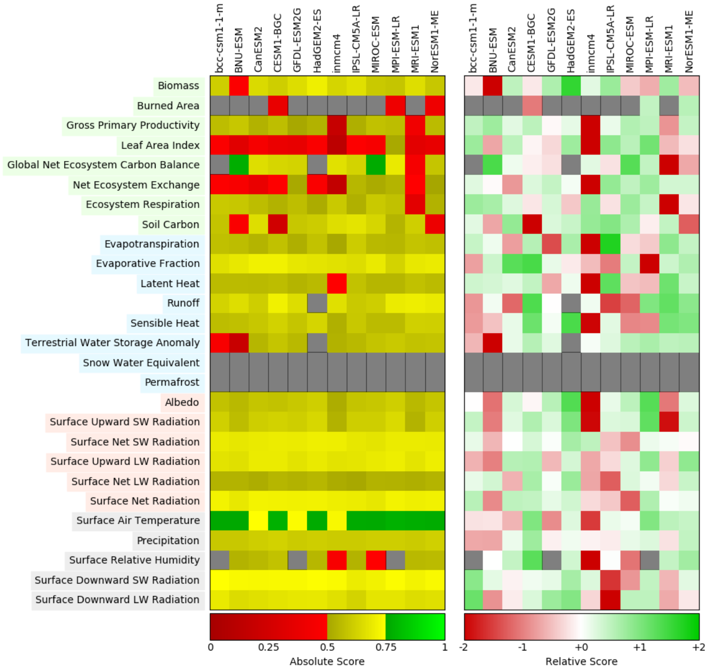
- Graphics and scoring systems:

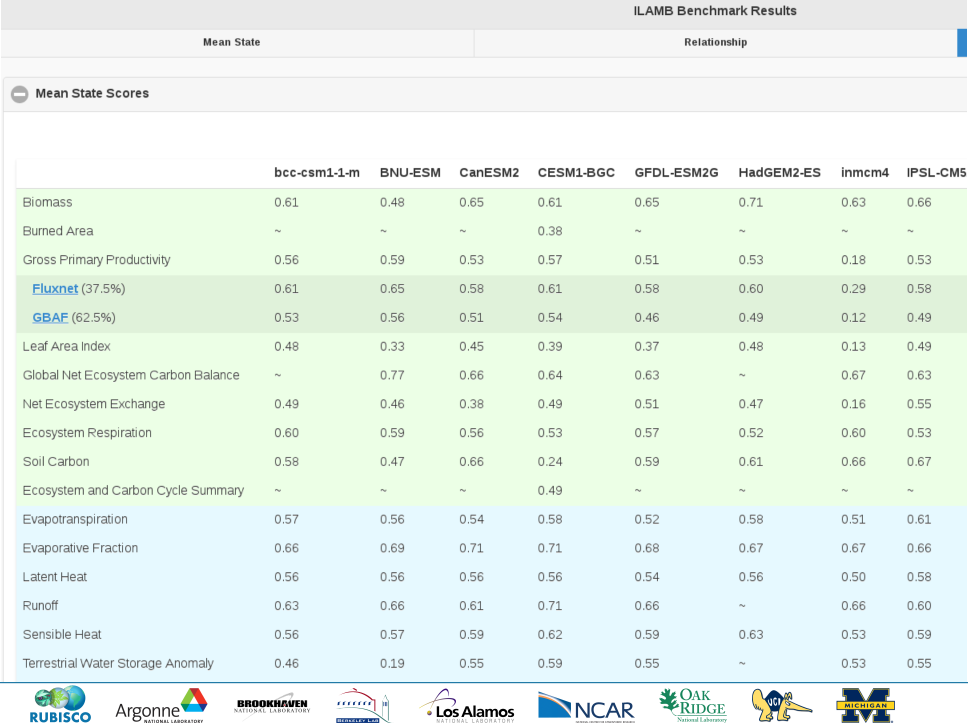
  • Annual mean, Bias, RMSE, seasonal cycle, spatial distribution, interannual coeff. of variation and variability, long-term trend scores

  • Global maps, variable to variable, and time series comparisons

- Software:

  Freely distributed, designed to be user friendly and to enable easy addition of new variables

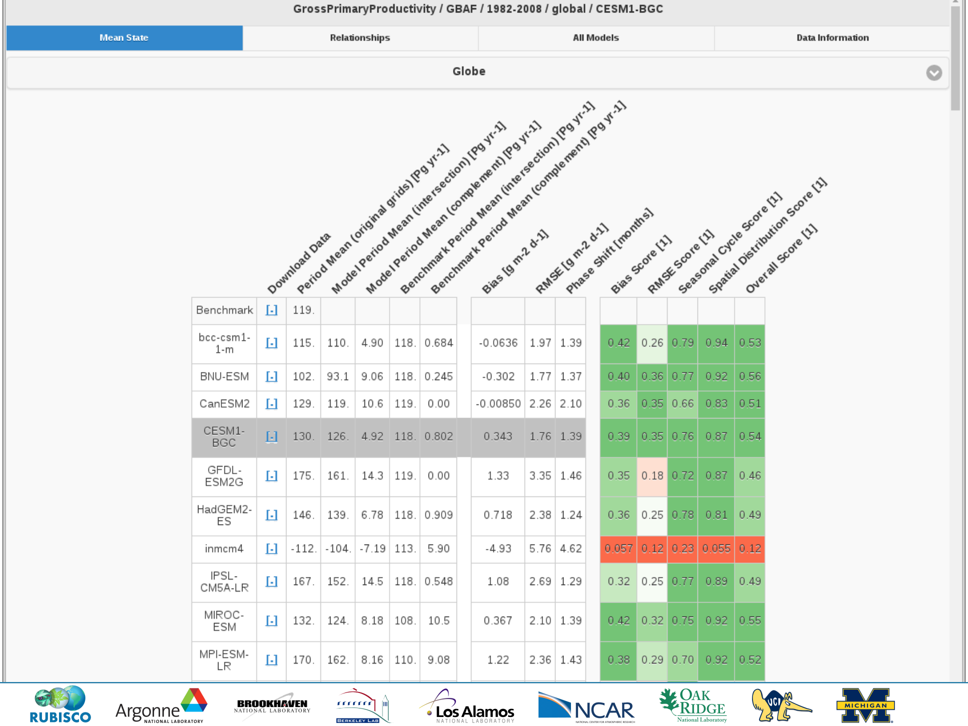

bcc-csm1-1-m
BNU-ESM
CanESM2
CESM1-BGC
GFDL-ESM2G
HadGEM2-ES
inmcm4
IPSL-CM5A-LR
MIROC-ESM
MPI-ESM-LR
MRI-ESM1
NorESM1-ME
Biomass
Burned Area
Gross Primary Productivity
Leaf Area Index
Global Net Ecosystem Carbon Balance
Net Ecosystem Exchange
Ecosystem Respiration
Soil Carbon
Evapotranspiration
Evaporative Fraction
Latent Heat
Runoff
Sensible Heat
Terrestrial Water Storage Anomaly
Snow Water Equivalent
Permafrost
Albedo
Surface Upward SW Radiation
Surface Net SW Radiation
Surface Upward LW Radiation
Surface Net LW Radiation
Surface Net Radiation
Surface Air Temperature
Precipitation
Surface Relative Humidity
Surface Downward SW Radiation
Surface Downward LW Radiation
0
0.25
0.5
0.75
1
Absolute Score
-2
-1
+0
+1
+2
Relative Score


Mean State
Relationship
RUBISCO
Argonne NATIONAL LABORATORY
BROOKHAVEN NATIONAL LABORATORY
BERKELEY LAB
Los Alamos NATIONAL LABORATORY
NCAR
OAK RIDGE National Laboratory
UCI
MICHIGAN

ILAMB Benchmark Results

Mean State Scores

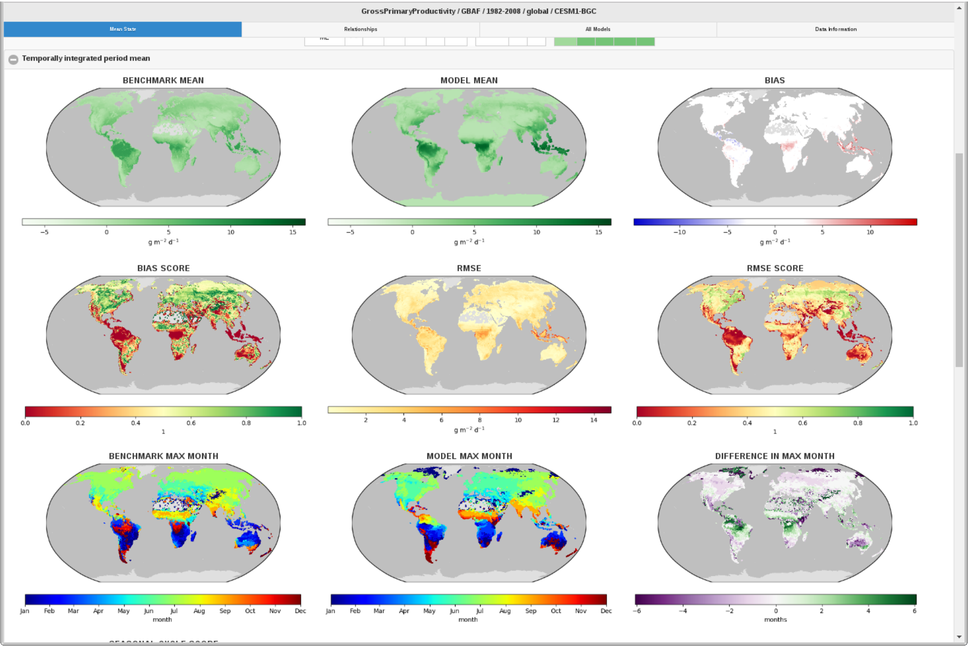
	bcc-csm1-1-m	BNU-ESM	CanESM2	CESM1-BGC	GFDL-ESM2G	HadGEM2-ES	inmcm4	IPSL-CM5
Biomass	0.61	0.48	0.65	0.61	0.65	0.71	0.63	0.66
Burned Area	~	~	~	0.38	~	~	~	~
Gross Primary Productivity	0.56	0.59	0.53	0.57	0.51	0.53	0.18	0.53
Fluxnet (37.5%)	0.61	0.65	0.58	0.61	0.58	0.60	0.29	0.58
GBAF (62.5%)	0.53	0.56	0.51	0.54	0.46	0.49	0.12	0.49
Leaf Area Index	0.48	0.33	0.45	0.39	0.37	0.48	0.13	0.49
Global Net Ecosystem Carbon Balance	~	0.77	0.66	0.64	0.63	~	0.67	0.63
Net Ecosystem Exchange	0.49	0.46	0.38	0.49	0.51	0.47	0.16	0.55
Ecosystem Respiration	0.60	0.59	0.56	0.53	0.57	0.52	0.60	0.53
Soil Carbon	0.58	0.47	0.66	0.24	0.59	0.61	0.66	0.67
Ecosystem and Carbon Cycle Summary	~	~	~	0.49	~	~	~	~
Evapotranspiration	0.57	0.56	0.54	0.58	0.52	0.58	0.51	0.61
Evaporative Fraction	0.66	0.69	0.71	0.71	0.68	0.67	0.67	0.66
Latent Heat	0.56	0.56	0.56	0.56	0.54	0.56	0.50	0.58
Runoff	0.63	0.66	0.61	0.71	0.66	~	0.66	0.60
Sensible Heat	0.56	0.57	0.59	0.62	0.59	0.63	0.53	0.59
Terrestrial Water Storage Anomaly	0.46	0.19	0.55	0.59	0.55	~	0.53	0.55



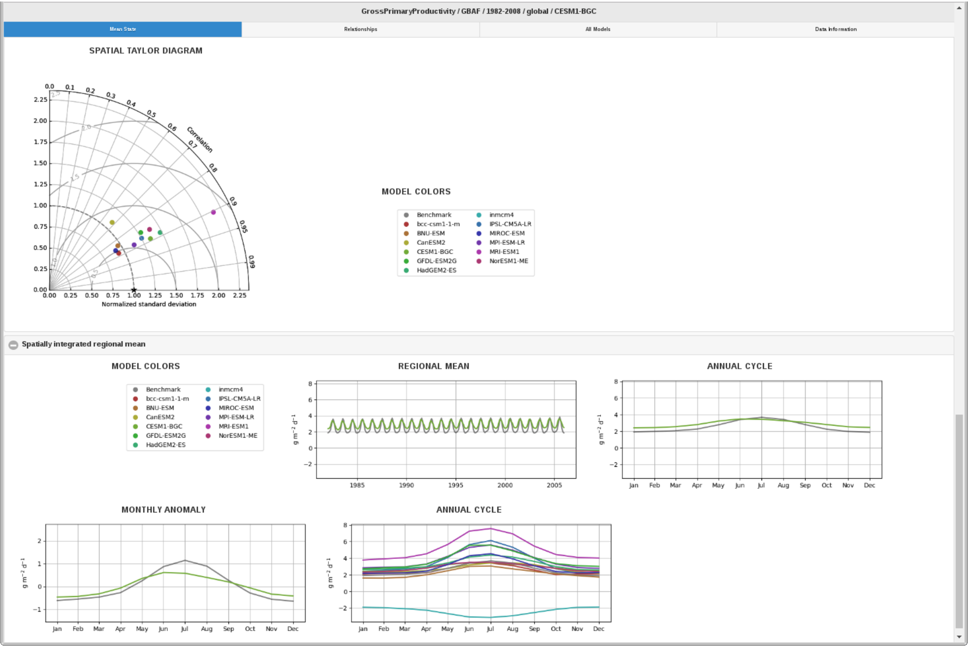
GrossPrimaryProductivity / GBAF / 1982-2008 / global / CESM1-BGC

Globe

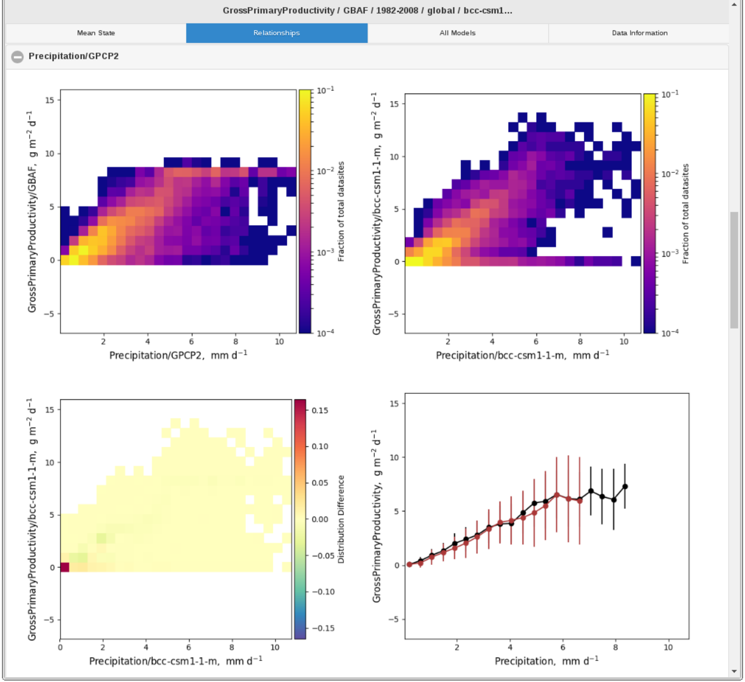
	Download Data	Period Mean (original grids) [Pg yr-1]	Model Period Mean (intersection) [Pg yr-1]	Model Period Mean (complement) [Pg yr-1]	Benchmark Period Mean (intersection) [Pg yr-1]	Benchmark Period Mean (complement) [Pg yr-1]	Bias [g m-2 d-1]	RMSE [g m-2 d-1]	Phase Shift [months]	Bias Score [1]	RMSE Score [1]	Seasonal Cycle Score [1]	Spatial Distribution Score [1]	Overall Score [1]
Benchmark	[-]	119.												
bcc-csm1-1-m	[-]	115.	110.	4.90	118.	0.684	-0.0636	1.97	1.39	0.42	0.26	0.79	0.94	0.53
BNU-ESM	[-]	102.	93.1	9.06	118.	0.245	-0.302	1.77	1.37	0.40	0.36	0.77	0.92	0.56
CanESM2	[-]	129.	119.	10.6	119.	0.00	-0.00850	2.26	2.10	0.36	0.35	0.66	0.83	0.51
CESM1-BGC	[-]	130.	126.	4.92	118.	0.802	0.343	1.76	1.39	0.39	0.35	0.76	0.87	0.54
GFDL-ESM2G	[-]	175.	161.	14.3	119.	0.00	1.33	3.35	1.46	0.35	0.18	0.72	0.87	0.46
HadGEM2-ES	[-]	146.	139.	6.78	118.	0.909	0.718	2.38	1.24	0.36	0.25	0.78	0.81	0.49
inmcm4	[-]	-112.	-104.	-7.19	113.	5.90	-4.93	5.76	4.62	0.057	0.12	0.23	0.055	0.12
IPSL-CM5A-LR	[-]	167.	152.	14.5	118.	0.548	1.08	2.69	1.29	0.32	0.25	0.77	0.89	0.49
MIROC-ESM	[-]	132.	124.	8.18	108.	10.5	0.367	2.10	1.39	0.42	0.32	0.75	0.92	0.55
MPI-ESM-LR	[-]	170.	162.	8.16	110.	9.08	1.22	2.36	1.43	0.38	0.29	0.70	0.92	0.52



RUBISCO Argonne NATIONAL LABORATORY BROOKHAVEN NATIONAL LABORATORY BERKELEY LAB Los Alamos NATIONAL LABORATORY NCAR OAK RIDGE National Laboratory UCI MICHIGAN



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



GrossPrimaryProductivity / GBAF / 1982-2008 / global / bcc-csm1...

Mean State
Relationships
All Models
Data Information
Precipitation/GPCP2
GrossPrimaryProductivity/GBAF, g m^{-2} d^{-1}
Precipitation/GPCP2, mm d^{-1}
Fraction of total datasites
10^{-1}
10^{-2}
10^{-3}
10^{-4}
GrossPrimaryProductivity/bcc-csm1-1-m, g m^{-2} d^{-1}
Precipitation/bcc-csm1-1-m, mm d^{-1}
Distribution Difference
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
GrossPrimaryProductivity, g m^{-2} d^{-1}
Precipitation, mm d^{-1}
15
10
5
0
−5
0
2
4
6
8
10

Dataset Weighting Rubric

Score	Certainty of data	Scale appropriateness and coverage	Overall important of constraint or process
1	No uncertainty, significant methodological issues affecting quality	Site level observations with limited space/time coverage	Observations that have limited influence on the targeted Earth system dynamics
2	No uncertainty, some methodological issues affecting quality	Partial regional coverage, up to 1 year	Observations have direct influence on the targeted Earth system dynamics
3	No uncertainty, methodology has some peer review	Regional coverage, at least 1 year	Observations useful to constrain processes that contribute to the targeted Earth system dynamics
4	Qualitative uncertainty, methodology accepted	Important regional coverage, at least 1 year	Observations well-suited to constrain important processes
5	Well-defined and relatively low uncertainty	Global scale spanning multiple years	Observations well-suited for discriminating critical processes among models

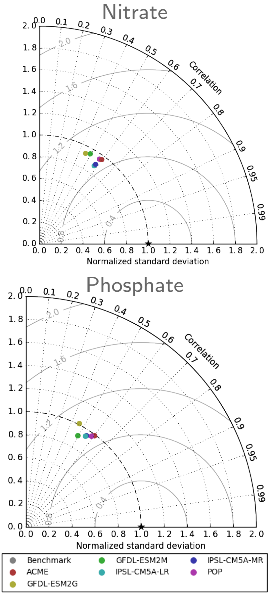
What is in the Overall Score?

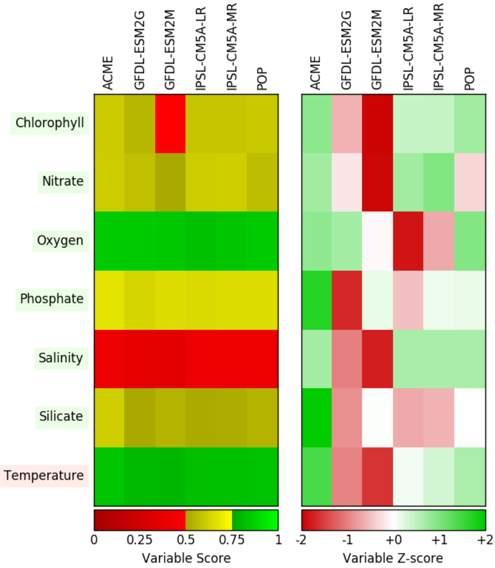
$$S_{overall} = \frac{S_{bias} + 2S_{rmse} + S_{phase} + S_{iav} + S_{dist}}{1+2+1+1+1}$$

Scores are based on the:

- S_{bias} - normalized bias
- S_{rmse} - normalized central RMSE
- S_{phase} - timing of the maximum of the annual cycle
- S_{iav} - interannual variability
- S_{dist} - spatial distribution of the period mean

Extending ILAMB for Ocean Model Evaluation





3rd Workshop, Washington, DC, USA, May 16–18, 2016

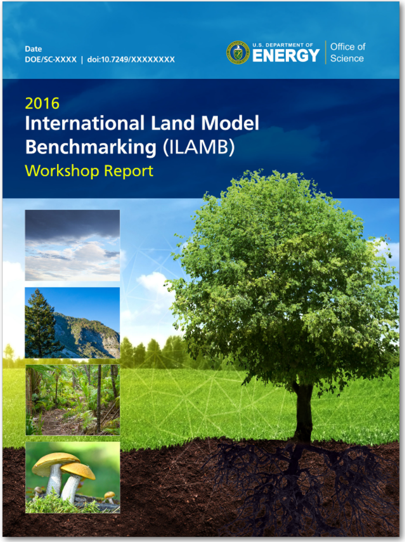
Overarching Workshop Goals

Engage the research community in defining scientific priorities for

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

Workshop Attendance

- 60+ participants from Australia, Japan, China, Germany, Sweden, Netherlands, UK, and US
- 10 modeling centers represented
- ~25 online attendees at any time



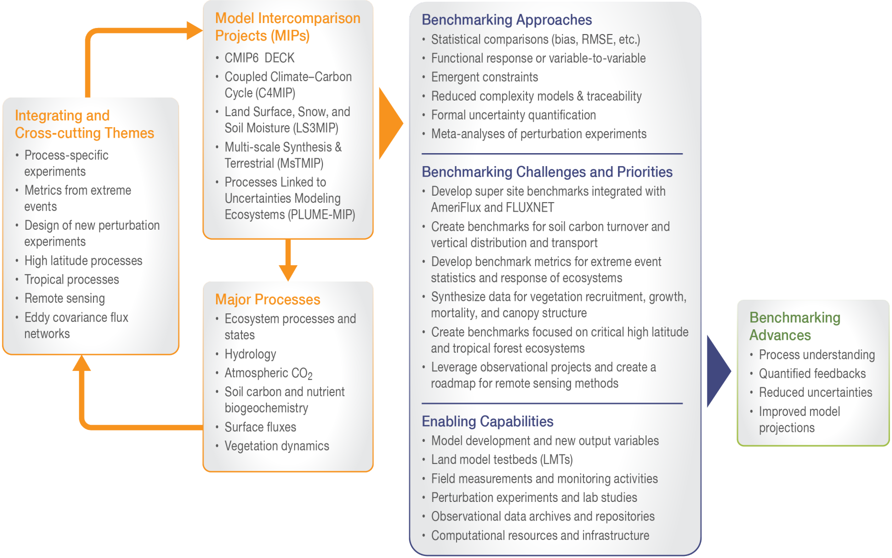
Date
DOE/SC-XXXX | doi:10.7249/XXXXXXXX

U.S. DEPARTMENT OF ENERGY | Office of Science

2016
International Land Model Benchmarking (ILAMB)
Workshop Report

doi:10.2172/1330803

2016 ILAMB Workshop Synthesis



Integrating and Cross-cutting Themes

- Process-specific experiments
- Metrics from extreme events
- Design of new perturbation experiments
- High latitude processes
- Tropical processes
- Remote sensing
- Eddy covariance flux networks

Model Intercomparison Projects (MIPs)

- CMIP6 DECK
- Coupled Climate–Carbon Cycle (C4MIP)
- Land Surface, Snow, and Soil Moisture (LS3MIP)
- Multi-scale Synthesis & Terrestrial (MsTMIP)
- Processes Linked to Uncertainties Modeling Ecosystems (PLUME-MIP)

Major Processes

- Ecosystem processes and states
- Hydrology
- Atmospheric CO_2
- Soil carbon and nutrient biogeochemistry
- Surface fluxes
- Vegetation dynamics

Benchmarking Approaches

- Statistical comparisons (bias, RMSE, etc.)
- Functional response or variable-to-variable
- Emergent constraints
- Reduced complexity models & traceability
- Formal uncertainty quantification
- Meta-analyses of perturbation experiments

Benchmarking Challenges and Priorities

- Develop super site benchmarks integrated with AmeriFlux and FLUXNET
- Create benchmarks for soil carbon turnover and vertical distribution and transport
- Develop benchmark metrics for extreme event statistics and response of ecosystems
- Synthesize data for vegetation recruitment, growth, mortality, and canopy structure
- Create benchmarks focused on critical high latitude and tropical forest ecosystems
- Leverage observational projects and create a roadmap for remote sensing methods

Enabling Capabilities

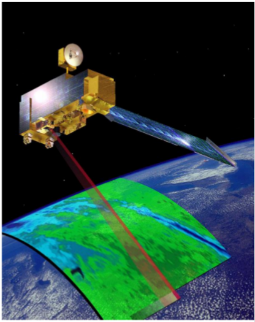
- Model development and new output variables
- Land model testbeds (LMTs)
- Field measurements and monitoring activities
- Perturbation experiments and lab studies
- Observational data archives and repositories
- Computational resources and infrastructure

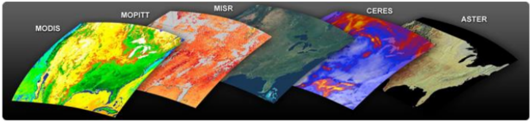
Benchmarking Advances

- Process understanding
- Quantified feedbacks
- Reduced uncertainties
- Improved model projections

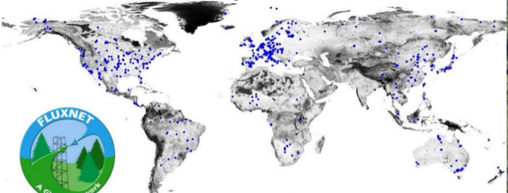
Benchmarking Challenges and Priorities

- **Super site benchmarks** for AmeriFlux and FLUXNET
- **Benchmarks for soil carbon** turnover, distribution, transport
- **Metrics for extreme events** & response of ecosystems
- **Data for vegetation** recruitment, growth, mortality, phenology, canopy structure
- Benchmarks for critical **high latitude & tropical ecosystems**
- Leverage **field projects & remote sensing methods**





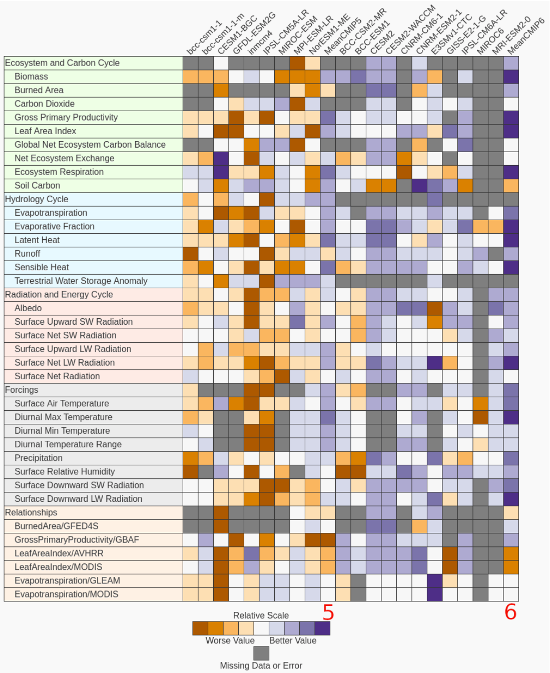
FLUXNET Representativeness



Kumar, Hoffman, and Hargrove (in prep)

CMIP5 vs. CMIP6

- The CMIP6 suite of land models (right) has improved over the CMIP5 suite of land models.
- 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.



Future ILAMB Development and Application

- ILAMBv1 and ILAMBv2 were applied to:
  - CMIP5 Historical and esmHistorical simulations
  - Model development of the Community Land Model (CLM)
  - E3SM Land Model (ELM) evaluation of BGC approaches
- Within U.S. Department of Energy projects:
  - NGEE Arctic, NGEE Tropics, and SPRUCE are adopting the framework for evaluating process parameterizations & integrating field observations
  - E3SM uses it for evaluation of new land model features
  - RUBISCO is developing the framework and benchmarking MIPs
- Ongoing projects: TRENDY, MsTMIP, CMIP6
- Others are using and contributing to ILAMB:
  - NASA-funded Permafrost Benchmarking System
  - In-house model evaluation at various modeling centers

Collier, Nathan, Forrest M. Hoffman, David M. Lawrence, Gretchen Keppel-Aleks, Charles D. Koven, William J. Riley, Mingquan Mu, and James 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.

Important Links

- Open source git repository

  https://bitbucket.org/ncollier/ilamb

- CLM (4/4.5/5)

  https://www.ilamb.org/CLM/

- CMIP5

  https://www.ilamb.org/CMIP5/esmHistorical/

- IOMB (Ocean benchmarking)

  https://www.ilamb.org/IOMB/

Acknowledgements

This research was supported through the Reducing Uncertainties in Biogeochemical Interactions through Synthesis and Computation Scientific Focus Area (RUBISCO SFA), which is sponsored by the Regional and Global Model Analysis (RGMA) Program in the Climate and Environmental Sciences Division (CESD) of the Office of Biological and Environmental Research (BER) in the U. S. Department of Energy Office of Science. Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC, for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725. The National Center for Atmospheric Research (NCAR) is managed by the University Corporation for Atmospheric Research (UCAR) on behalf of the National Science Foundation (NSF). Lawrence Berkeley National Laboratory (LBNL) is managed and operated by the Regents of the University of California under Contract No. DE-AC02-05CH11231.

References

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

F. M. Hoffman, 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. 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. International Land Model Benchmarking (ILAMB) 2016 workshop report. Technical Report DOE/SC-0186, U.S. Department of Energy, Office of Science, Germantown, Maryland, USA, Apr. 2017.

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

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