The International Land Model Benchmarking (ILAMB) Project

Forrest M. Hoffman[1, 2], James T. Randerson[1], David M. Lawrence[3], Eleanor M. Blyth[4], Mingquan Mu[1], Gretchen Keppel-Aleks[1], Kathe Todd-Brown[1], Brendan M. Rogers[1], Miguel D. Mahecha[5], Nuno Carvalhais[5], Jannis von Buttlar[5], Markus Reichstein[5], Martin Best[6], James Ehleringer[7], Yiqi Luo[8], and others

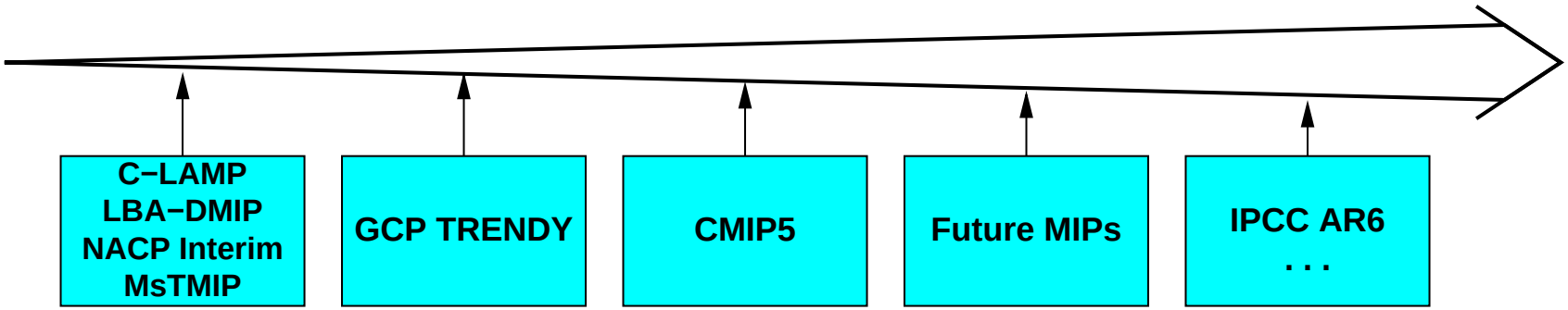
[1]University of California-Irvine, [2]Oak Ridge National Laboratory, [3]National Center for Atmospheric Research, [4]Centre for Ecology & Hydrology, [5]Max Planck Institute for Biogeochemistry, [6]United Kingdom Met Office, [7]University of Utah, and [8]University of Oklahoma

January 23, 2013
DataONE Exploration, Visualization, and Analysis (EVA) Workshop
Polytechnic Institute of New York University, Brooklyn, New York, USA

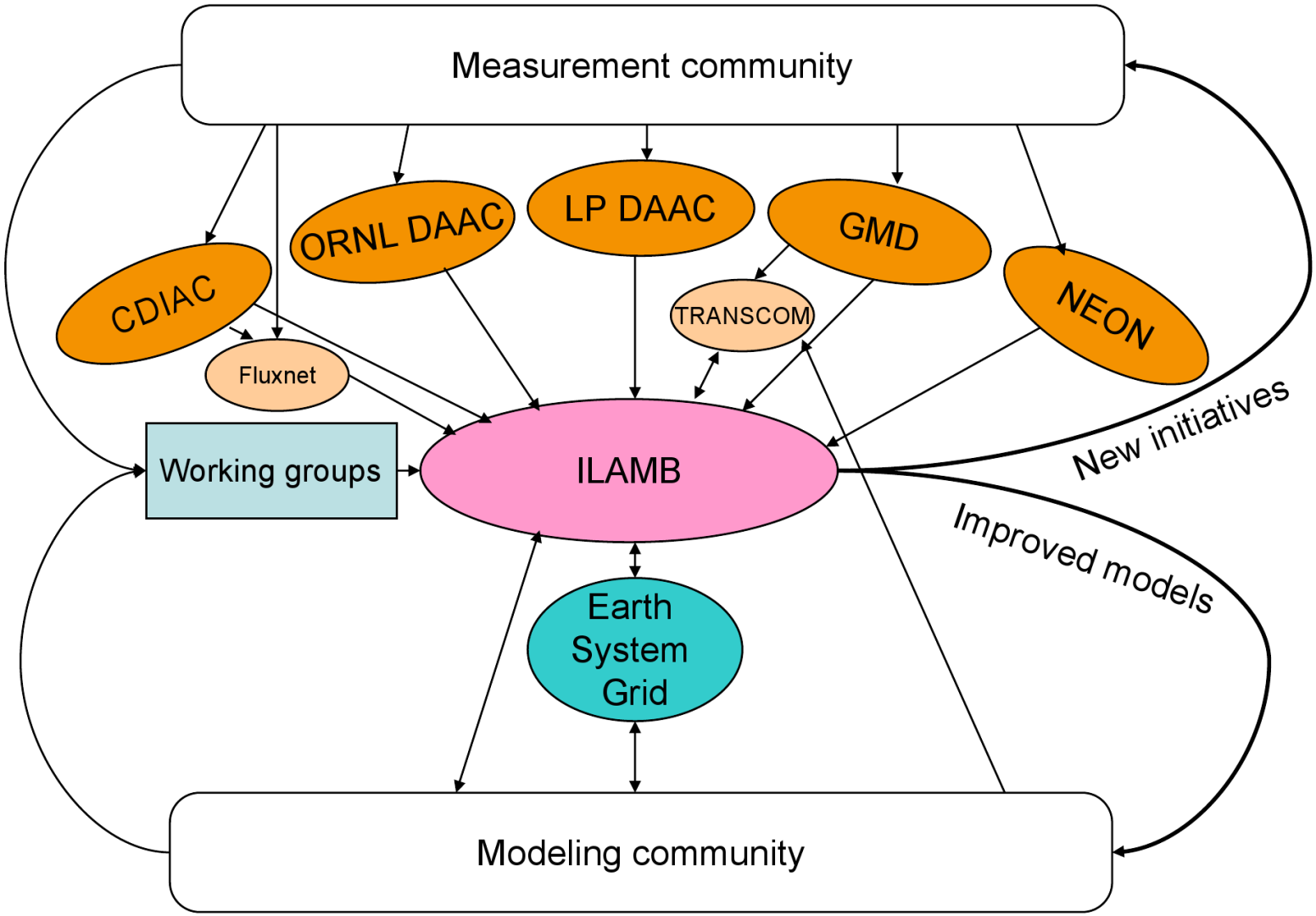
Why Benchmark?

- to show the broader science community and the public that the representation of the carbon cycle in climate models is improving;
- to provide a means, in Earth System models, to quantitatively diagnose impacts of model development in related fields on carbon cycle and land surface processes;
- to guide synthesis efforts, such as the Intergovernmental Panel on Climate Change (IPCC), in the review of mechanisms of global change in models that are broadly consistent with available contemporary observations;
- to increase scrutiny of key datasets used for model evaluation;
- to identify gaps in existing observations needed for model validation;
- to provide a quantitative, application-specific set of minimum criteria for participation in model intercomparison projects (MIPs);
- to provide an optional weighting system for multi-model mean estimates of future changes in the carbon cycle.

An Open Source Benchmarking Software System



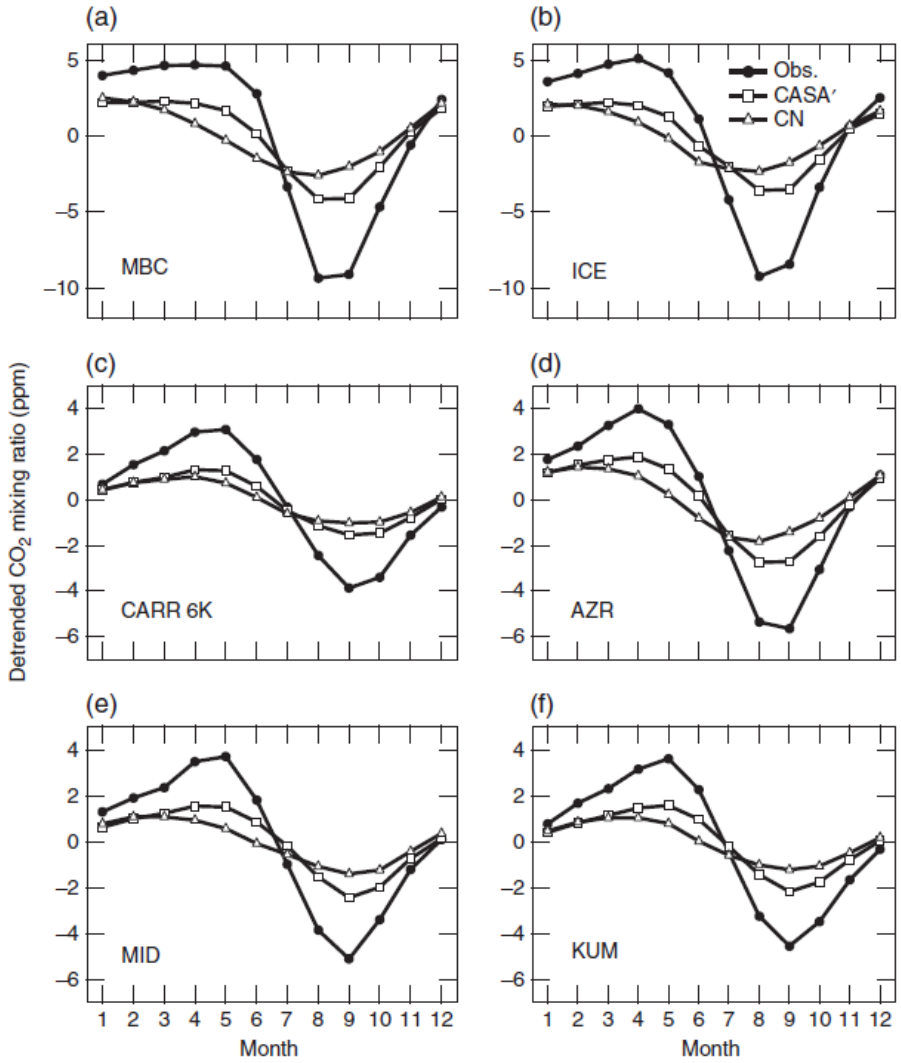
- Human capital costs of making rigorous model-data comparisons is considerable and constrains the scope of individual MIPs.
- Many MIPs spend resources "reinventing the wheel" in terms of variable naming conventions, model simulation protocols, and analysis software.
- **Need for ILAMB:** Each new MIP 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 increase MIP efficiency.



International Land Model Benchmarking project and diagnostic system

What is a Benchmark?

- A benchmark is a quantitative test of model function, for which the uncertainties associated with the observations can be quantified.
- Acceptable performance on benchmarks **is a necessary but not sufficient condition** for a fully functioning model.
- Since all datasets have strengths and weaknesses, an effective benchmark is one that draws upon a broad set of independent observations to evaluate model performance on multiple temporal and spatial scales.



(Randerson et al., 2009)

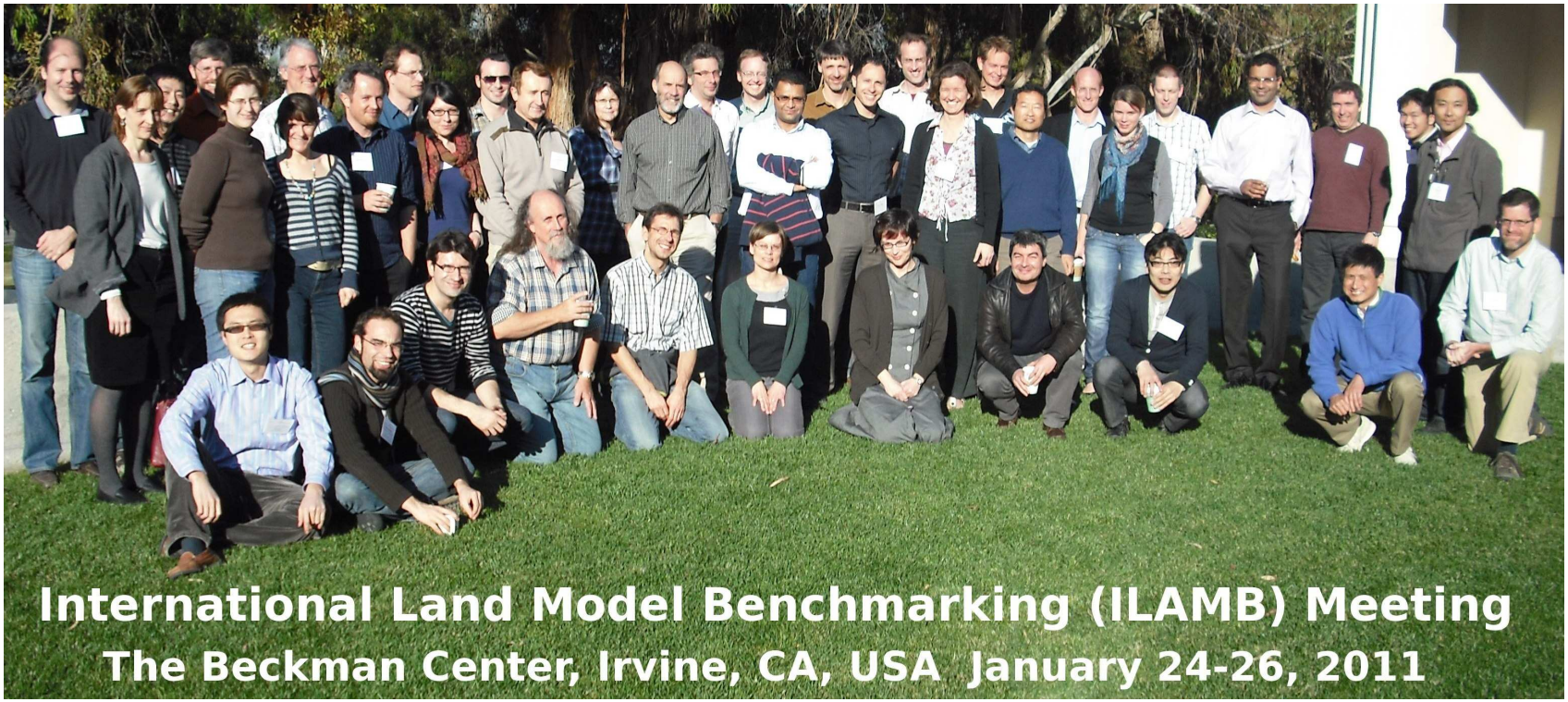
Example Benchmark Score Sheet from C-LAMP

Models →

BGC Datasets ↓

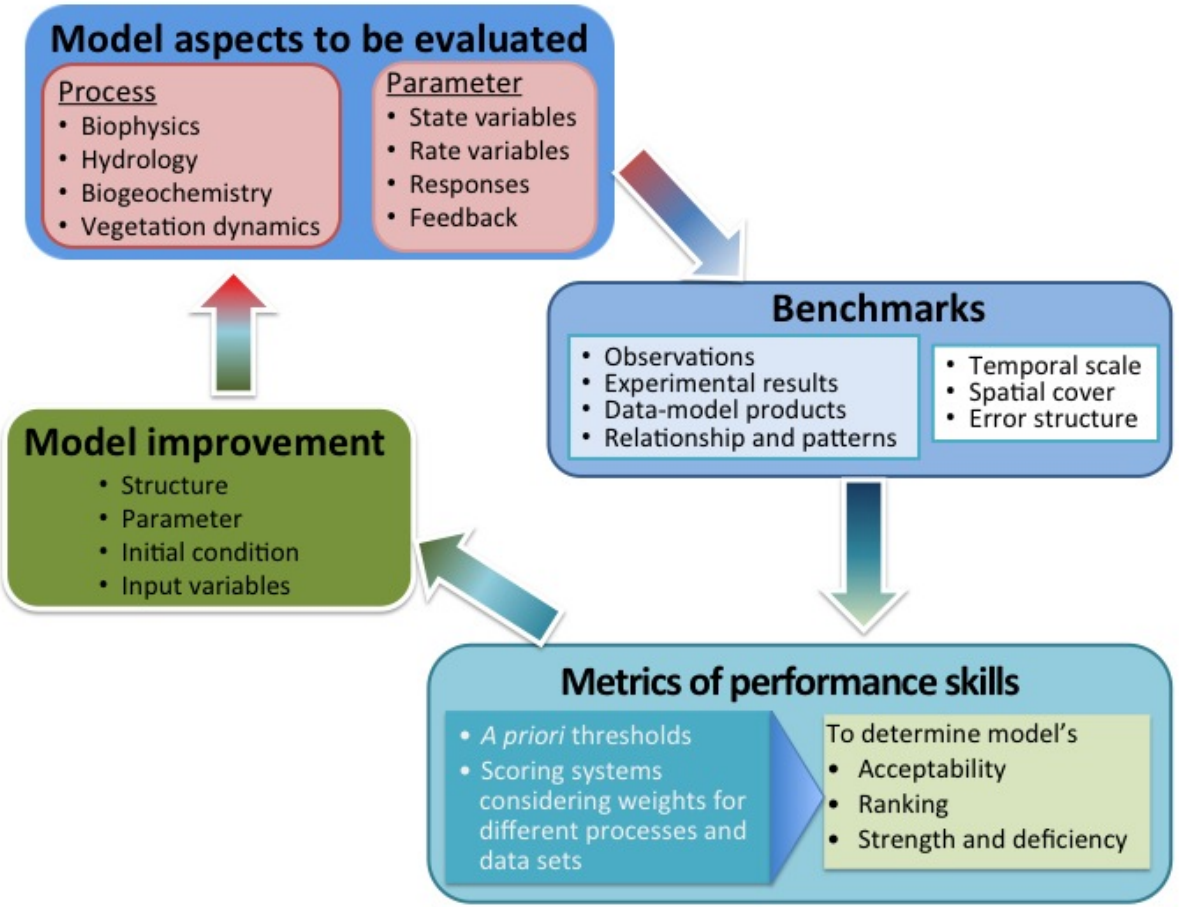
Metric	Metric components	Uncertainty of obs.	Scaling mismatch	Total score	Sub-score	CASA′	CN
LAI	Matching MODIS observations			15.0		13.5	12.0
	• Phase (assessed using the month of maximum LAI)	Low	Low		6.0	5.1	4.2
	• Maximum (derived separately for major biome classes)	Moderate	Low		5.0	4.6	4.3
	• Mean (derived separately for major biome classes)	Moderate	Low		4.0	3.8	3.5
NPP	Comparisons with field observations and satellite products			10.0		8.0	8.2
	• Matching EMDI Net Primary Production observations	High	High		2.0	1.5	1.6
	• EMDI comparison, normalized by precipitation	Moderate	Moderate		4.0	3.0	3.4
	• Correlation with MODIS (r^2)	High	Low		2.0	1.6	1.4
	• Latitudinal profile comparison with MODIS (r^2)	High	Low		2.0	1.9	1.8
CO_2 annual cycle	Matching phase and amplitude at Globalview flash sites			15.0		10.4	7.7
	• 60°–90°N	Low	Low		6.0	4.1	2.8
	• 30°–60°N	Low	Low		6.0	4.2	3.2
	• 0°–30°N	Moderate	Low		3.0	2.1	1.7
Energy & CO_2 fluxes	Matching eddy covariance monthly mean observations			30.0		17.2	16.6
	• Net ecosystem exchange	Low	High		6.0	2.5	2.1
	• Gross primary production	Moderate	Moderate		6.0	3.4	3.5
	• Latent heat	Low	Moderate		9.0	6.4	6.4
	• Sensible heat	Low	Moderate		9.0	4.9	4.6
Transient dynamics	Evaluating model processes that regulate carbon exchange on decadal to century timescales			30.0		16.8	13.8
	• Aboveground live biomass within the Amazon Basin	Moderate	Moderate		10.0	5.3	5.0
	• Sensitivity of NPP to elevated levels of CO_2: comparison to temperate forest FACE sites	Low	Moderate		10.0	7.9	4.1
	• Interannual variability of global carbon fluxes: comparison with TRANSCOM	High	Low		5.0	3.6	3.0
	• Regional and global fire emissions: comparison to GFEDv2	High	Low		5.0	0.0	1.7
			Total:	**100.0**		**65.9**	**58.3**

(Randerson et al., 2009)



- Meeting Co-organized by Forrest Hoffman (UC-Irvine and ORNL), Chris Jones (UK Met Office), Pierre Friedlingstein (U. Exeter and IPSL-LSCE), and Jim Randerson (UC-Irvine).
- About 45 researchers participated from the United States, Canada, the United Kingdom, the Netherlands, France, Germany, Switzerland, China, Japan, and Australia.

General Benchmarking Procedure

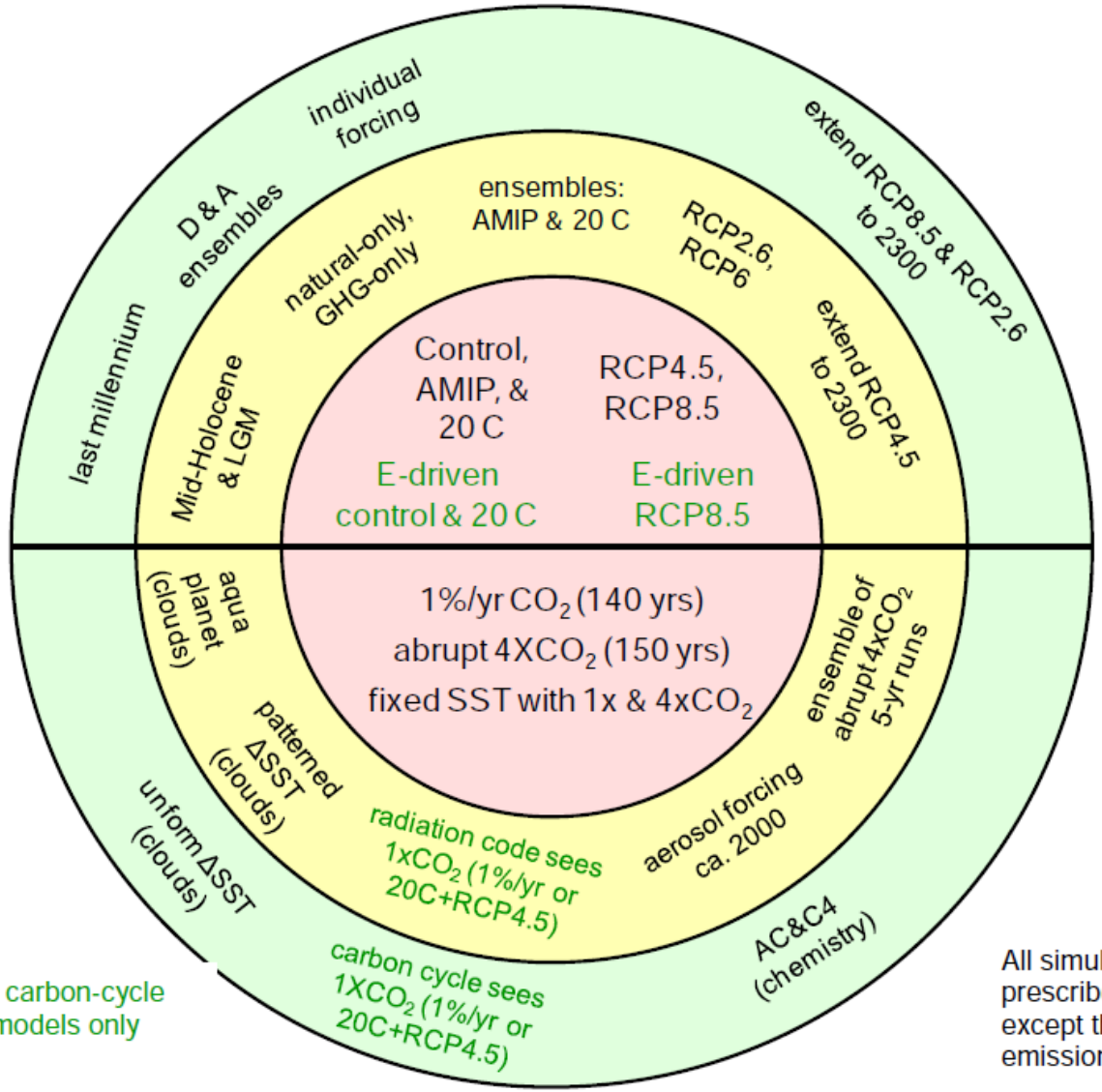


(Luo et al., 2012)

ILAMB 1.0 Benchmarks Now Under Development

	Annual Mean	Seasonal Cycle	Interannual Variability	Trend	Data Source
Atmospheric CO_2					
Flask/conc. + transport		✓	✓	✓	NOAA, SIO, CSIRO
TCCON + transport		✓	✓	✓	Caltech
Fluxnet					
GPP, NEE, TER, LE, H, RN	✓	✓	✓		Fluxnet, MAST-DC
Gridded: GPP	✓	✓	?		MPI-BGC
Hydrology/Energy					
runoff ratio (R/P) ~~river flow~~	✓		✓		GRDC, Dai, GFDL
global runoff/ocean balance	✓				Syed/Famiglietti
albedo (multi-band)		✓	✓		MODIS, CERES
soil moisture		✓	✓		de Jeur, SMAP
column water		✓	✓		GRACE
snow cover	✓	✓	✓	✓	AVHRR, GlobSnow
snow depth/SWE	✓	✓	✓	✓	CMC (N. America)
T_{air} & P	✓	✓	✓	✓	CRU, GPCP and TRMM
Gridded: LE, H	✓	✓			MPI-BGC, dedicated ET
Ecosystem Processes & State					
soil C, N	✓				HWSD, MPI-BGC
litter C, N	✓				LIDET
soil respiration	✓	✓	✓	✓	Bond-Lamberty
FAPAR	✓	✓			MODIS, SeaWIFS
biomass & change	✓			✓	Saatchi, Pan, Blackard
canopy height	✓				Lefsky, Fisher
NPP	✓				EMDI, Luyssaert
Vegetation Dynamics					
fire — burned area	✓	✓	✓		GFED3
wood harvest	✓			✓	Hurtt
land cover	✓				MODIS PFT fraction

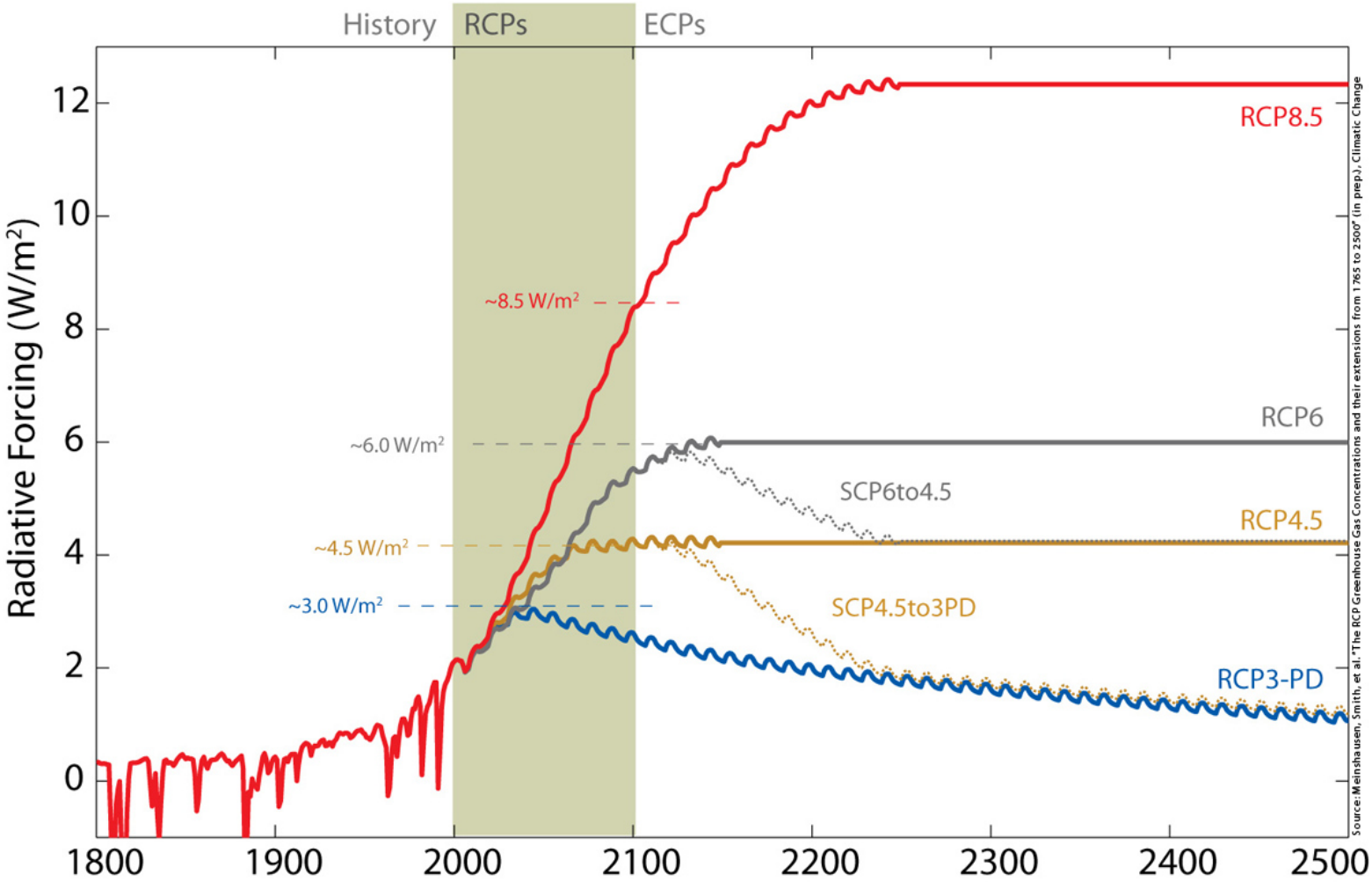
CMIP5 Long-Term Experiments (Taylor et al., 2012)



Coupled carbon-cycle climate models only

All simulations are forced by prescribed concentrations except those "E-driven" (i.e., emission-driven)

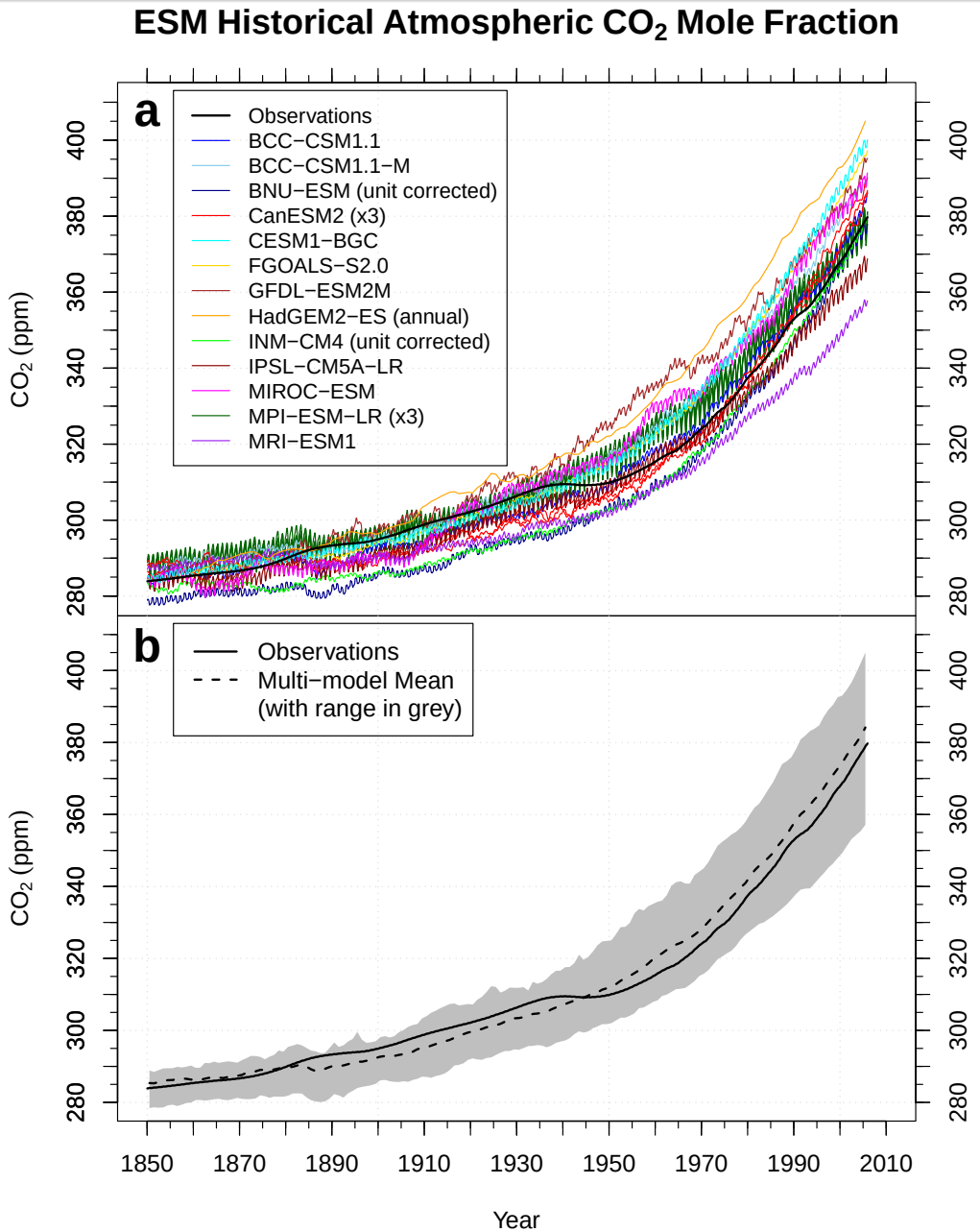
Total Radiative Forcing Estimates for RCPs and ECPs

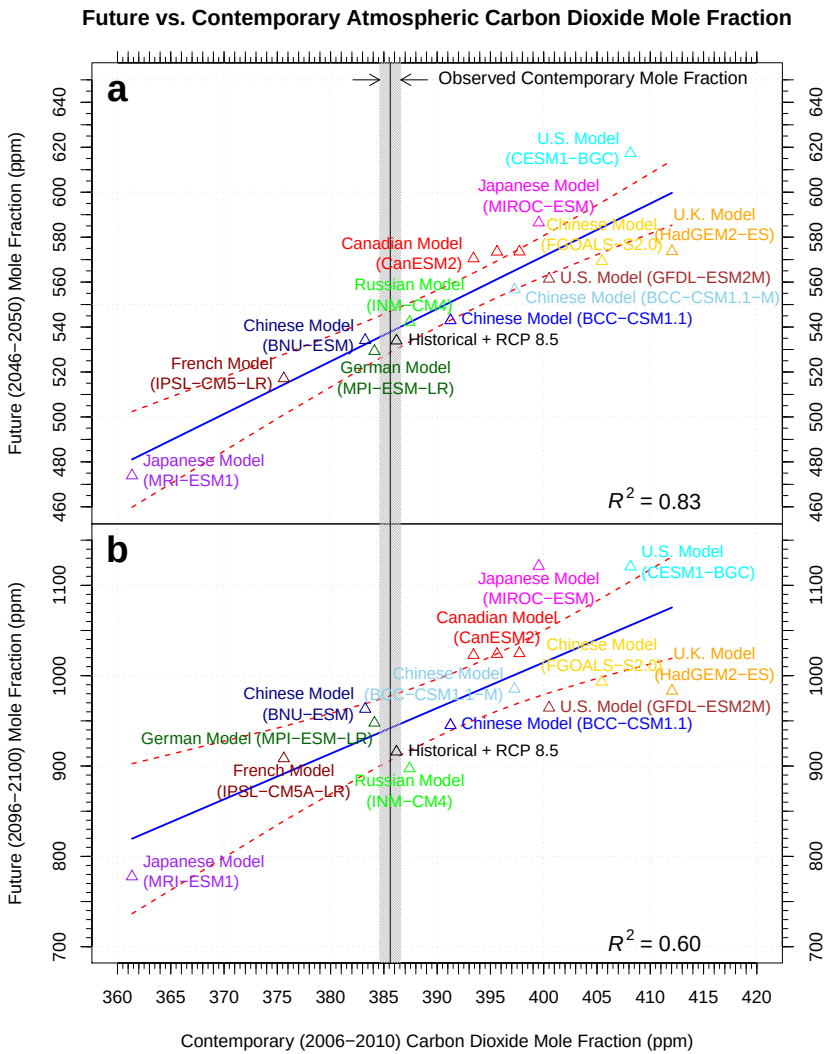


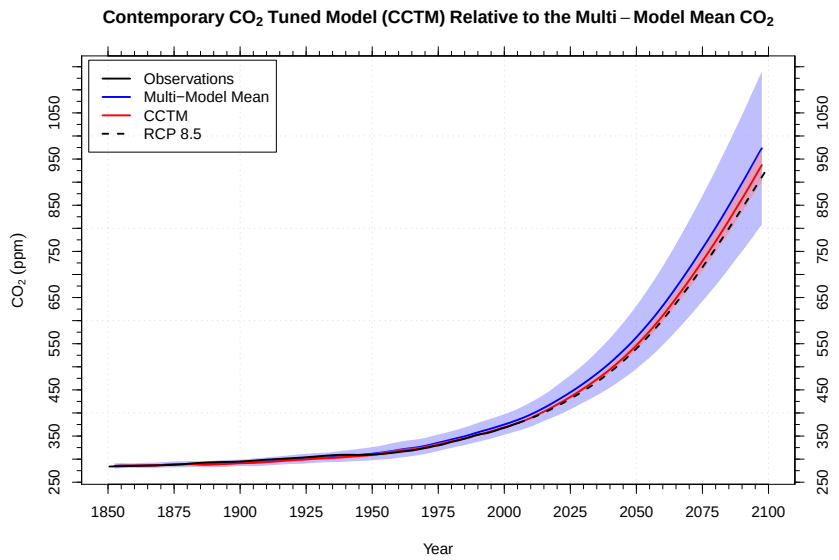
Meinshausen et al. (2011) extended the RCP forcings out to 2500. Data are available at `http://www.pik-potsdam.de/~mmalte/rcps/`

Most CMIP5 Earth System Models (ESMs) exhibit a positive bias in atmospheric CO_2 by the end of the observational era.

(Hoffman et al., in prep.)



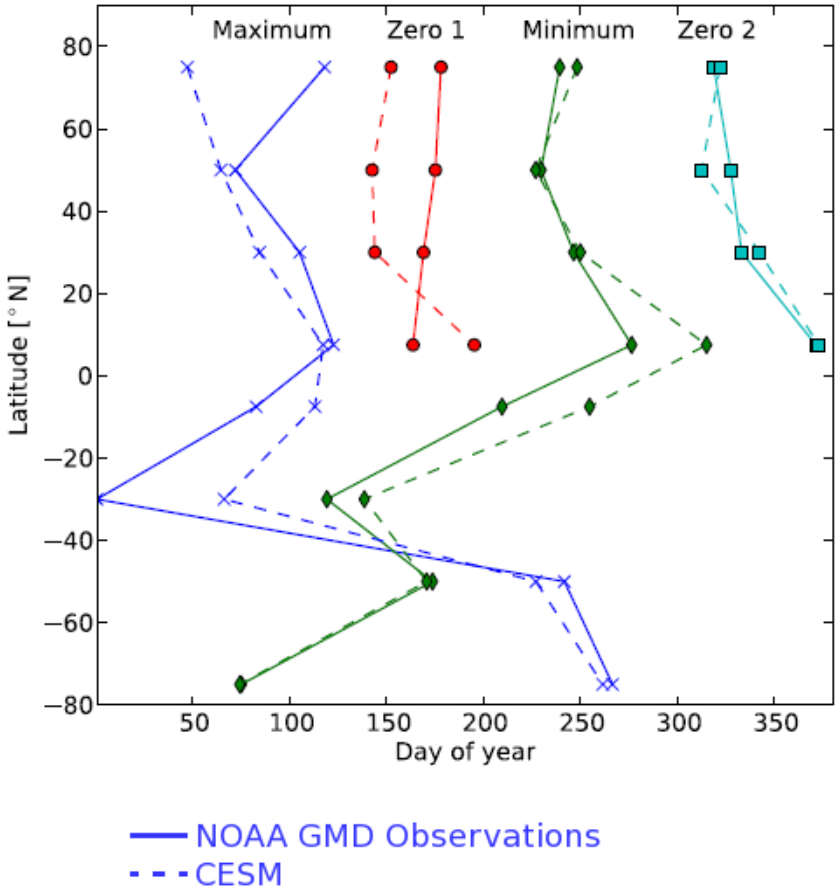




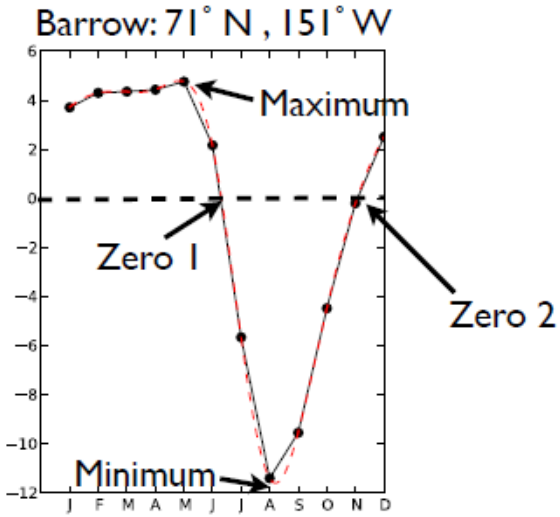
(Hoffman et al., in prep.)

Multi-model estimates and contemporary observations can be used to reduce uncertainties in future scenarios.

The phase of the annual cycle of atmospheric CO_2



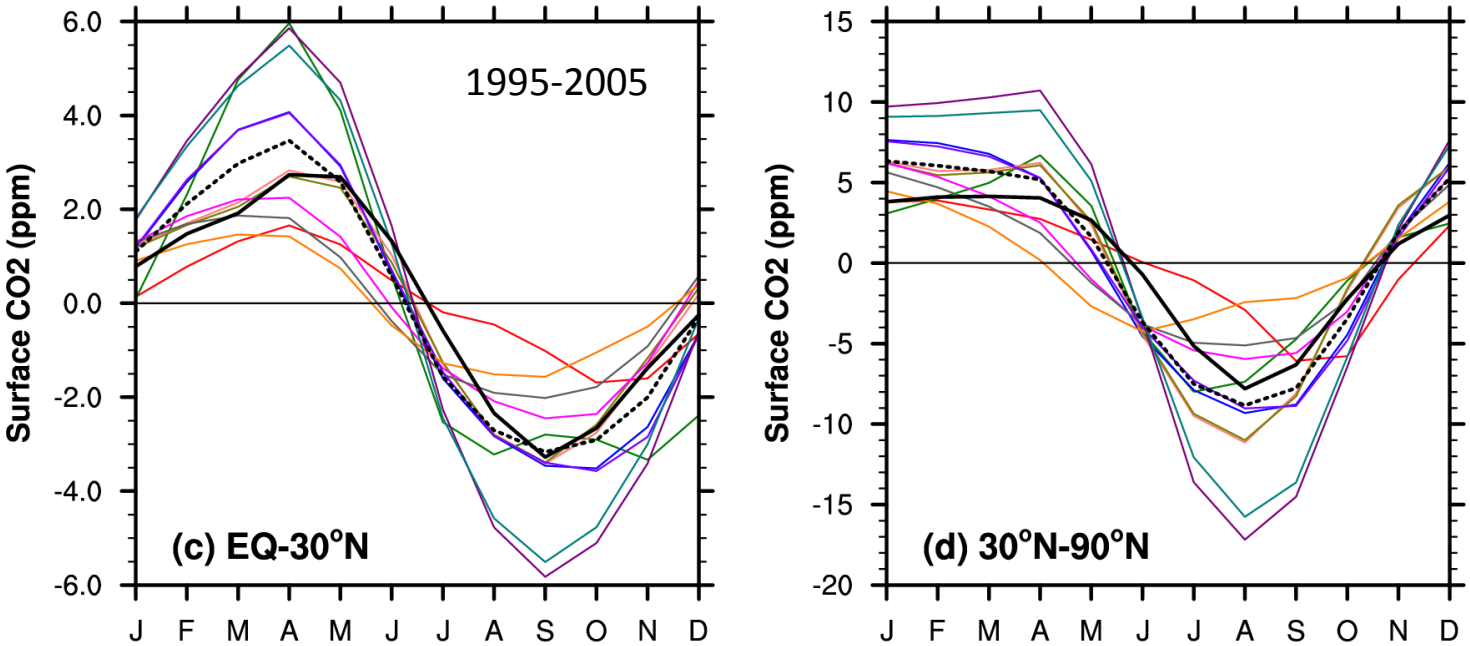
The onset of the growing season occurs too early in CESM.



(Keppel-Aleks et al., *J. Clim.*, in press)

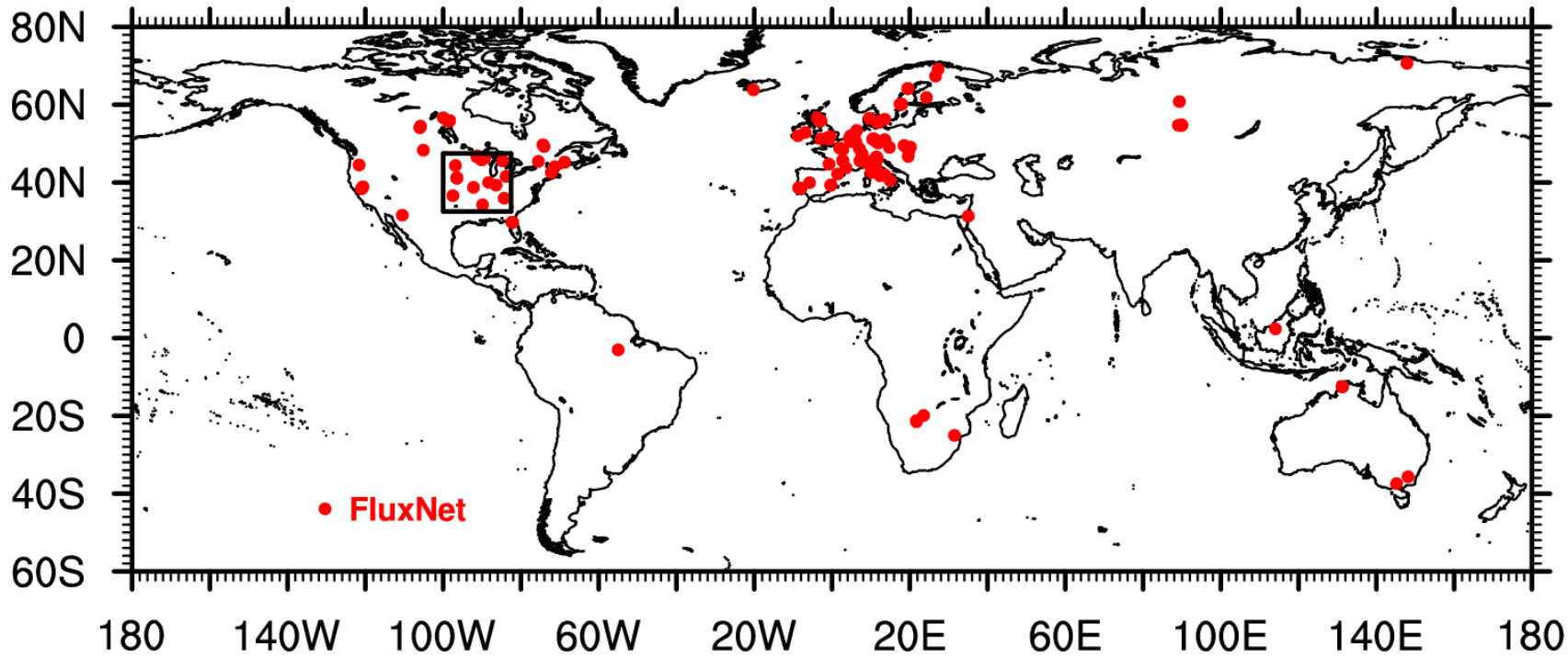
Atmospheric CO_2 is drawn down too early in spring in most CMIP5 Earth system models

- GEOS-Chem with CMIP5 NEE and prescribed ocean and fossil fuel fluxes, sampled at NOAA GMD stations and compared with observations (1995-2005)



(Randerson, Mu et al., in prep.)

What are the causes of the early season uptake bias?
Eddy covariance observations from FLUXNET provide constraints

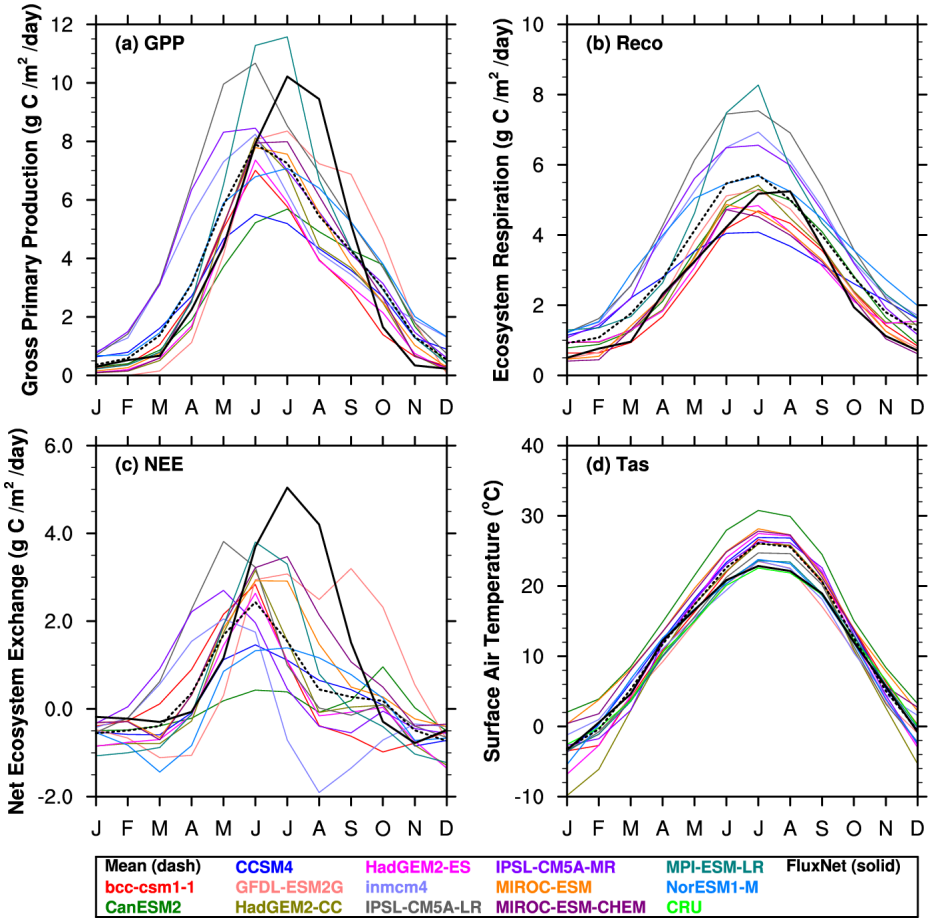


(Randerson, Mu et al., in prep.)

GPP appears to be the primary culprit for the early NEE uptake and CO_2 drawdown

- Fluxnet sites in North America between 35N and 45N

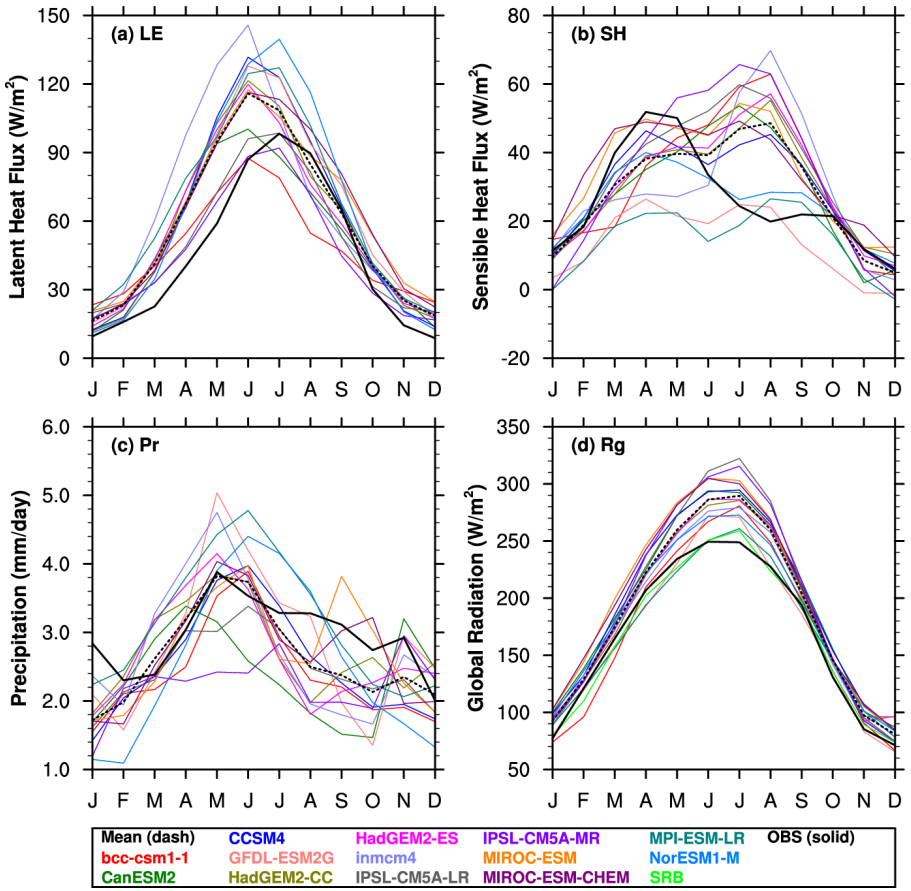
Model grid cells extracted and sampled at all measurement sites



(Randerson, Mu et al., in prep.)

Early onset of photosynthesis may have consequences for the seasonal dynamics of surface energy exchange

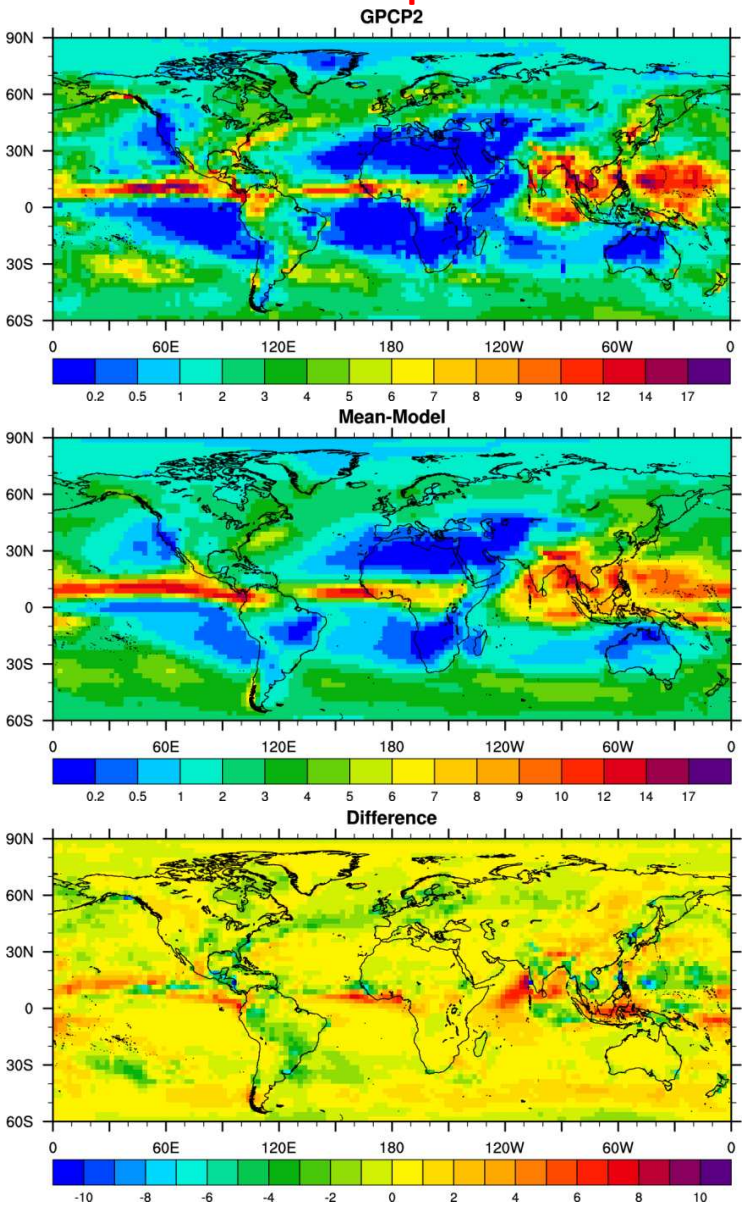
- Fluxnet sites in North America between 35N and 45N



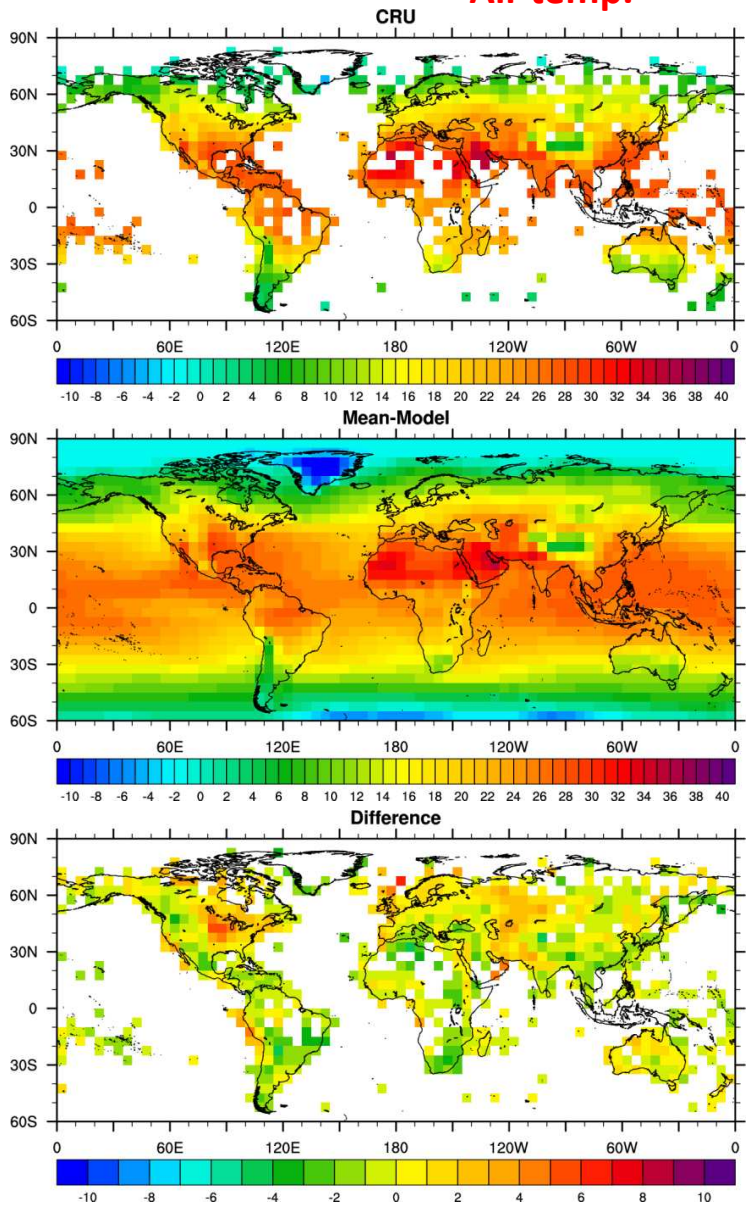
(Randerson, Mu et al., in prep.)

Multi-year (1986-2005) mean of Precip and Tas in August

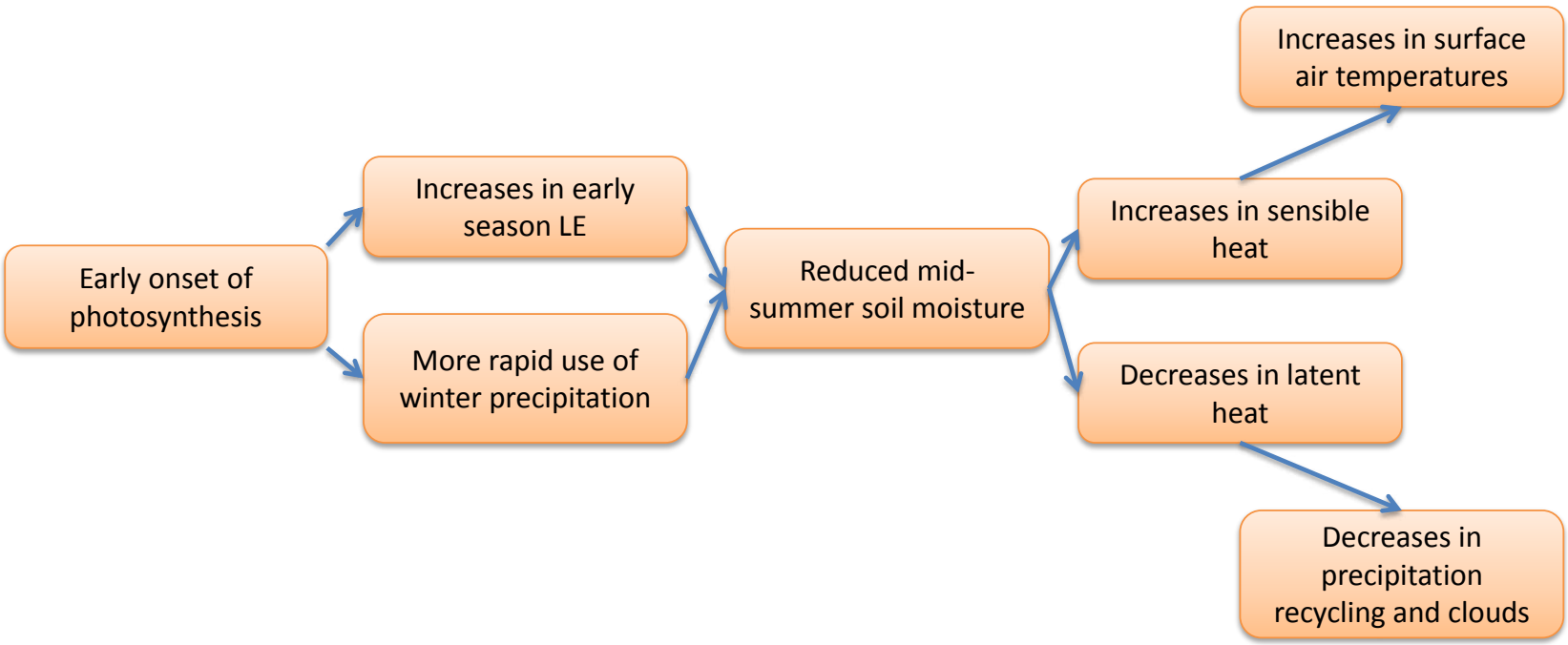
Precip.



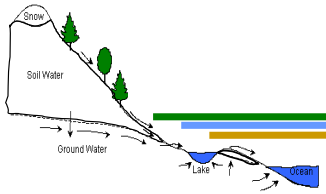
Air temp.



How much of the mid-summer climate and GPP biases are a consequence of missing physiology vs. issues with other parts of the climate system?



(Randerson, Mu et al., in prep.)



Example: Metric based on global RMSE

$$RMSE = \sqrt{\frac{1}{W}\sum_i\sum_j\sum_t w_{ijt}(Model_{ijt} - Obs_{ijt})^2}$$

$$M = \quad 0 < \quad 1 - \frac{RMSE}{c\sigma_{obs,ijt}}$$

RMSE Metric tests spatial pattern and annual cycle

Not a direct test of a land model process

w_{ijt} are the area and month weights, W is the total of the weights

i,j,t are the lon,lat, and month

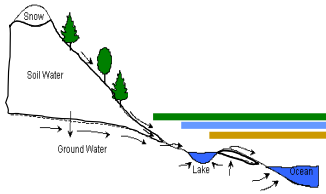
Model and Obs are climatological data (20yr avg)

Amplitude of σ_{ijt} controls average position of M (0-1),

influence of σ_{ijt} can be adjusted with c

What is best/most meaningful normalization ?

(Lawrence et al., in prep.)



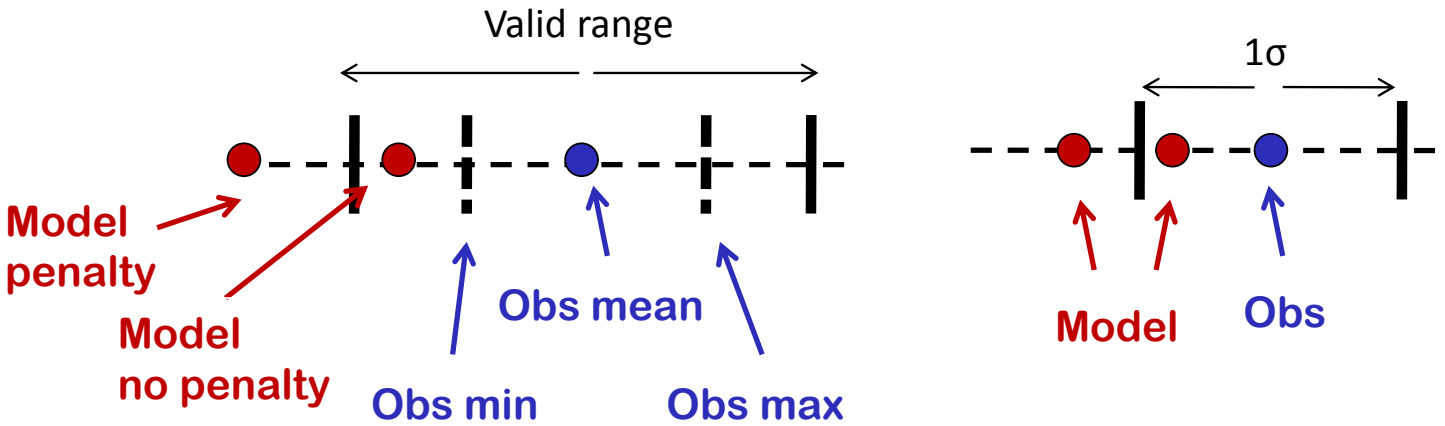
Accounting for observational error

Two methods:

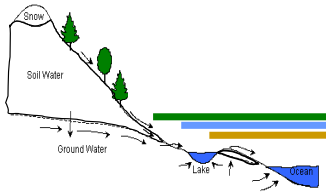
If multiple datasets available

Valid range = mean(multiple obs datasets) +- max-min(multiple datasets)

If dataset has observational error estimate (like FLUXNET-MTE does, see next slide, then use that

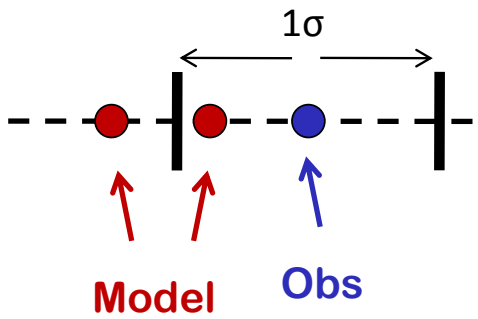


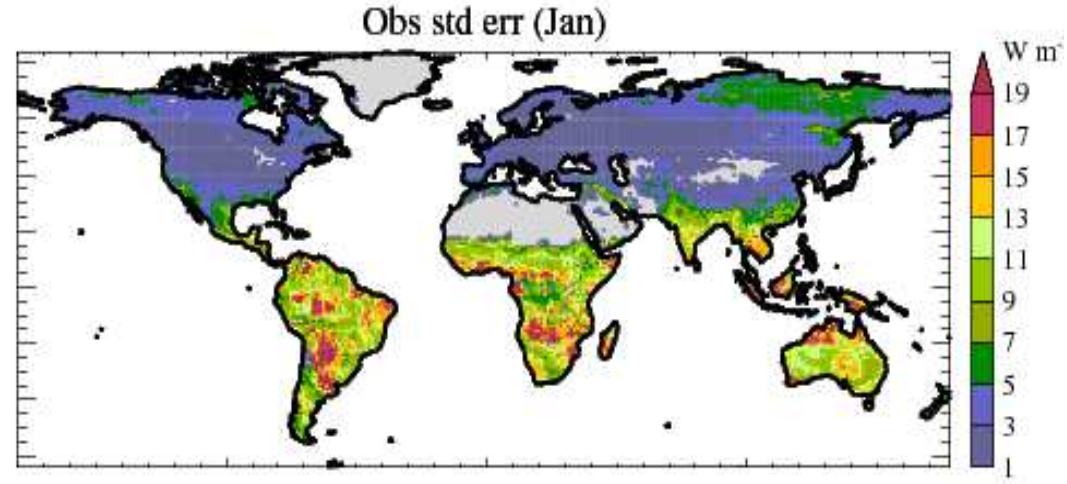
(Lawrence et al., in prep.)

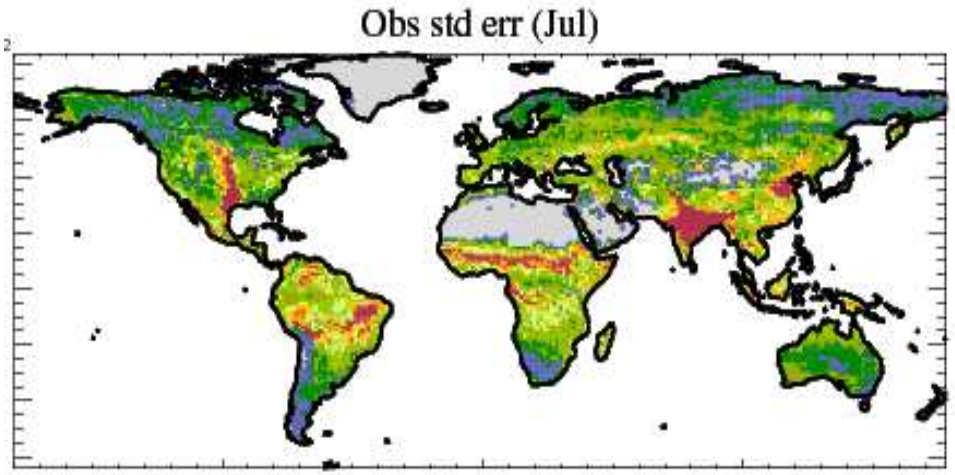


FLUXNET-MTE (1 σ error estimate)

FLUXNET-MTE product error varies geographically and in time based in part on FLUXNET sampling biases







(Lawrence et al., in prep.)

Global RMSE (LH, centered); FLUXNET-MTE (Jung et al. 2010)

bcc-csm1-1:
RMSE = 14.56W m^{-2}, M = 0.31

CCSM:
RMSE = 15.02W m^{-2}, M = 0.28

NorESM1-M:
RMSE = 15.61W m^{-2}, M = 0.26

CanESM2:
RMSE = 15.96W m^{-2}, M = 0.24

HadGEM2-CC:
RMSE = 14.10W m^{-2}, M = 0.33

HadGEM2-ES:
RMSE = 14.29W m^{-2}, M = 0.32

inmcm4:
RMSE = 19.67W m^{-2}, M = 0.06

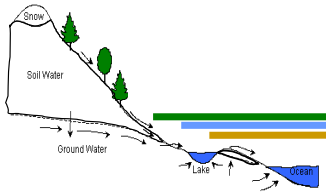
IPSL-CM5A-MR:
RMSE = 12.71W m^{-2}, M = 0.39

MIROC-ESM-CHEM:
RMSE = 16.40W m^{-2}, M = 0.22

MIROC-ESM:
RMSE = 16.48W m^{-2}, M = 0.21

GFDL-ESM2G:
RMSE = 19.29W m^{-2}, M = 0.08

MPI-ESM-LR:
RMSE = 15.76W m^{-2}, M = 0.25

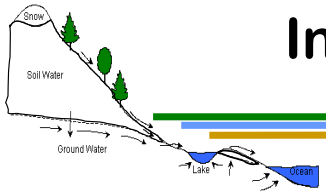


Global land RMSE metric: Latent Heat Flux

obs: FLUXNET-MTE LH (Jung et al. 2010)

	RMSE	M
CCSM4(CN)	25.1	0.25
CLM3.5(SP)	15.6	0.54
CLM4(CN)	18.3	0.45
CLM4(SP)	15.4	0.51

(Lawrence et al., in prep.)



Index of performance (MAE) (Willmott et al., 2011, Int. J. Climatology)

$$d_{ij} = 1 - \frac{\sum_t \left| Model_{ijt} - Obs_{ijt} \right|}{c \sum_t \left| Obs_{ijt} - \overline{Obs_{ij}} \right|}$$

when numerator > denominator

$$d_{ij} = \frac{c \sum_t \left| Obs_{ijt} - \overline{Obs_{ij}} \right|}{\sum_t \left| Model_{ijt} - Obs_{ijt} \right|} - 1$$

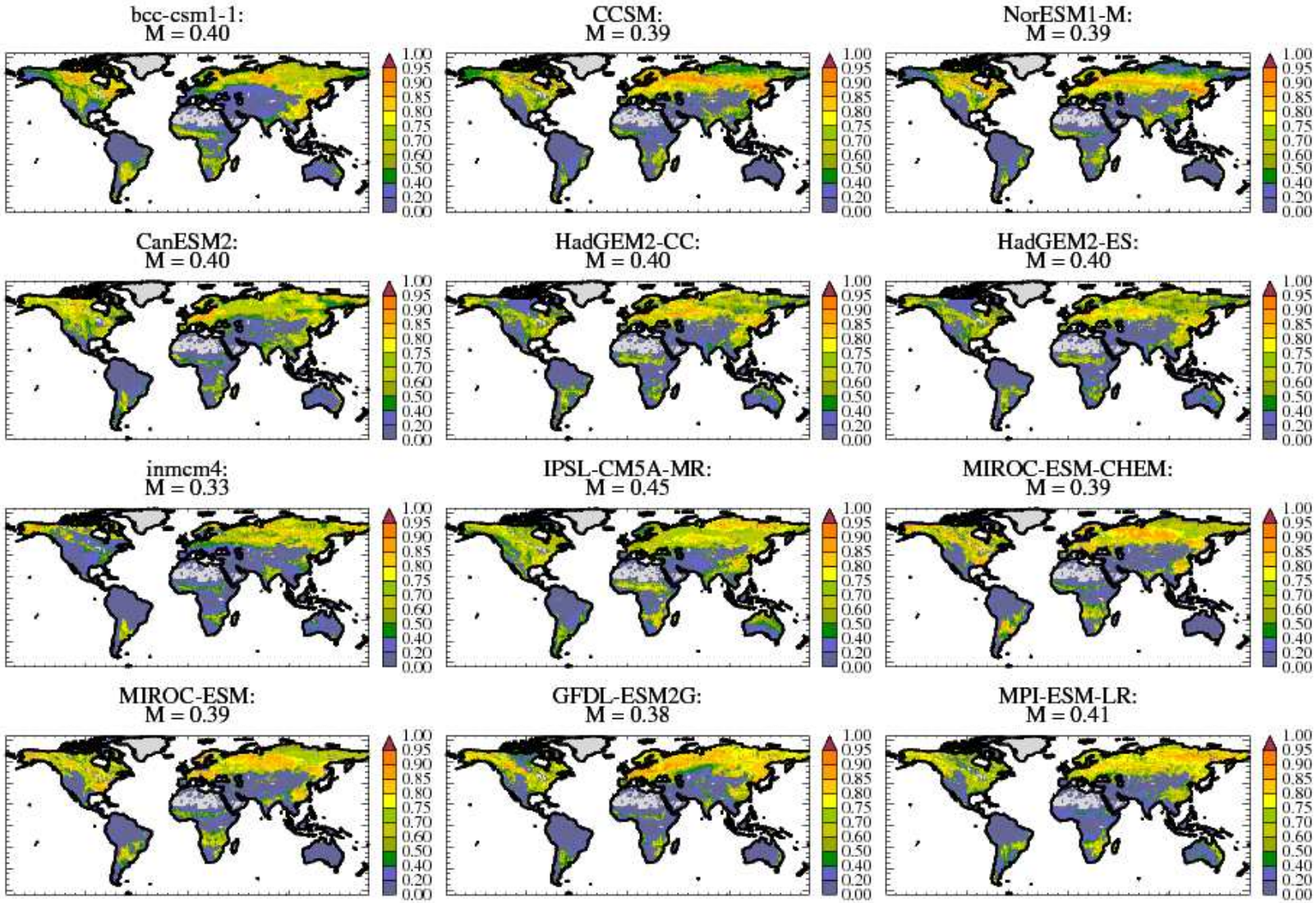
when denominator > numerator

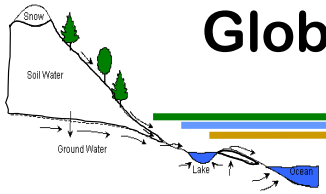
$$M = \sum_{ij} w_{ij} \left(d_{ij} + 1 \right) / 2$$

Willmott et al. argue that metrics based on RMS overweight the influence of a few large errors and that MAE metrics are more indicative of general model performance

(Lawrence et al., in prep.)

Global MAE (LH, centered); FLUXNET-MTE (Jung et al. 2010)



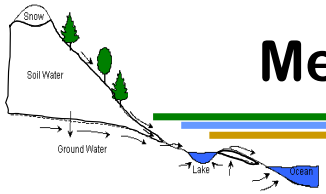


Global land RMSE and MSE metrics: Latent Heat Flux

obs: FLUXNET-MTE LH (Jung et al. 2010)

	M (RMSE)	M (MAE)
CCSM4(CN)	0.25	0.62
CLM3.5(SP)	0.54	0.72
CLM4(CN)	0.45	0.69
CLM4(SP)	0.54	0.72
CLM4(SP-GPP)	0.58	0.74

(Lawrence et al., in prep.)

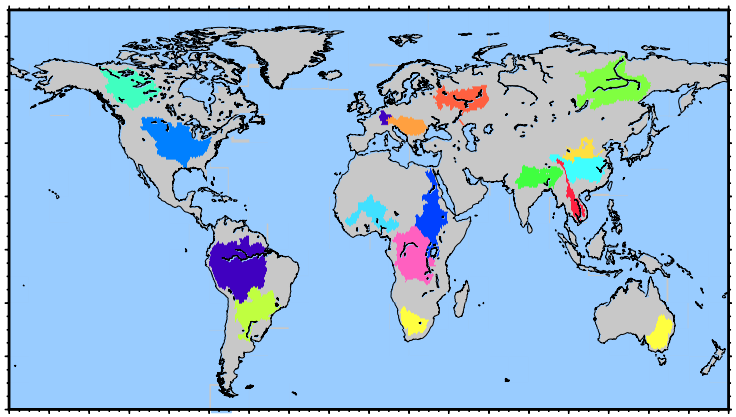


Metric for Runoff Ratio (R/P) (top 20 river basins)

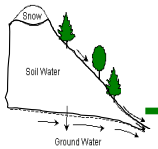
$$\mathrm{MSE} = \frac{1}{\mathrm{A}} \sum_{n=1}^{20} a_n (\mathrm{Model}_n - \mathrm{Obs}_n)^2$$

$$\mathrm{M} = 0 < 1 - \frac{\mathrm{MSE}}{\sigma_{\mathrm{obs}}^2} < 1$$

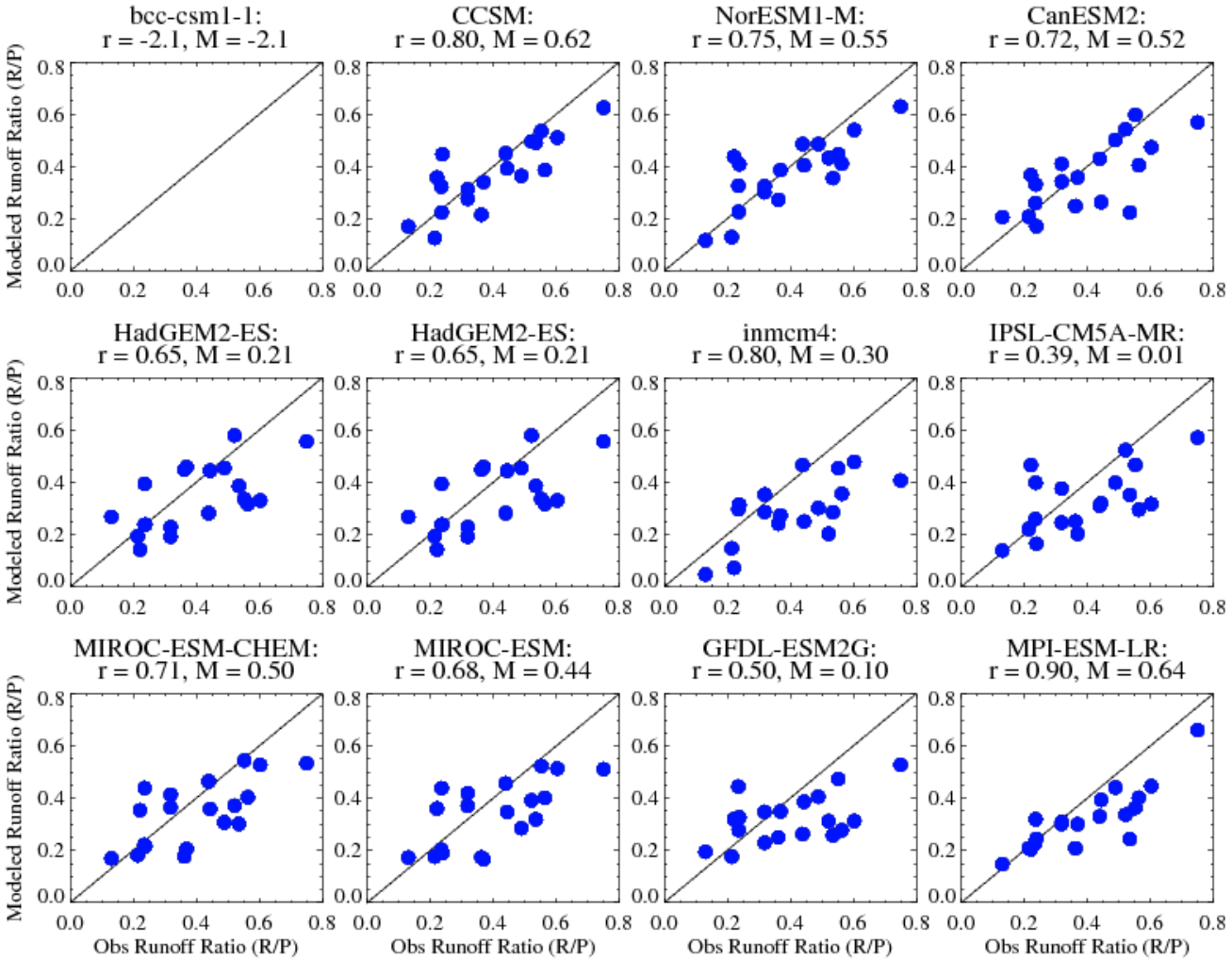
a_n is river basin area, A is total area across all basins,
σ is variance of obs weighted for basin size



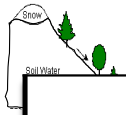
(Lawrence et al., in prep.)



Runoff ratio (R/P): Top 20 rivers



(Lawrence et al., in prep.)

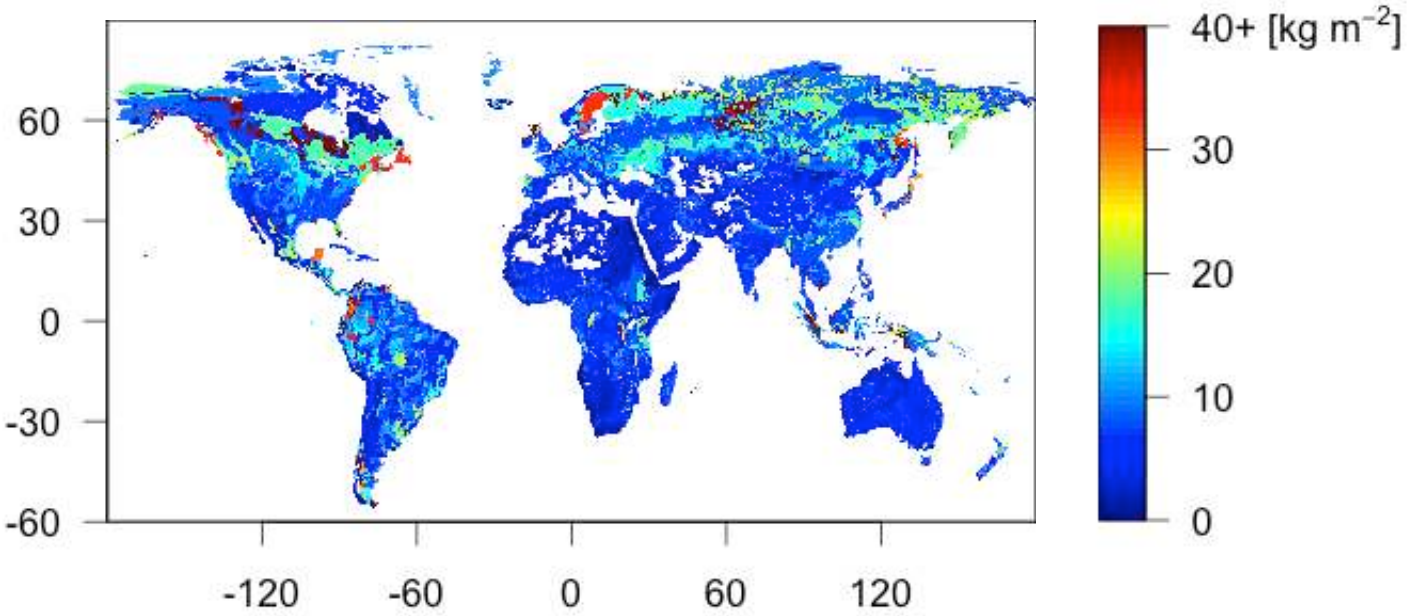


Synthesis

Class	Variable	Obs dataset	W (1-5)	CCSM4	CLM3.5 SP	CLM4 CN	CLM4 SP	CLM4 SPGP P
Global (or regional) RMSE	LH	FLUXNET-MTE	4	0.68	0.71	0.63	0.71	0.74
	SCF	AVHRR	4	0.68	0.64	0.75	0.74	0.74
	Snow Depth	CMC	2	0.53	0.70	0.73	0.70	0.71
	Albedo	MODIS	5	0.44	0.35	0.52	0.55	0.55
	P	CMAP	3	0.48	0.93	0.93	0.93	0.93
	T_{air}	CRU	3	0.91	0.93	0.93	0.93	0.93
Global Interannual Variability	LH	FLUXNET-MTE	3	0.15				
Basin Runoff (Top 20 biggest river basins)	R / P	riv discharge, CMAP	5	0.63	0.49	0.57	0.55	0.48
	R	river discharge	3	0.22	0.65	0.66	0.68	0.65
	P	CMAP	3	0.62	0.95	0.95	0.95	0.95
Total				18.54	21.38	22.80	23.17	22.87

(Lawrence et al., in prep.)

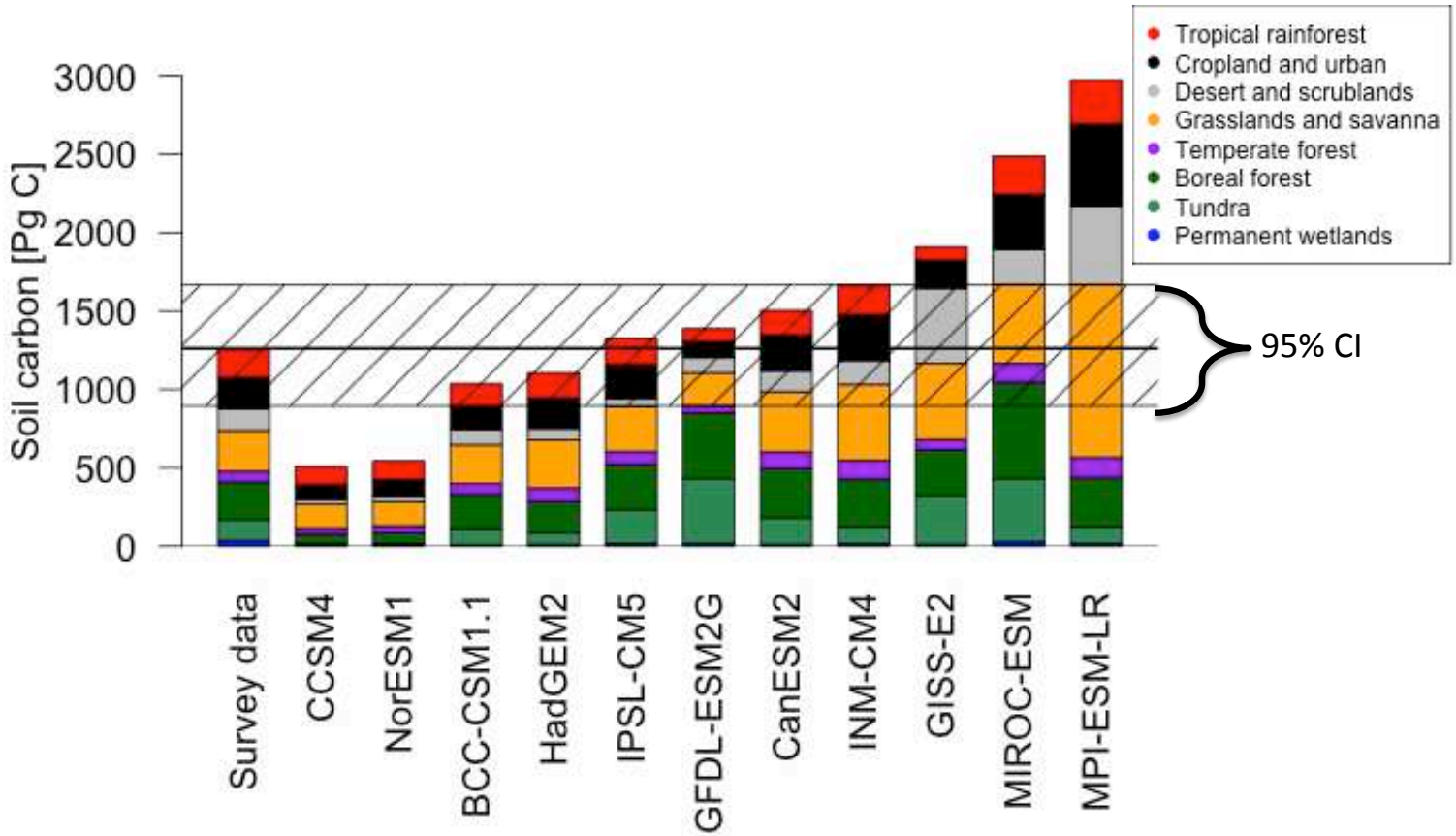
Harmonized World Soil Database 1.2



- Merge of European Soil Data base, Soil Map of China, regional databases, and Soil Map of the World
- Depth of 1 meter on a 0.5° x 0.5° grid

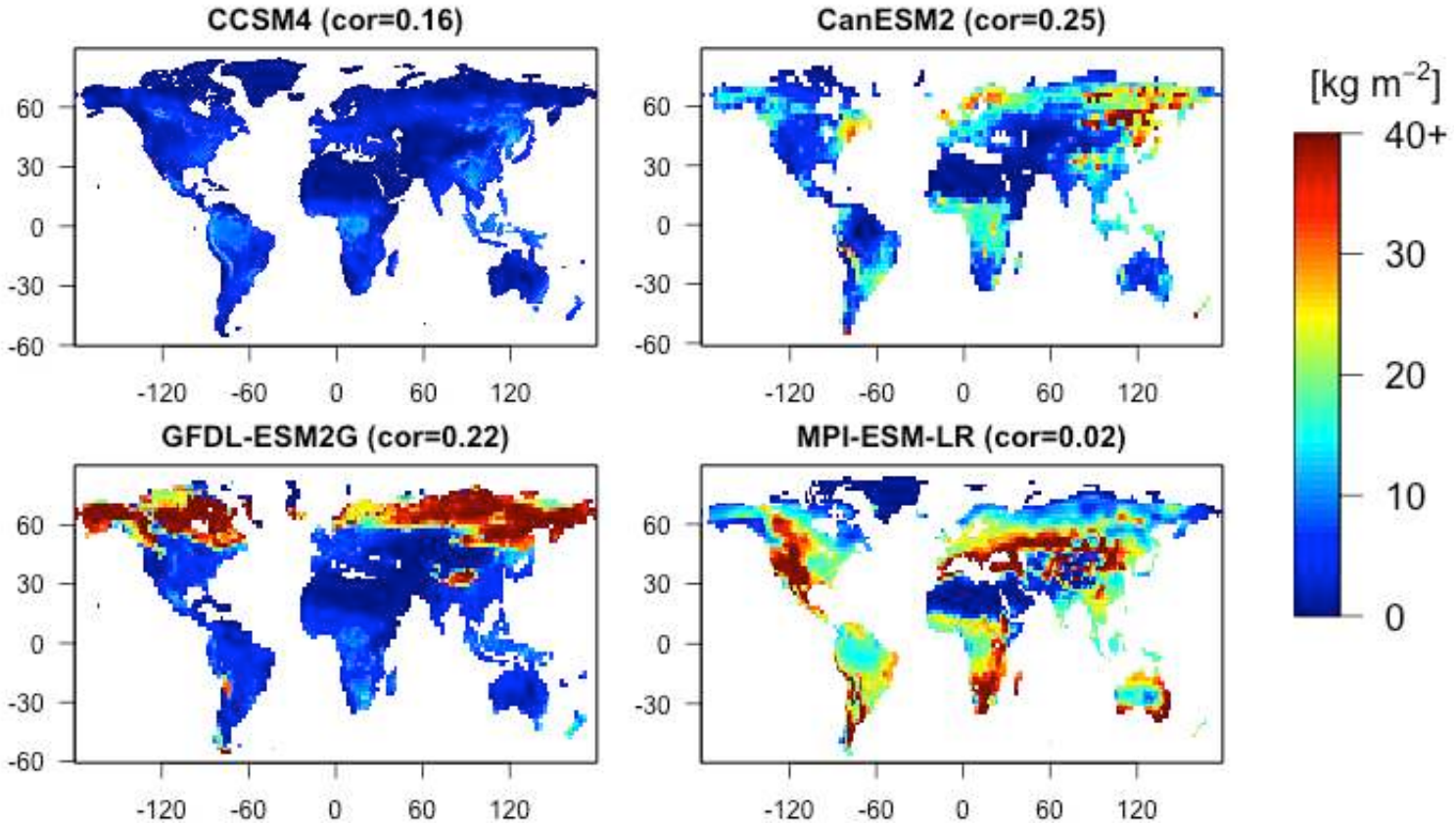
FAO/IIASA/ISRIC/ISSCAS/JRC (2012)

6 of 11 models fall with in survey data



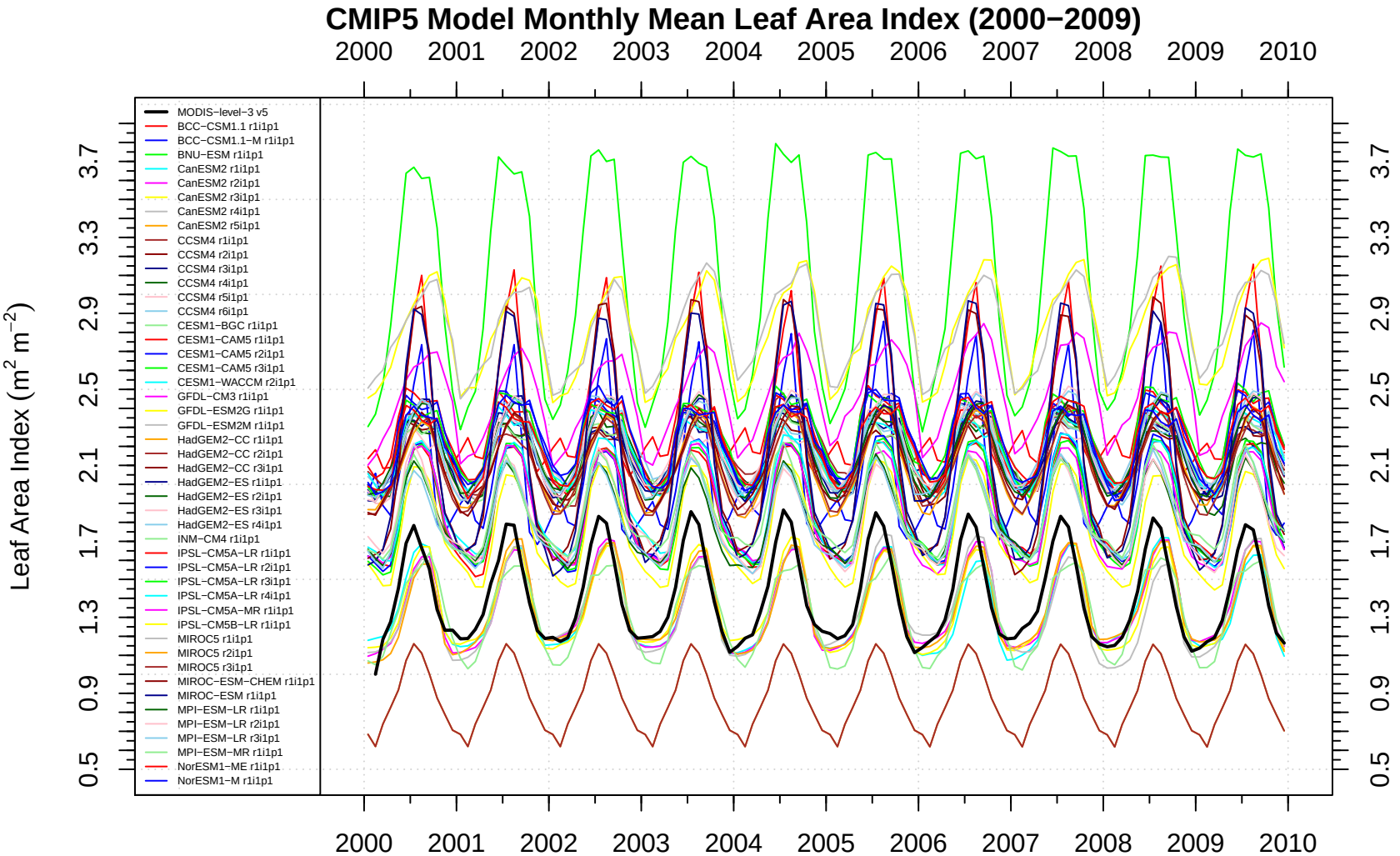
Todd-Brown et al, *Biogeosciences Discussion* (2012)

Grid-by-grid correlation between models and survey data < 0.4



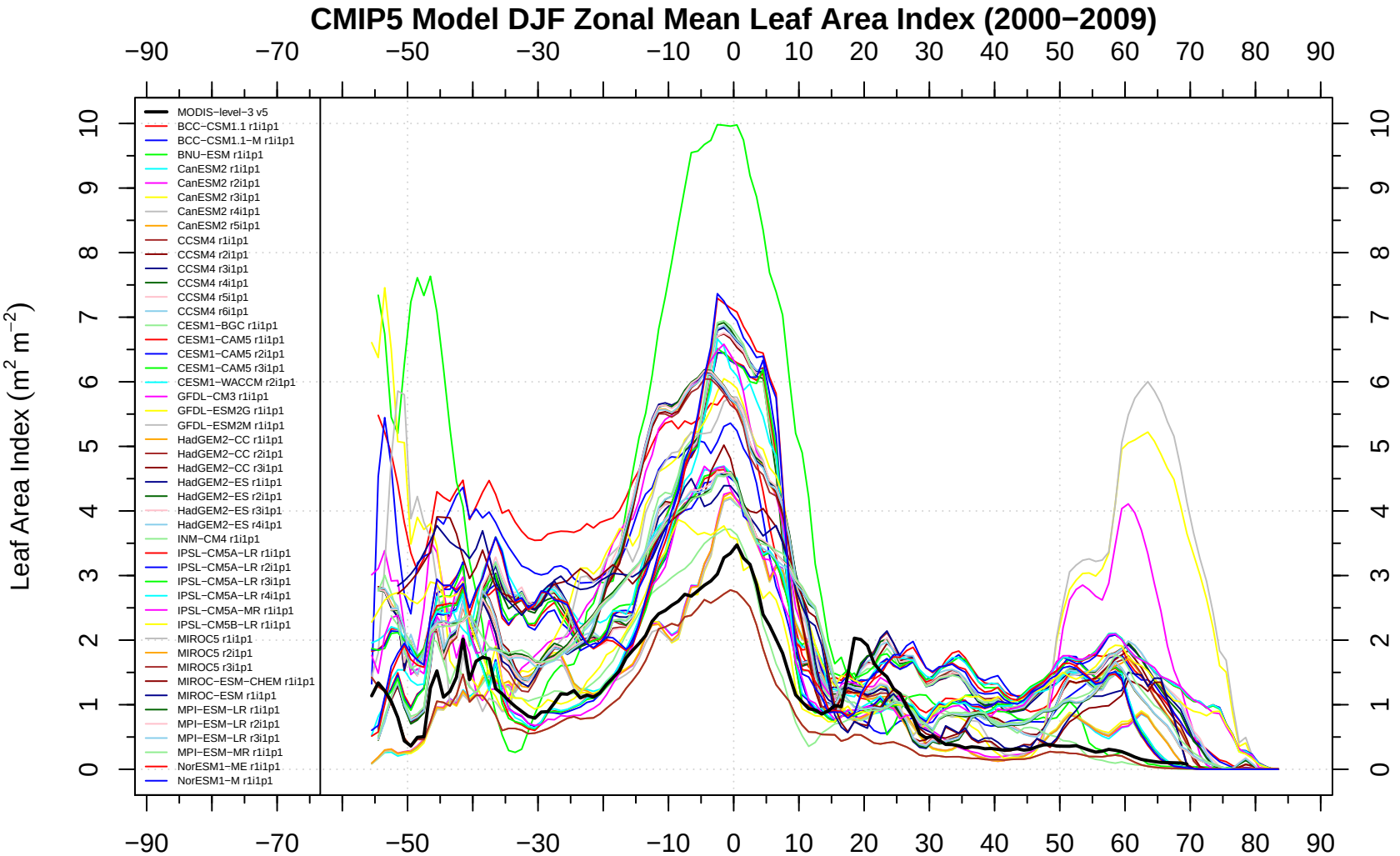
Todd-Brown et al, *Biogeosciences Discussion* (2012)

Global LAI for 47 CMIP5 Simulations Compared to MODIS



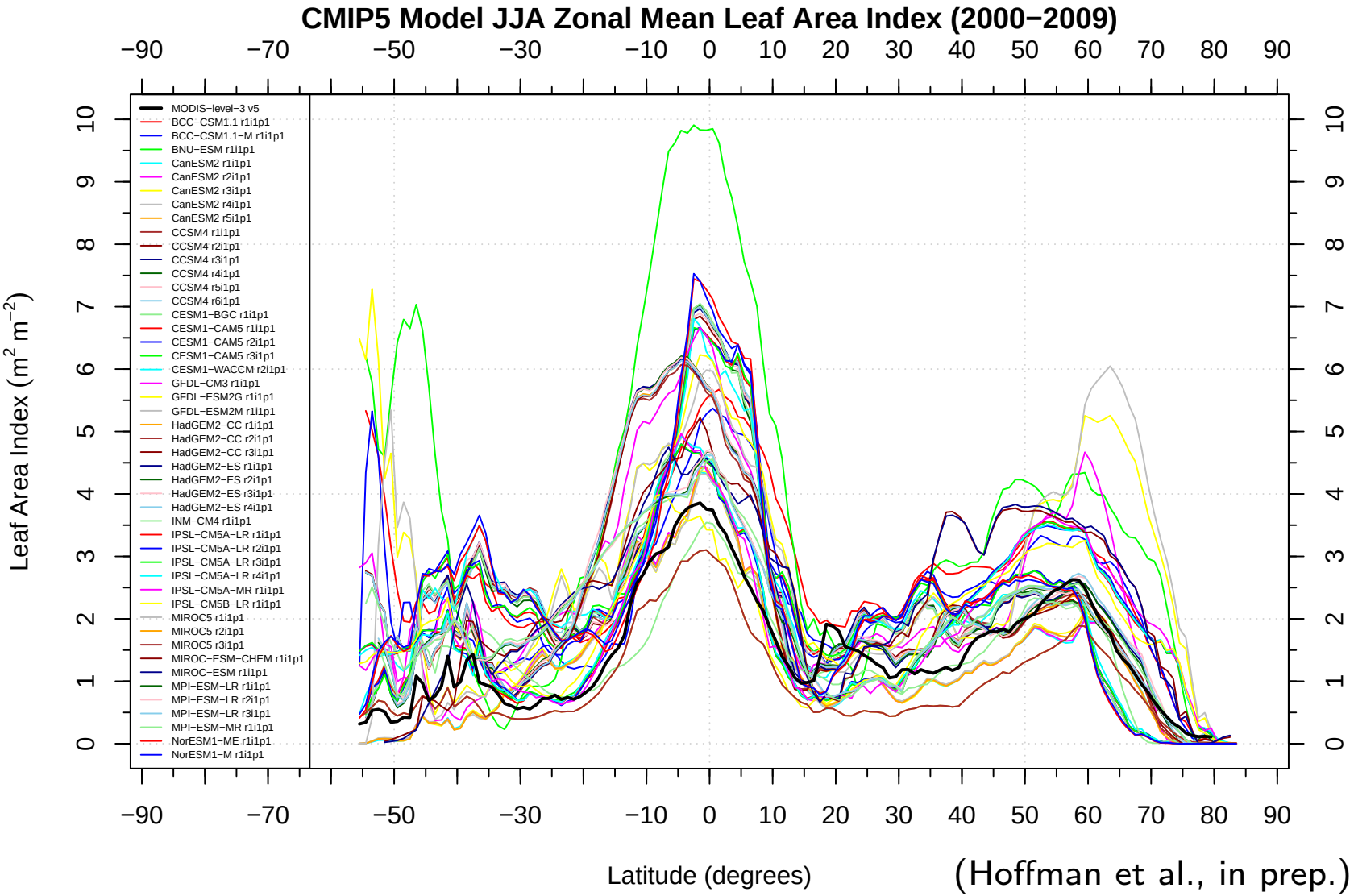
(Hoffman et al., in prep.)

Zonal LAI for 47 CMIP5 Simulations Compared to MODIS



(Hoffman et al., in prep.)

Zonal LAI for 47 CMIP5 Simulations Compared to MODIS



(Hoffman et al., in prep.)

Summary

- Our international collaboration has made significant progress on development of metrics and diagnostics for ILAMB 1.0.
- As CMIP5 papers come out, we need to collect cost functions and algorithms for integration into an ILAMB 1.0 package.
- Much more work is needed on
  - diagnostics for full suite of variables and time scales,
  - combining metrics into model skill scores,
  - applying skill scores to weight models for multi-model statistics, and
  - writing papers.
- Greater participation is welcome!
- ILAMB Meeting in 2013? With ICDC-9 or GLASS/GSWP?

International Land Model Benchmarking (ILAMB) Project

http://www.ilamb.org/

References

M. Jung, M. Reichstein, P. Ciais, S. I. Seneviratne, J. Sheffield, M. L. Goulden, G. Bonan, A. Cescatti, J. Chen, R. de Jeu, A. J. Dolman, W. Eugster, D. Gerten, D. Gianelle, N. Gobron, J. Heinke, J. Kimball, B. E. Law, L. Montagnani, Q. Mu, B. Mueller, K. Oleson, D. Papale, A. D. Richardson, O. Roupsard, S. Running, E. Tomelleri, N. Viovy, U. Weber, C. Williams, E. Wood, S. Zaehle, and K. Zhang. Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature*, 467(7318):951–954, Oct. 2010. doi:10.1038/nature09396.

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.

M. Meinshausen, S. Smith, K. Calvin, J. Daniel, M. Kainuma, J.-F. Lamarque, K. Matsumoto, S. Montzka, S. Raper, K. Riahi, A. Thomson, G. Velders, and D. P. van Vuuren. The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Clim. Change*, 109(1):213–241, Nov. 2011. doi:10.1007/s10584-011-0156-z.

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.

K. E. Taylor, R. J. Stouffer, and G. A. Meehl. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.*, 93(4):485–498, Apr. 2012. doi:10.1175/BAMS-D-11-00094.1.

C. J. Willmott, S. M. Robeson, and K. Matsuura. A refined index of model performance. *Int. J. Climatol.*, 32(13): 2088–2094, Nov. 2012. doi:10.1002/joc.2419.