Scoring Methods in the International Land Benchmarking (ILAMB) Package

Nathan Collier[1], Forrest M. Hoffman[1], Gretchen Keppel-Aleks[2], Dave Lawrence[3], Mingquan Mu[4], William J. Riley[5], James T. Randerson[4],

[1]Oak Ridge National Laboratory, [2]University of Michigan, [3]National Center for Atmospheric Research
[4]University of California, Irvine, [5]Lawrence Berkley National Laboratory

The International Land Model Benchmarking (ILAMB) project is a model-data intercomparison and integration project designed to improve the performance of the land component of Earth system models. ILAMB is more than a workflow system that automates the generation of common scalars and plot comparisons to observational data. We aim to provide scientists and model developers with a tool to gain insight into model behavior. Thus, a salient feature of the ILAMB package is our synthesis methodology, which provides users with a high-level understanding of model performance.

Fundamental Difficulty

There are a few difficulties in comparing models to observational datasets.

- The observational datasets and models are not discretized on the same spatial grid. This requires some form of interpolation to make the data comparable which has an effect on the scores and integrated values.
- Even once interpolated to the same grid, the observational datasets and models all define land in different ways. Comparisons can only be done on shared areas which are particular to the variables being compared.

For example, consider the following plots of gross primary productivity (gpp), plotted over portions of Central and South America for emphasis. We have included mean gpp values in the plots of Figure 1. From these plots it is easy to appreciate the disparity in resolutions and definitions of land. We handle these differences by composing the breaks of each grid into a single composite. Consider two spatial grids (for example, from an observational dataset and a model result) whose cells are defined by the outer product of one-dimensional vectors representing the cell breaks in spherical coordinates,

$$\mathcal{G}_{\text{obs}} := \theta_{\text{obs}} \otimes \varphi_{\text{obs}}$$
$$\mathcal{G}_{\text{mod}} := \theta_{\text{mod}} \otimes \varphi_{\text{mod}}$$

where θ refers to the latitude and φ to longitude. Then we may define a composite grid which consists of the outer product of the union of these two grids' cell breaks,

$$\mathcal{G}_c := (\theta_{\text{obs}} \cup \theta_{\text{mod}}) \otimes (\varphi_{\text{obs}} \cup \varphi_{\text{mod}}) \, .$$

Once constructed, quantities defined on both $\mathcal{G}_{\text{obs}}$ and $\mathcal{G}_{\text{mod}}$ may be interpolated to $\mathcal{G}_c$ by nearest neighbor interpolation with zero interpolation error as shown in the left panel of Figure 2. Once on a equivalent grid, this allows us to represent the land areas of each source in a common grid and make comparisons. We will use $\mathcal{L}$ to denote the set of cells in a grid $\mathcal{G}$ designated as land. We report integrated values on various sets of overlapping areas: the intersection of the two grids $\mathcal{L}_{\text{obs}} \cap \mathcal{L}_{\text{mod}}$ and each complement, $\mathcal{L}_{\text{obs}} \setminus \mathcal{L}_{\text{mod}}$ and $\mathcal{L}_{\text{mod}} \setminus \mathcal{L}_{\text{obs}}$. All scores and other integrated quantities are computed on the intersection of the grids' land definitions.

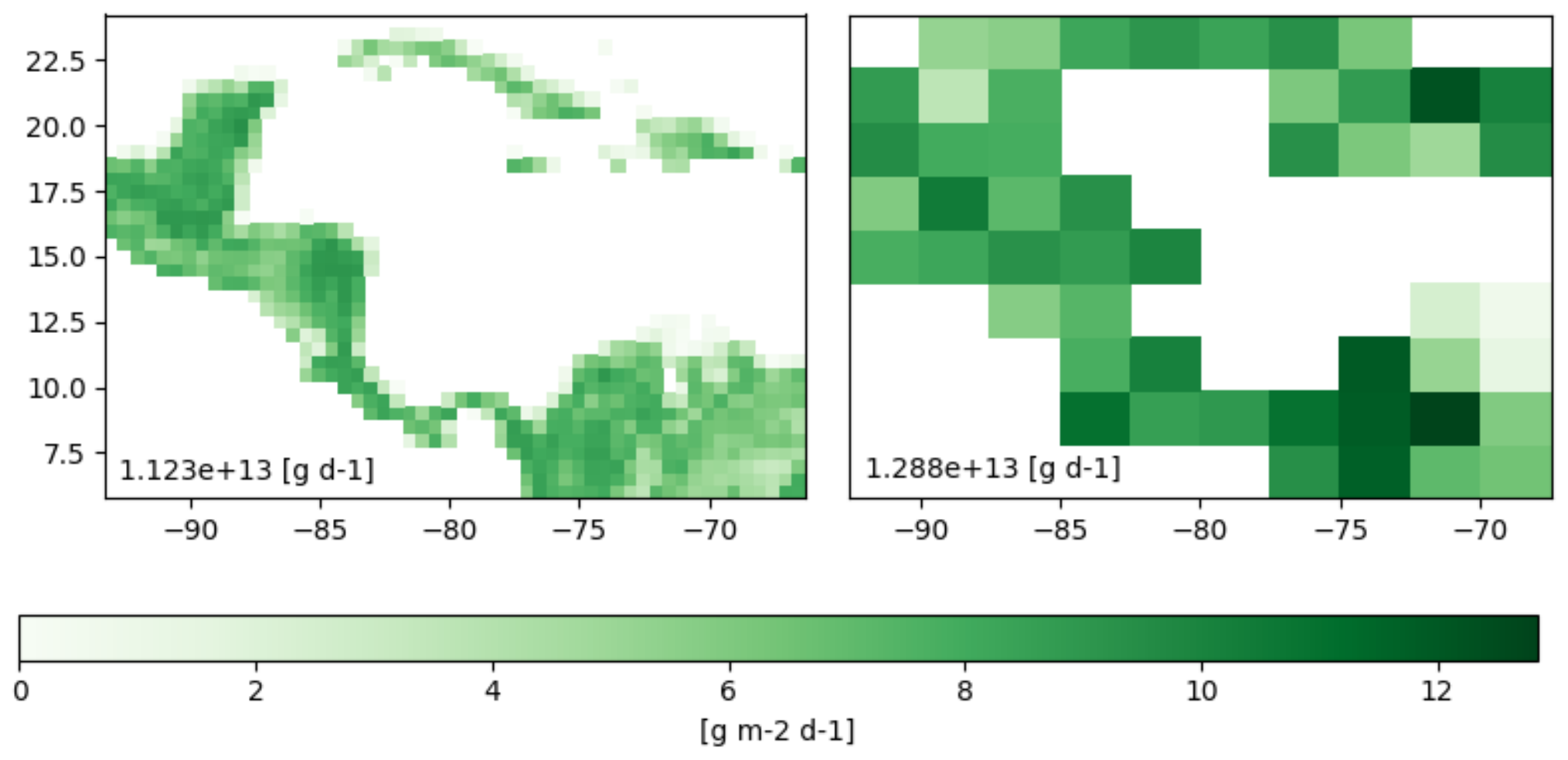

Figure 1: *Gross primary productivity values from the Fluxnet-MTE data product (left) and a model, CLM45 (right).*

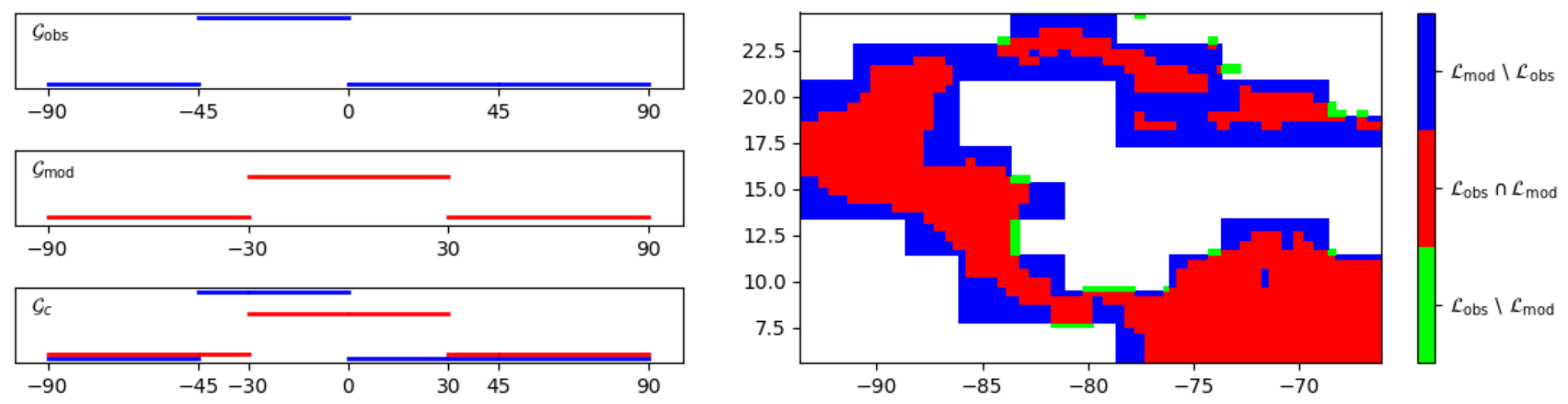

Figure 2: *(left) Step functions on two grids interpolated to a composite grid with zero interpolation error. (right) Differences in representation of land among data sources.*

Converting Errors to Scores

In ILAMB, we map measures of relative error ε to a score using a exponential mapping,

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

where S is a score on the interval $[0,1]$ and α is a parameter which can be used to tuned the mapping of error to score. While we currently take $\alpha = 1$, other values of α may be taken to assign meaning to the scores. If you want a relative error of $\hat{\varepsilon}$ to equate to a score of $\hat{S}$, then

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

Mean State Scores

Within ILAMB, we calculate a non-dimensional score of model performance in a given dimension of the physics, chemistry, or biology with respect to an observational dataset. The following table lists a number of scalars and scores which we use to gauge performance with respect to a given benchmark dataset.

	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												
CLM40	128	123	5.06	118	0.802	0.291	1.60	1.34	0.70	0.62	0.77	0.88	0.72
CLM45	109	104	5.33	118	0.802	-0.103	1.48	1.33	0.74	0.66	0.77	0.90	0.74
CLM50	114	109	5.30	118	0.802	-0.00890	1.60	1.35	0.75	0.66	0.78	0.89	0.75

Table 1: *Sample table found on ILAMB dataset pages, in this case comparing versions of CLM to the Fluxnet-MTE gpp product. We show mean values over the time period, integrated over the different areas as well as other important scalars and scores.*

Period Mean The mean value we compute is an integral over space and time,

$$\text{mean} = \int_\Omega \left(\frac{1}{t_f - t_0} \int_{t_0}^{t_f} v(t, \mathbf{x}) \; dt \right) \; d\Omega$$

where $[t_0, t_f]$ is the time interval and Ω is the spatial domain. We integrate both the observations and model over the intersection and complement of land representations for complete accounting of the variable.

Bias We compute the bias at each spatial point $\mathbf{x}$ as

$$\text{bias}(\mathbf{x}) = \frac{1}{t_f - t_0} \int_{t_0}^{t_f} \text{mod}(t, \mathbf{x}) - \text{obs}(t, \mathbf{x}) \; dt$$

where obs represents the observational or reference dataset, and mod represents the model or comparison dataset. The relative error in the bias is then computed with the following normalization.

$$\varepsilon_{\text{bias}}(\mathbf{x}) = \frac{|\text{bias}(\mathbf{x})|}{\text{bias}(\mathbf{x}) - \min(\text{bias}(\mathbf{x})) + 1 \times 10^{-12}}$$

The score is then computed using

$$S_{\text{bias}} = \int_\Omega e^{-\varepsilon_{\text{bias}}(\mathbf{x})} \; d\Omega$$

RMSE We compute the RMSE at each spatial point $\mathbf{x}$ as

$$\text{RMSE}(\mathbf{x}) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} (\text{mod}(t, \mathbf{x}) - \text{obs}(t, \mathbf{x}))^2 \; dt}$$

The relative error in the RMSE is then computed with the following normalization.

$$\varepsilon_{\text{RMSE}}(\mathbf{x}) = \frac{\text{RMSE}(\mathbf{x})}{\text{RMS}(\mathbf{x})}$$

where the RMS is given as

$$\text{RMS}(\mathbf{x}) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} (\text{obs}(t, \mathbf{x}))^2 \; dt}$$

The score is then computed using

$$S_{\text{RMSE}} = \int_\Omega e^{-\varepsilon_{\text{RMSE}}(\mathbf{x})} \; d\Omega$$

Seasonal Cycle We compute a score for the seasonal cycle by first computing a mean annual cycle across the comparison time period and finding the difference in the annual timing of the maximum value, represented by $\theta(\mathbf{x})$ in terms of days. The a score can be computed using a cosine function,

$$S_{\text{cycle}}(\mathbf{x}) = \frac{1}{2} \left(1 + \cos \left(\frac{2\pi\theta(\mathbf{x})}{365} \right) \right)$$

Spatial Distribution We score the spatial distribution of the time averaged obs and mod by computing the normalized standard deviation,

$$\sigma = \frac{stdev\,(\text{mod})}{stdev\,(\text{obs})}$$

and the correlation R, and then assigning a score by the following relationship

$$S_{\text{dist}} = \frac{2(1+R)}{(\sigma + \frac{1}{\sigma})^2}$$

where the main idea is that we penalize when R and σ deviate from a value of 1.

Overall Score The overall score is then a weighted blend of all these scores,

$$S_{\text{overall}} = \frac{S_{\text{bias}} + 2S_{\text{RMSE}} + S_{\text{cycle}} + S_{\text{dist}}}{1 + 2 + 1 + 1}$$

where the RMSE score is doubly weighted to emphasize its importance.

Relationship Scores

As many models are calibrated using these scalar measures with respect to observational datasets, we also score the relationships among relevant variables in the model. For example, in the case of GPP, we also consider its relationship to precipitation, evapotranspiration, and temperature. We do this by creating a two-dimensional distribution based on the observational data and model results (left two panels of Figure 3) as well as a mean response curve (right panel).

Hellinger Distance The distributions are scored using the so-called Hellinger distance. If the observational distribution is given as $P = (p_1, ..., p_k)$ and the model is given as $Q = (q_1, ..., q_k)$, then

$$S_{\mathsf{H}}(P, Q) = 1 - \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}$$

RMSE Score The response curves are then scored using a relative measure of the root mean squared error and the exponential as before. For an observational curve $p(x)$ and a model curve $q(x)$, then

$$S_{\mathsf{RMSE}}(p, q) = e^{-\sqrt{\frac{\int (p(x) - q(x))^2 \; dx}{\int p(x)^2 \; dx}}}$$

where as before we have used the exponential to map the relative error in the RMSE.

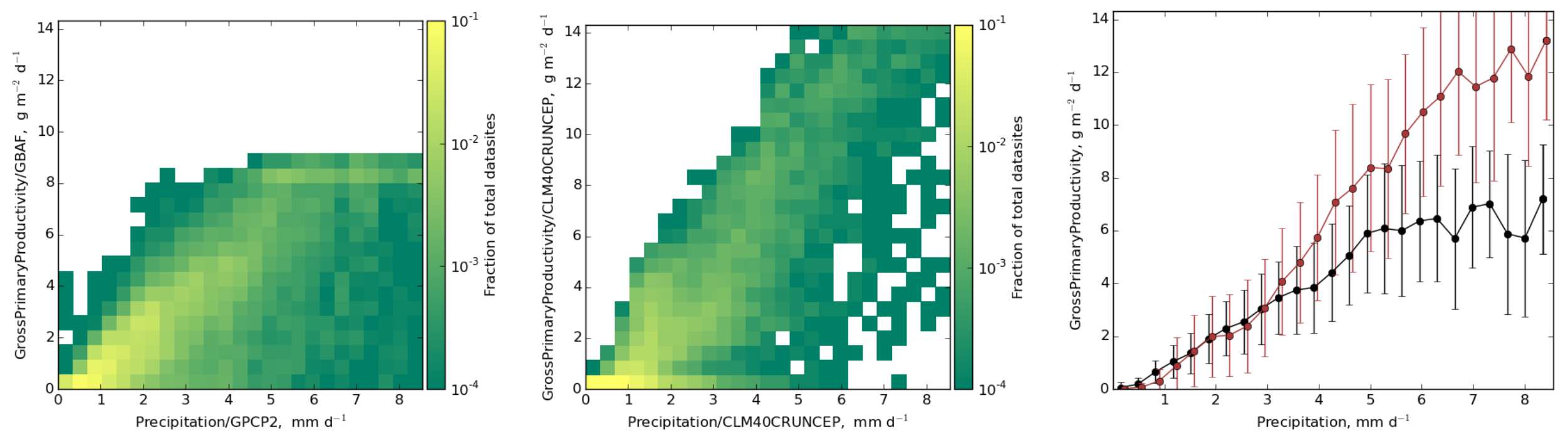

Figure 3: *(left) Observational dataset 2D distribution P, (middle) Model 2D distribution Q, (right) Observational and model functional relationship p and q.*

Overall Score The overall score is then a weighted blend of all these scores,

$$S_{\text{overall}} = \frac{1}{2} (S_{\mathsf{H}} + S_{\text{RMSE}})$$

Summary

The overall scores computed are then combined to form an overall assement of how well a model performs with respect to a given variable. The ILAMB system then makes a plot as shown in Figure 4. On the left side of the plot we show the model's overall score in a particular variable. However, as these scores tend to be close together, we also provide the right panel which shows a relative assessment among the models being compared.

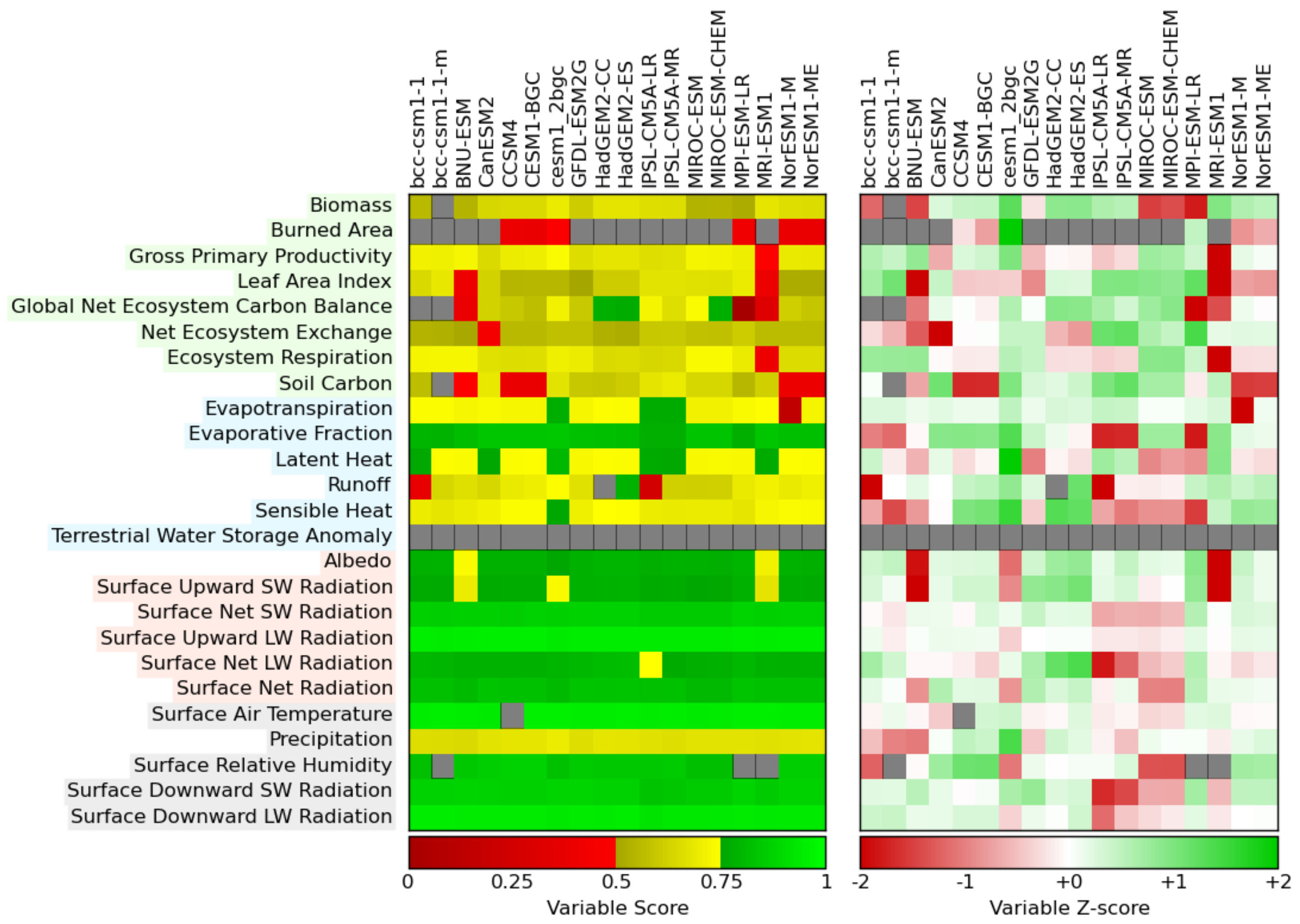

Figure 4: *(left) the absolute overall score for the given model and variable (right) the relative performance for the given variable across models.*

Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC005-00OR22725.