Scoring Methods in the ILAMB Benchmarking Package

Nathan Collier, Forrest M. Hoffman, Gretchen Keppel-Aleks, David M. Lawrence, Charlie Koven, William J. Riley, and James T. Randerson

Climate Change Science Institute
Oak Ridge National Laboratory

6 Feb 2018

International Land Model Benchmarking (ILAMB)

What is ILAMB?

- group of researchers who develop internationally accepted benchmarks for land model performance and promote their use
- collection of datasets and benchmark techniques
- a general, open-source python package, which disseminates this research

High level summary of model performance

- Measure model performance against **63** datasets across a wide swath of measurable quantities from land models **25** variables
- Left: absolute performance in terms of an *overall* score
- Right: relative performance with respect to other models

International Land Model Benchmarking (ILAMB)

What is ILAMB?

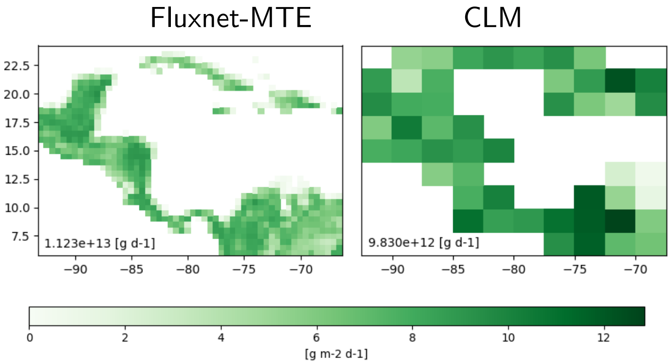
- group of researchers who develop internationally accepted benchmarks for land model performance and promote their use
- collection of datasets and benchmark techniques
- a general, open-source python package, which disseminates this research

This talk is about challenges

- Resolution - observations tend to be high resolution, models are relatively lower resolution
- Representation of land - what each data source calls land varies
- Converting errors to normalized scores

Resolution differences

Consider a plot of gpp zoomed into Central America for emphasis.



Interpolate to a composite grid

If we take two grids defined by the latitude cell breaks, θ, and longitude cell breaks φ,

$$\mathcal{G}_1 := \theta_1 \otimes \varphi_1$$
$$\mathcal{G}_2 := \theta_2 \otimes \varphi_2$$

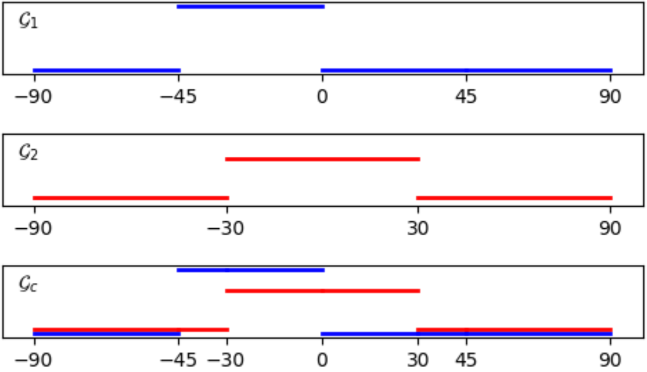
Then we can define a composite grid made up of the union of both grids' breaks,

$$\mathcal{G}_c := (\theta_1 \cup \theta_2) \otimes (\varphi_1 \cup \varphi_2)$$

and interpolate by nearest neighbor with zero interpolation error.

Interpolate to a composite grid

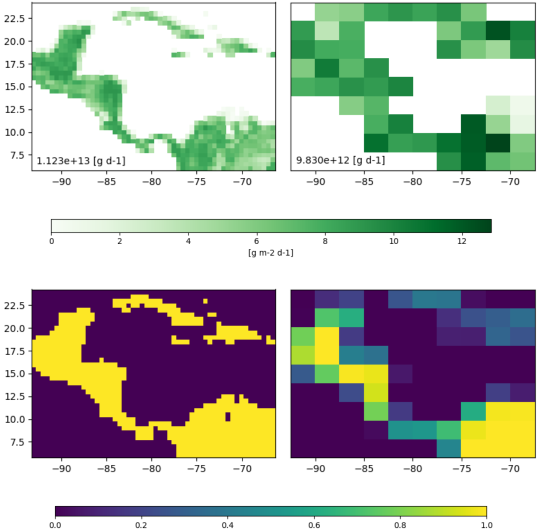
Steps functions from $\mathcal{G}_1$ and $\mathcal{G}_2$ interpolate perfectly to $\mathcal{G}_c$.



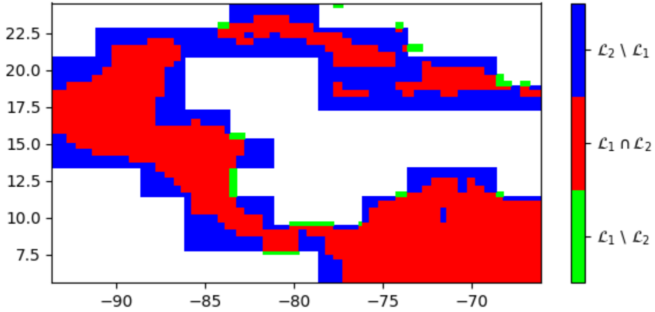
Land definition differences

gpp

land fraction



Represented areas of two sources



Land definition differences

	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.												
CESM1(LENS1)	[-]	131.	125.	5.28	118.	0.802	0.358	1.68	1.37	0.41	0.35	0.76	0.90	0.55
CESM1.2	[-]	112.	107.	5.00	118.	0.802	-0.0501	1.65	1.46	0.40	0.36	0.76	0.94	0.56
CESM2(227)	[-]	107.	103.	4.81	118.	0.774	-0.157	1.71	1.48	0.42	0.36	0.79	0.93	0.57

Converting Errors to Scores

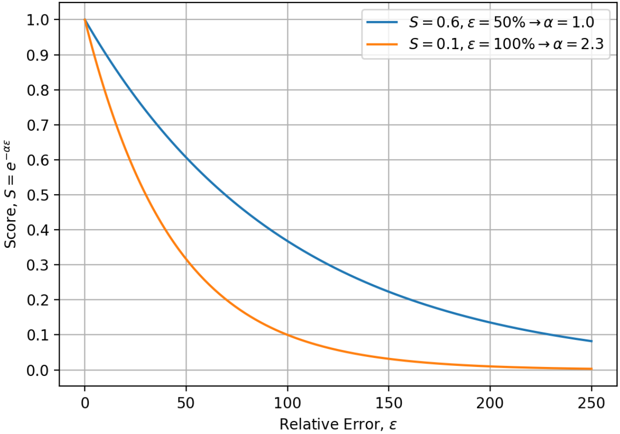
We map a measure of relative error ε to a score by,

$$S = e^{-\alpha\varepsilon}$$

where α can be chosen such that a particular error equates to a given score,

$$\begin{aligned} \hat{S} &= e^{-\alpha\hat{\varepsilon}} \\ \ln(\hat{S}) &= -\alpha\hat{\varepsilon} \\ \alpha &= -\frac{\ln(\hat{S})}{\hat{\varepsilon}} \end{aligned}$$

Converting Errors to Scores



Bias Score

We compute a relative error by normalizing the bias,

$$\varepsilon_{\text{bias}}(\mathbf{x}) = |\text{bias}(\mathbf{x})| / \text{crms}(\mathbf{x})$$

where

$$\text{crms}(\mathbf{x}) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} \left(v_{\text{ref}}(t, \mathbf{x}) - \overline{v_{\text{ref}}}(\mathbf{x})\right)^2 \, dt}$$

Then the score of the bias is

$$s_{\text{bias}}(\mathbf{x}) = e^{-\varepsilon_{\text{bias}}(\mathbf{x})}$$

RMSE Score

We compute a relative error by normalizing the centralized RMSE,

$$\text{crmse}(\mathbf{x}) = \sqrt{\frac{1}{t_f - t_0}\int_{t_0}^{t_f}\left(\left(v_{\text{com}}(t,\mathbf{x}) - \overline{v_{\text{com}}}(\mathbf{x})\right) - \left(v_{\text{ref}}(t,\mathbf{x}) - \overline{v_{\text{ref}}}(\mathbf{x})\right)\right)^2\, dt$$

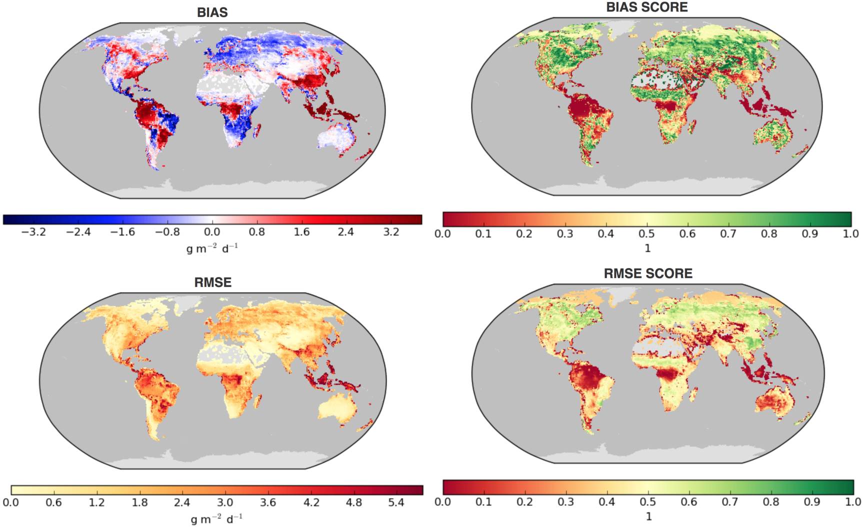
again by the centralized RMS of the reference dataset

$$\varepsilon_{\text{rmse}}(\mathbf{x}) = \text{crmse}(\mathbf{x})/\text{crms}(\mathbf{x})$$

which leads to a score

$$s_{\text{rmse}}(\mathbf{x}) = e^{-\varepsilon_{\text{rmse}}(\mathbf{x})}$$

Errors and Scores

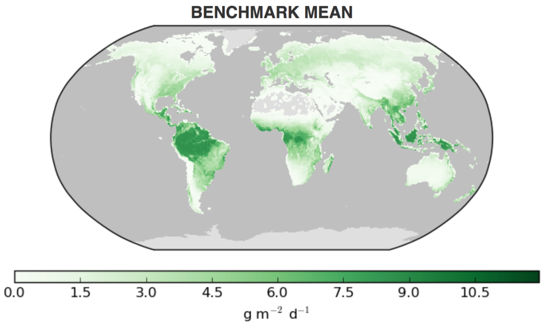


Mass weighting for a scalar score

When computing a score for variables which represent a mass (like carbon or water),

$$S_{\rm rmse} = \frac{1}{\int_\Omega w(\mathbf{x})\ d\Omega} \int_\Omega s_{\rm rmse}(\mathbf{x}) w(\mathbf{x})\ d\Omega$$

where $w(\mathbf{x}) = \overline{v_{\rm ref}}(\mathbf{x})$.



Remarks

- You can compute many things, but it is challenging to come up with a general methodology which can be broadly applied.
- The quality of the conclusions you can draw from ILAMB depends on the dirty details of the methodology we employ.
- We encourage the use of the ILAMB framework. It is more than a flexible software framework, it encapsulates the collective wisdom of the community.
- We encourage community involvement (development is open, regular conference calls to discuss issues). The ILAMB methodology benefits heavily from close interactions with NCAR. The more critical eyes we have on results, the more useful we can make the product.

ILAMB Software

- Open source git repository

  `https://bitbucket.org/ncollier/ilamb`

- CLM (4/4.5/5)

  `http://ilamb.ornl.gov/CLM/`

- CMIP5

  `http://ilamb.ornl.gov/CMIP5/`

- IOMB (Ocean benchmarking)

  `http://ilamb.ornl.gov/IOMB/`