AGU100 ADVANCING EARTH AND SPACE SCIENCE

Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE

10.1029/2018MS001354

The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation

Nathan Collier[1], **Forrest M. Hoffman**[1,2], **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]Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, USA, [2]Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, Knoxville, TN, USA, [3]Climate and Global Dynamics Division, National Center for Atmospheric Research, Boulder, CO, USA, [4]Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI, USA, [5]Climate Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, [6]Department of Earth System Science, University of California, Irvine, CA, USA

Key Point:

- The ILAMB benchmarking system broadly compares models to observational data sets and provides a synthesis of overall performance

Correspondence to:
N. Collier,
nathaniel.collier@gmail.com

Citation:
Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., et al. (2018). The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, *10*, 2731–2754. https://doi.org/10.1029/2018MS001354

Received 24 APR 2018
Accepted 10 OCT 2018
Accepted article online 12 OCT 2018
Published online 7 NOV 2018

Abstract The increasing complexity of Earth system models has inspired efforts to quantitatively assess model fidelity through rigorous comparison with best available measurements and observational data products. Earth system models exhibit a high degree of spread in predictions of land biogeochemistry, biogeophysics, and hydrology, which are sensitive to forcing from other model components. Based on insights from prior land model evaluation studies and community workshops, the authors developed an open source model benchmarking software package that generates graphical diagnostics and scores model performance in support of the International Land Model Benchmarking (ILAMB) project. Employing a suite of in situ, remote sensing, and reanalysis data sets, the ILAMB package performs comprehensive model assessment across a wide range of land variables and generates a hierarchical set of web pages containing statistical analyses and figures designed to provide the user insights into strengths and weaknesses of multiple models or model versions. Described here is the benchmarking philosophy and mathematical methodology embodied in the most recent implementation of the ILAMB package. Comparison methods unique to a few specific data sets are presented, and guidelines for configuring an ILAMB analysis and interpreting resulting model performance scores are discussed. ILAMB is being adopted by modeling teams and centers during model development and for model intercomparison projects, and community engagement is sought for extending evaluation metrics and adding new observational data sets to the benchmarking framework.

1. Introduction

As Earth system models (ESMs) become increasingly complex and observational data volumes rapidly expand, there is a growing need for comprehensive and multifaceted evaluation of model fidelity. Process-rich ESMs pose challenges to developers implementing new parameterizations or tuning process representations, and to the broader community seeking information about the skill of model predictions. Model developers and software engineers require a systematic means for evaluating changes in model results to ensure that developments improve the scientific performance of target process representations while not adversely affecting results in other, possibly less familiar, parts of the model. To advance understanding and predictability of terrestrial biogeochemical processes and their interactions with hydrology and climate under conditions of increasing atmospheric carbon dioxide, rigorous analysis methods, employing best available observational data, are required to objectively assess and constrain model predictions, inform model development, and identify needed measurements and field experiments (Hoffman et al., 2017).

Building upon past model evaluation work (Randerson et al., 2009), we developed an extensible model benchmarking package in support of the goals of the International Land Model Benchmarking (ILAMB; https://www.ilamb.org/) activity. ILAMB's goals are to

1. develop internationally accepted benchmarks for land model performance by drawing upon international expertise and collaboration;
2. promote the use of these benchmarks by the international community for model intercomparison and development;

Table 1
The ILAMB Rubric Used to Assign Relative Weights of a Data Set

Score	Certainty	Scale	Process
1	No given 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 given 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 given 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

Note. A score for each data set is assigned in each of three areas. These scores are then combined multiplicatively and used to determine relative importance for a data set with respect to a given variable. ILAMB = International Land Model Benchmarking.

3. strengthen linkages among experimental, remote sensing, and climate modeling communities in the design of new model tests, benchmarks, and measurement programs; and
4. support the design and development of a new, open source, benchmarking software system for use by the international community.

Three ILAMB workshops have been held—in Exeter, UK, in 2009; Irvine, California, USA, in 2011 (Luo et al., 2012); and Washington, DC, USA, in 2016 (Hoffman et al., 2017)—to engage the modeling, measurements, and remote sensing communities in the identification of observational data sets and the design of model evaluation metrics. In this way, community consensus was sought for the curation of observational data and the methodology of model evaluation and scoring, which are described below.

Recognition that the capacities of the terrestrial and marine biosphere to store anthropogenic carbon will weaken under climate warming (Cox et al., 2000; Denman et al., 2007; Friedlingstein et al., 2001; Fung et al., 2005; Mahowald et al., 2017; Moore et al., 2018; Randerson et al., 2015) and that uncertainties in carbon cycle feedbacks must be quantified and reduced to improve projections of future climate change (Arora et al., 2013; Ciais et al., 2013; Friedlingstein et al., 2006, 2014; Gregory et al., 2009; Hoffman et al., 2014) has inspired efforts to quantitatively evaluate model performance through comparison with in situ and remote sensing observations (Anav et al., 2013; Eyring et al., 2016). Multimodel simulation results from the third Coupled Model Intercomparison Project (CMIP3; Meehl et al., 2007) and fifth CMIP (CMIP5; Taylor et al., 2012), which informed the Intergovernmental Panel on Climate Change Fourth and Fifth Assessment Reports (AR4 and AR5), provided opportunities for developing and testing model evaluation diagnostics, formal metrics, and exploration of benchmarking concepts and techniques. Early work on coupled model evaluation and establishing formal metrics focused primarily on atmospheric variables (Gleckler et al., 2008; Reichler & Kim, 2008). Following the first two ILAMB workshops, the land modeling community began exploring standardized and comprehensive benchmarking for terrestrial carbon cycle models (Abramowitz, 2012; Anav et al., 2013; Blyth et al., 2011; Bouskill et al., 2014; Cadule et al., 2010; Dalmonech & Zaehle, 2013; Ghimire et al., 2016; Kelley et al., 2013; Piao et al., 2013). While some researchers define benchmarking as a series of model tests based on a predefined expected level of performance (Abramowitz, 2005; Best et al., 2015), most of the systematic benchmarking strategies explored by the land modeling community to date do not depend upon the establishment of an expected level of performance.

The ILAMB software package, hereafter referred to as ILAMB, shares some of the same goals as existing model diagnostic and evaluation tools, such as the Protocol for the Analysis for Land Surface models (Abramowitz, 2012), the Program for Climate Model Diagnosis and Intercomparison Metrics Package (Gleckler et al., 2016),

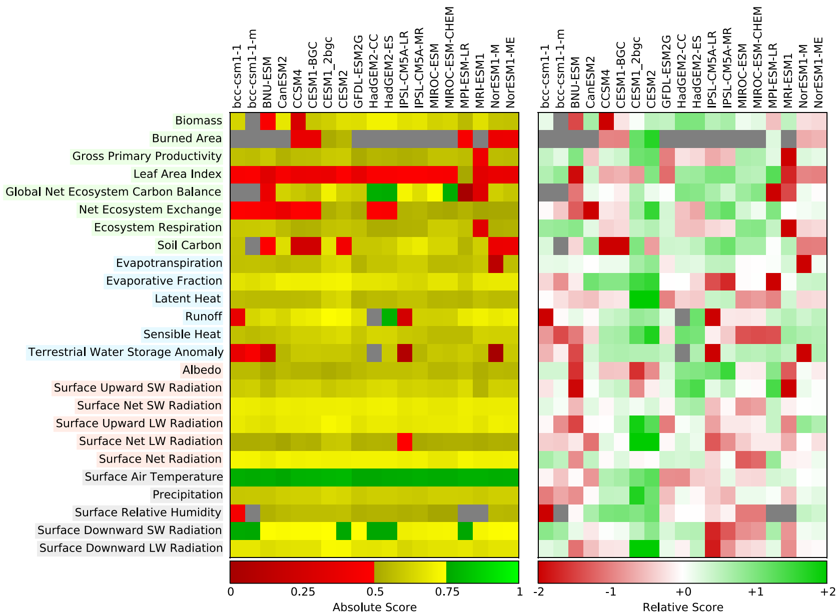


Figure 1. The International Land Model Benchmarking top-level graphic uses stoplight colors to show how different models or model versions (across the top) score with respect to each variable (down the left) in an absolute sense (left rectangle) and with respect to each other (right rectangle). Gray boxes reflect missing or unavailable data.

the ESM Evaluation Tool (Eyring et al., 2016), the Land surface Verification Toolkit (Kumar et al., 2012), and a wide variety of often custom-developed diagnostic packages in use at international modeling centers. Some of these tools provide model-to-model comparisons, a large collection of stand-alone graphical diagnostics, or workflow infrastructure that allows one to regenerate analysis results from previously published studies but with new model outputs. In contrast, ILAMB was designed to compare multiple models or model versions with observations simultaneously, assess functional relationships between prognostic variables and one or more forcing variables through variable-to-variable comparisons (e.g., gross primary production vs. precipitation), and score model performance across a suite of metrics, variables, and data sets. Model performance is evaluated for variables in categories of biogeochemistry (Table 2), hydrology (Table 3), radiation and energy (Table 4), and climate forcing (Table 5).

For every variable, ILAMB generates graphical diagnostics (spatial contour maps, time series line plots, and Taylor diagrams; Taylor, 2001) and scores model performance for the period mean, bias, root-mean-square error (RMSE), spatial distribution, interannual coefficient of variation, seasonal cycle, and long-term trend. Model performance scores are calculated for each metric and variable and are scaled based on the degree of certainty of the observational data set, the scale appropriateness, and the overall importance of the constraint or process to model predictions, following a customizable rubric described below (Table 1). Scores are aggregated across metrics and data sets, producing a single scalar score for each variable for every model or model version. As shown in Figure 1, these scalar scores are presented graphically. On the left side we use a stoplight color scheme to indicate aggregate performance for each model by variable. On the right, we show relative performance (i.e., *Z* score), indicating which models or model versions perform better with respect to others contained in the overall analysis.

We do not view these aggregate absolute scores as a determinant of *good* or *bad* models. We envision the scores as a tool to more quickly identify relative differences among models and model versions which the scientist must then interpret. As in any evaluation methodology, many of our choices are subjective and must be considered as the scores are interpreted. Where possible, the ILAMB implementation allows for users to customize weights and diagnostics in order to incorporate aspects of model performance relevant to their scientific goals. ILAMB may be thought of as a framework which may be expanded to incorporate community ideas regarding model benchmarking. Thus, while our choices are subjective, they are informed by the preferences of a larger community and can be considered as an initial suggestion.

The remainder of this paper describes the ILAMB methodology used to compute aggregate absolute scores. First we describe how we compare an individual observational data set to model output (section 2). Then we explain how scores are aggregated across data sets for each variable and present the data sets used in the land model evaluation (section 3). In section 4 we present some salient points about how the ILAMB software is designed. Finally, in section 5 we discuss what ILAMB scores mean and how they should be used.

2. Methodology

In this section we describe the methodology used to assess how well a model captures information contained in a reference (e.g., observational) data set. For the purposes of this section, we discuss the analysis of a generalized variable $v(t, \mathbf{x})$, which we assume represents a piecewise discontinuous function of constants in space and time. This means that the temporal domain, represented by the variable t, is defined by the beginning and ending of time intervals and the spatial domain, represented by the variable $\mathbf{x}$ (in bold to emphasize it is a vector quantity), represents the areas created by cell boundaries or the areas associated with data sites. When necessary, we use the subscript *ref* to reflect a variable whose source is a reference or observational data set, and the subscript *mod* for model data sets.

While many statistical quantities may be computed, the goal of our initial methodology is to examine the mean state and variability around the mean over monthly to decadal time scales and grid cell to global spatial scales. While we intend to uniformly apply this analysis procedure to all variables, we also implement a mechanism to skip certain aspects when deemed inappropriate. For example, if a reference data set only contains average information across a span of years, the annual cycle is undefined and automatically skipped in our implementation. The implementation also allows users to skip aspects of the analysis that are deemed inappropriate even if it is possible to compute these metrics using the available data. For example, the interannual variability may be poorly characterized in a reference data set even though the quantity could be computed.

2.1. Preliminary Definitions

Before presenting the specifics of the ILAMB methodology, we first present some definitions used throughout the paper. While the following definitions are widely used in the community, there are many subtle choices in their implementation that affect the interpretation of the results. We present them here with precise meanings to emphasize where a choice has been made and our reasoning for making it.

2.1.1. Mean Values Over Time

When calculating mean values over the time period of the benchmark data set, denoted by a bar superscribing the variable, we use the midpoint quadrature rule to approximate the integral,

$$\begin{aligned}\bar{v}(\mathbf{x}) &= \frac{1}{t_f - t_0}\int_{t_0}^{t_f} v(t, \mathbf{x})\,\mathrm{d}t \\ &\approx \frac{1}{T(\mathbf{x})}\sum_{i=1}^{n} v(t_i, \mathbf{x})\Delta t_i\end{aligned} \tag{1}$$

where n represents the number of time intervals on which v is defined between the initial time, t_0, and the final time, t_f, and Δt_i is the size of the i^{th} time interval, modified to exclude time, which falls outside of the integral limits,

$$\Delta t_i = \min(t_f, t_f^i) - \max(t_0, t_0^i) \tag{2}$$

where t_0^i and t_f^i are the initial and final times of each time interval. The average value is obtained by dividing through by the amount of time in the interval, $t_f - t_0$, replaced in our discrete approximation by the following function.

$$T(\mathbf{x}) = \sum_{i=1}^{n} \Delta t_i \text{ if } v(t_i, \mathbf{x}) \text{ is valid} \tag{3}$$

In words, equation (3) addresses temporally discontinuous data by summing all the time step interval sizes only if the corresponding variable data are marked as valid. This means that if a function has some values masked or marked as invalid at some locations, we do not penalize the averaged value by including this as a time at which a value is expected. If an integral (or sum) is desired instead of an average, then we simply omit the division by $T(\mathbf{x})$ in equation (1).

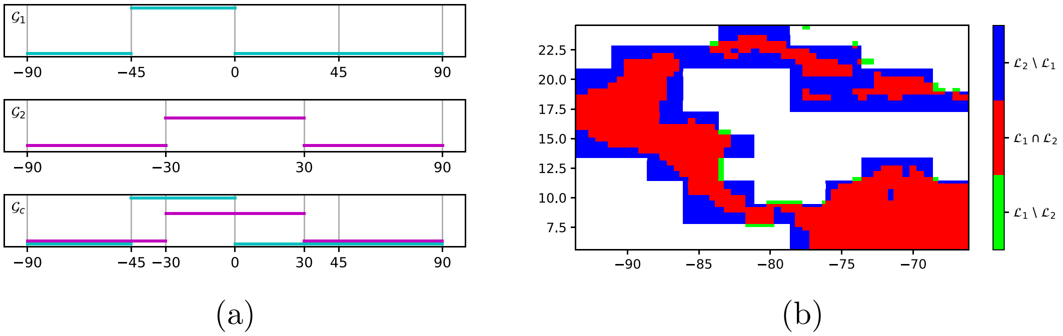


Figure 2. When comparing two spatial variables of varying resolution, we interpolate both to a common grid composed of the cell breaks of both variables over the intersection of what both variables agree is land. (a) Interpolation of sample step functions defined on grids $\mathcal{G}_1$ and $\mathcal{G}_2$ both interpolated to a composite grid $\mathcal{G}_c$ using nearest neighbor interpolation with zero interpolation error. The vertical grid lines reflect the cell boundaries in each grid. (b) Differences in the representation of land from a reference and model data set zoomed into Central America for emphasis. The red region represents where both sources are in agreement, the blue is land for the model but not the reference, and the green is land for the reference but not the model.

2.1.2. Mean Values Over Space

When computing spatial means over various regions of interest, denoted by a double bar over a variable, we use the midpoint rule for integration to approximate the following weighted spatial integral,

$$\begin{aligned}\bar{\bar{v}}(t) &= \frac{1}{\int_\Omega w(\mathbf{x})\, d\Omega} \int_\Omega v(t,\mathbf{x}) w(\mathbf{x})\, d\Omega \\ &\approx \frac{1}{A(\Omega)} \sum_{i=1}^{n(\Omega)} v(t,\mathbf{x}_i) w(\mathbf{x}_i) a_i\end{aligned} \tag{4}$$

over a region Ω, also referred to as a area-weighted mean. Here the function $w(\mathbf{x})$ is an optional generic weighting function defined over space. The summation is over $n(\Omega)$, that is, the integer number of spatial cells whose centroids fall into the region of interest. A function evaluation at a location $\mathbf{x}_i$ refers to the constant value which corresponds to that spatial cell. The value of a_i is the area of the cell, which could be some fraction of the total cell area if integrating over land in coastal regions. We then divide through by the measure, the sum of the grid areas with the weights,

$$A(\Omega) = \sum_{i=1}^{n(\Omega)} w(\mathbf{x}_i) a_i \text{ if } v(t,\mathbf{x}_i) \text{ is valid} \tag{5}$$

Note that if no weighting is required, this is a normalization by the sum of the area over which we integrate. As with the temporal mean, if an integral only is required, we simply omit the division by $A(\Omega)$. In cases where a mean over a collection of sites is needed, the spatial integral reduces to an arithmetic mean across the sites.

If we are spatially integrating a variable from a single source, then its spatial grid is clearly defined and equation (4) can be directly applied to compute the quantity of interest. However, if the integrand involves quantities from two different sources, as in computing the global bias or RMSE, then there is likely a disparity in both resolution and representation of land areas. We address resolution differences by interpolating both sources to a grid composed of the cell breaks, the location at which two neighboring cells meet, of both data sources. Consider two spatial grids whose cells are defined by the outer product of 1-D vectors representing the cell breaks in spherical coordinates,

$$\mathcal{G}_1 := \theta_1 \otimes \varphi_1 \tag{6}$$

$$\mathcal{G}_2 := \theta_2 \otimes \varphi_2 \tag{7}$$

where θ refers to the latitude, φ to longitude, and $\otimes$ a operator which creates a two-dimensional grid from one-dimensional vectors. We address differences in resolution by defining a composite grid, which consists of the outer product of the union of these two grids' cell breaks,

$$\mathcal{G}_c := \left(\theta_1 \cup \theta_2\right) \otimes \left(\varphi_1 \cup \varphi_2\right). \tag{8}$$

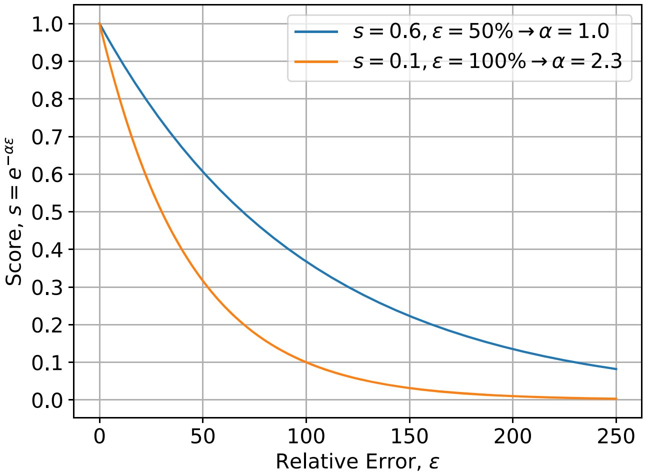


Figure 3. Mapping function of relative error ε to a score s on the unit interval. Two choices of α are shown: $\alpha = 1$, shown in blue, which equates a score of 0.6 to a relative error of 50%; and $\alpha = 2.3$, shown in orange, which equates a score of 0.1 to a relative error of 100%.

Once constructed, quantities defined on both $\mathcal{G}_1$ and $\mathcal{G}_2$ may be interpolated to $\mathcal{G}_c$ by nearest neighbor interpolation with zero interpolation error due to the nested nature of the grids. This can be seen visually by comparing the three plots shown in Figure 2a. In each plot, the tick marks along the x axis represent the cell breaks of the particular one-dimensional grid left coarse for illustration. The cyan curve represents a step function defined on the grid of a reference data set $\mathcal{G}_1$ and the magenta curve on that of the model data set $\mathcal{G}_2$. Both are interpolated to the composed grid $\mathcal{G}_c$ without loss of information, albeit on a new grid containing more cells of variable size. Once on a composite grid, the quantities may be compared directly. As the ILAMB methodology has been envisioned for comparisons with model output from CMIP5, we have made an implicit assumption that each source grid, $\mathcal{G}_1$ and $\mathcal{G}_2$, is regular and can be represented by one-dimensional vectors. While the implementation does provide naive interpolation for nonregular grids, the user is encouraged to employ a conservative interpolation scheme of their choosing prior to applying the ILAMB methodology.

In addition to resolution differences, we observe that data sources vary in the underlying representation of the distinction between land and water. We illustrate this concept in Figure 2b where we compare a fine scale representation of land $\mathcal{L}_1$ to a relatively coarse representation $\mathcal{L}_2$. This is a typical situation encountered when comparing high-resolution observational data to lower-resolution model output. The red region represents the intersection of land areas $\mathcal{L}_1 \cap \mathcal{L}_2$, that is, where both sources report the presence of land. However, there are missed land areas from both sources, represented by the blue and green colors. As much of the disagreement over what is considered land occurs around islands in tropical regions (e.g., Central America and Equatorial Asia), these nonrepresented areas can constitute a nontrivial percentage of the total represented variable v.

For transparency, the ILAMB implementation is built with the capability of reporting integrals over each of these three land areas. Unless specifically stated otherwise, when spatially integrating a quantity from a single source, we use the original grid and land areas given by that source. This is to remain as true to the original intent of the provider as we can. However, when comparing two data sources of varying resolution and land representation, we perform this integration over what both report to be land, $\mathcal{L}_1 \cap \mathcal{L}_2$ (the red area in Figure 2b).

2.1.3. Computing Normalized Scores From Errors

In sections 2.2 and 2.3, we detail how we compute errors and transform them into normalized scores on the unit interval. This approach is intended to synthesize model performance across a range of dimensions with respect to a given data set. We achieve this by taking a measure of the relative error, generically represented here as ε, and passing it through the exponential function,

$$s = e^{-\alpha\varepsilon} \tag{9}$$

where s is a score on the interval [0,1] and α is a parameter which can be used to tune the mapping of error to score. The classic expression of relative error is prone to numerical instabilities for denominator values near or which cross zero. Furthermore, the magnitude of the error can depend on the units selected. For this reason we depart from the standard definition of relative error and develop specialized expressions in equations (13), (18), and (26).

While the choice of the exponential function is arbitrary, it was chosen because it maps zero error to a score of one and smoothly reduces the score as the error grows, never reaching exactly zero. This is important as we want to improve the score when the error improves, no matter how large of error we observe. If the user wants a relative error of $\hat{\varepsilon}$ to equate to a score of $\hat{s}$, then

$$\alpha = -\frac{\ln(\hat{s})}{\hat{\varepsilon}} \tag{10}$$

In Figure 3 we plot this function with two choices for α, which illustrates how the relative error may be controlled. Unless stated otherwise, we use an implicit $\alpha = 1$ throughout the manuscript.

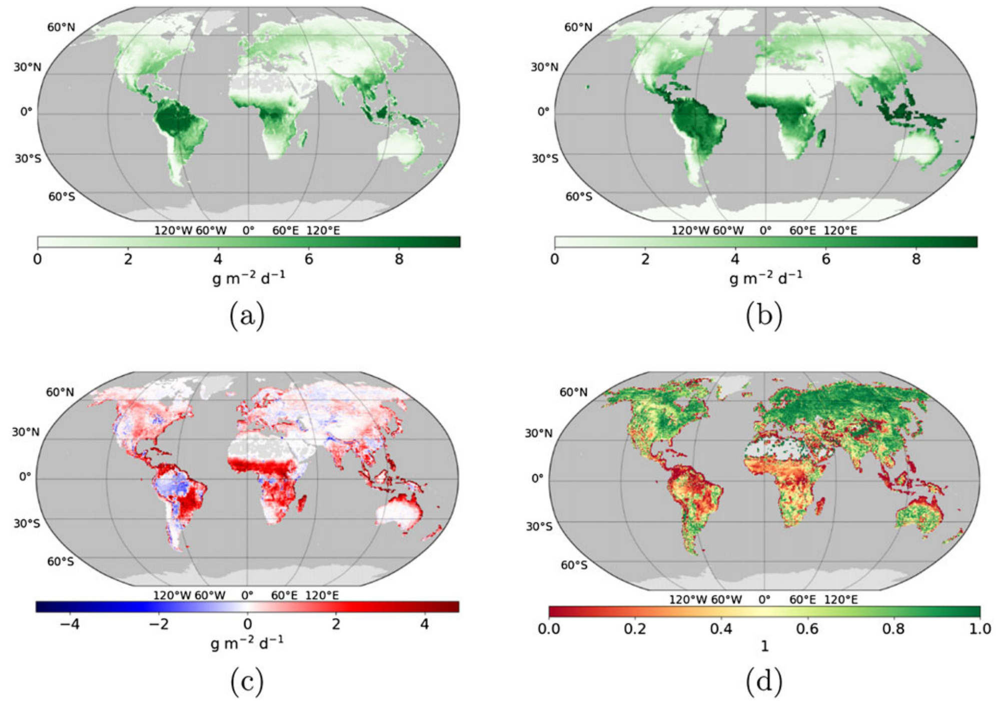


Figure 4. Comparisons of gross primary productivity between the reference (GBAF) and the model (CLM4.5) data sets. Each period mean is plotted over the original grid of the data set. We highlight here that the reference (a) is not defined over Antarctica, Greenland, and part of the Sahara desert, whereas the model (b) is defined over all land areas. Yet when the bias (c) and its score (d) is reported, the area represented is what both the reference and model agree on as land. (a) Reference period mean, $\overline{v_{\mathrm{ref}}}(\mathbf{x})$; (b) model period mean, $\overline{v_{\mathrm{mod}}}(\mathbf{x})$; (c) bias, $bias(\mathbf{x})$; (d) bias score, $s_{\mathrm{bias}}(\mathbf{x})$.

2.2. Mean State Analysis

In this section, we describe the various metrics and plots that our methodology generates. While presented in terms of the abstract variable v, we also include sample plots of a comparison of the GBAF (Jung et al., 2010) gross primary productivity (GPP) with CLM4.5 (Oleson et al., 2013) for the purpose of illustration. In practice, ILAMB produces thousands of such plots and scalars, which are browsable in a website designed to aid modelers in understanding the benchmarking results.

2.2.1. Bias

We find the mean value in time, $\overline{v_{\mathrm{ref}}}(\mathbf{x})$, over the time period of the reference, as well as that of the model, $\overline{v_{\mathrm{mod}}}(\mathbf{x})$, over the same time period. These are spatial variables that are included in the standard output as plots, as shown in Figures 4a and 4b. We also compute the bias,

$$bias(\mathbf{x}) = \overline{v_{\mathrm{mod}}}(\mathbf{x}) - \overline{v_{\mathrm{ref}}}(\mathbf{x}) \tag{11}$$

as well as its mean over a given region, $\overline{\overline{bias}}(\mathbf{x})$. To score the bias, we need to nondimensionalize it as a relative error. We have chosen to do this by using the centralized RMS of the reference data,

$$crms(\mathbf{x}) = \sqrt{\frac{1}{t_f - t_0}\int_{t_0}^{t_f} \left(v_{\mathrm{ref}}(t,\mathbf{x}) - \overline{v_{\mathrm{ref}}}(\mathbf{x})\right)^2 \, \mathrm{d}t}, \tag{12}$$

which makes the relative error in bias given as

$$\varepsilon_{\mathrm{bias}}(\mathbf{x}) = |bias(\mathbf{x})|/crms(\mathbf{x}) \tag{13}$$

where the || operator represents the absolute value. The bias score as a function of space is

$$s_{\mathrm{bias}}(\mathbf{x}) = e^{-\varepsilon_{\mathrm{bias}}(\mathbf{x})} \tag{14}$$

and the scalar score

$$S_{\mathrm{bias}} = \overline{\overline{s_{\mathrm{bias}}}}(\mathbf{x}), \tag{15}$$

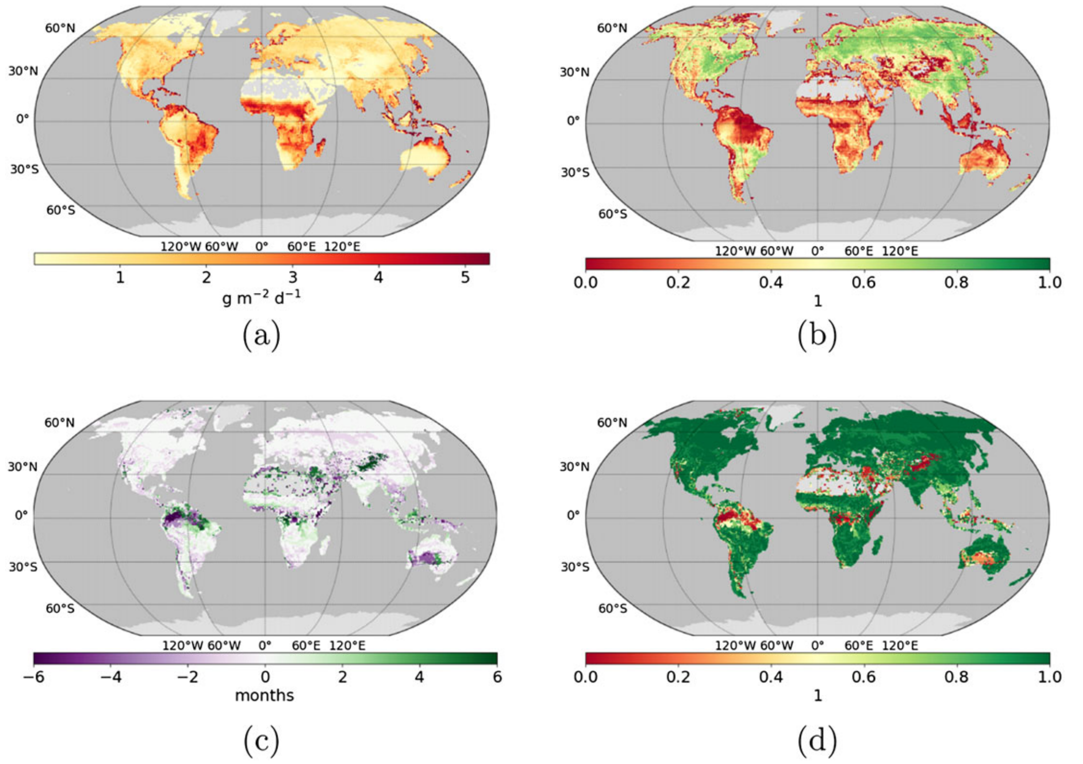


Figure 5. Comparisons of the root-mean-square error (RMSE) and phase of gross primary productivity between the reference (GBAF) and the model (CLM4.5) data sets. (a) RMSE, $rmse(\mathbf{x})$; (b) RMSE score, $s_{\text{rmse}}(\mathbf{x})$; (c) phase shift, $\theta(\mathbf{x})$; (d) phase shift score, $s_{\text{cycle}}(\mathbf{x})$.

that is, the spatially integrated bias score. The motivation behind equation (13) is to normalize the bias by the variability at any given spatial location. However, this also leads to the consequence that in areas where the given variable v has a small magnitude, simple noise can lead to large relative errors. For example, in Figure 4d we observe a poor score in the dry regions of Australia where GPP is small. Given the small contribution, it is undesirable that these errors induce a large negative contribution to the overall score. To address this issue, we introduce the concept of *mass weighting*. That is, when performing the spatial integral to obtain a scalar score (equation (15)), we weigh the integral by the period mean value of the reference variable using equation (4) with $w = \overline{v_{\text{ref}}}$. In some instances the variable is truly a mass, but other times a flux or rate. The main motivation is to weigh in areas where the variable is active. So while in our conceptual example, there is large relative error in GPP over deserts, these values will not negatively contribute to the overall score as the value of GPP is low in this area.

We apply mass weighting when the variable v represents a mass or flux of carbon or water as in GPP or precipitation. For variables representing energy states or quantities, such as temperature and radiation, we omit the weighting and perform a spatial integral only. We report plots of the bias and its score as well as the scalar integrated mean values.

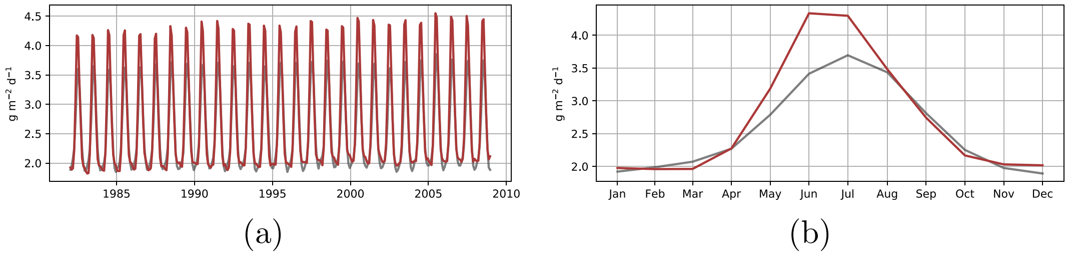


Figure 6. Spatial means of gross primary productivity of the reference (GBAF) shown in gray and the model (CLM4.5) in maroon. (a) Spatially integrated mean, $\overline{v_{\text{ref}}}(t)$ and $\overline{v_{\text{mod}}}(t)$; (b) mean annual cycle, $\overline{v_{\text{ref}}}(t)$ and $\overline{v_{\text{mod}}}(t)$.

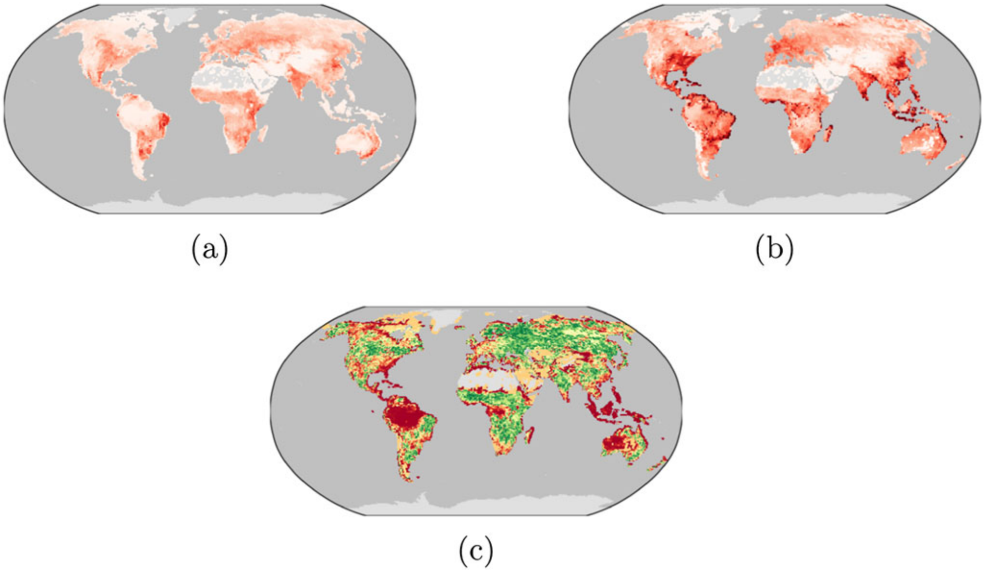


Figure 7. Comparisons of the interannual variability of gross primary productivity between the reference (GBAF) and the model (CLM4.5) data sets. (a) Reference interannual variability, $iav_{\text{ref}}(\mathbf{x})$; (b) model interannual variability, $iav_{\text{mod}}(\mathbf{x})$; (c) interannual variability score, $s_{\text{iav}}(\mathbf{x})$.

2.2.2. RMSE

For reference data sets with seasonal and interannual variability, we compute the RMSE over the time period of the reference data set,

$$rmse(\mathbf{x}) = \sqrt{\frac{1}{t_f - t_0}\int_{t_0}^{t_f}\left(v_{\text{mod}}\,(t,\mathbf{x}) - v_{\text{ref}}\,(t,\mathbf{x})\right)^2\,\mathrm{d}t} \tag{16}$$

and include plots and the scalar $\overline{\overline{rmse(\mathbf{x})}}$ in the standard output (Figure 5a). To score the RMSE, we normalize the centralized RMSE,

$$crmse(\mathbf{x}) = \sqrt{\frac{1}{t_f - t_0}\int_{t_0}^{t_f}\left(\left(v_{\text{mod}}(t,\mathbf{x}) - \overline{v_{\text{mod}}}(\mathbf{x})\right) - \left(v_{\text{ref}}(t,\mathbf{x}) - \overline{v_{\text{ref}}}(\mathbf{x})\right)\right)^2\,\mathrm{d}t} \tag{17}$$

by the centralized RMS of the reference data set, equation (12). This leads to a relative error of

$$\varepsilon_{\text{rmse}}(\mathbf{x}) = crmse(\mathbf{x})/crms(\mathbf{x}) \tag{18}$$

and a spatial RMSE score

$$s_{\text{rmse}}(\mathbf{x}) = e^{-\varepsilon_{\text{rmse}}(\mathbf{x})}. \tag{19}$$

The scalar score is obtained by

$$S_{\text{rmse}} = \overline{\overline{s_{\text{rmse}}}}(\mathbf{x}), \tag{20}$$

where we again employ mass weighting when necessary. We score the centralized RMSE to decouple the bias score from the RMSE score. Computing the RMSE score by normalizing the RMSE would lead to a double counting of errors. That is, a large error in bias also leads to a large error in RMSE. By scoring the centralized RMSE, we remove the bias from the RMSE, allowing the RMSE score to focus on an orthogonal aspect of model performance.

2.2.3. Phase Shift

We evaluate the phase shift of the annual cycle of many data sets that have monthly variability by comparing the timing of the maximum of the annual cycle of the variable, $c(v)$, at each spatial cell across the time period of the reference data set. We then approximate the phase shift of the reference and model data sets by subtracting these two values,

$$\theta(\mathbf{x}) = \arg\max_t(c_{\text{mod}}(t,\mathbf{x})) - \arg\max_t(c_{\text{ref}}(t,\mathbf{x})) \tag{21}$$

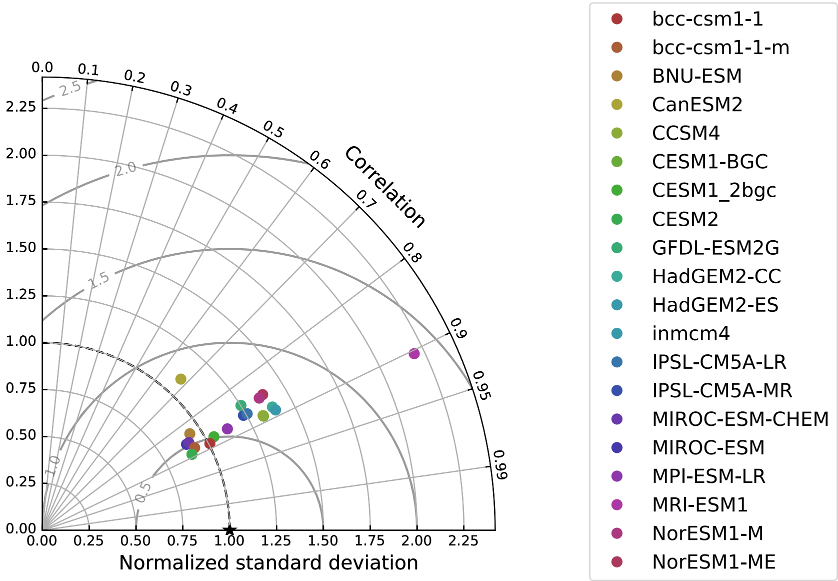


Figure 8. Taylor diagram comparing the spatial distribution of gross primary productivity of the reference (GBAF) shown as a black star to the CMIP5 models shown in colors.

expressed in days. As the units for phase shift are consistent across all variables, no normalization is needed and we can remap the shift to the unit interval by

$$s_{\text{phase}}(\mathbf{x}) = \frac{1}{2}\left(1 + \cos\left(\frac{2\pi\theta(\mathbf{x})}{365}\right)\right) \tag{22}$$

and then spatially integrate the score over the appropriate region to find the scalar score,

$$S_{\text{phase}} = \overline{\overline{s_{\text{phase}}}}(\mathbf{x}), \tag{23}$$

where again mass weighting is employed when appropriate. We include plots of the phase shift and its score in the standard output and represent them here in Figures 5c and 5d. In addition to plots which show the time averaged variables as a map, we include line plots of the mean annual cycle and the spatially averaged variables, $\overline{\overline{v_{\text{ref}}}}(t)$ and $\overline{\overline{v_{\text{mod}}}}(t)$ shown in Figure 6.

2.2.4. Interannual Variability

A score for the interannual variability is computed by removing the annual cycle from both the reference and the model,

$$iav_{\text{ref}}(\mathbf{x}) = \sqrt{\frac{1}{t_f - t_0}\int_{t_0}^{t_f} \left(v_{\text{ref}}\left(t,\mathbf{x}\right) - c_{\text{ref}}\left(t,\mathbf{x}\right)\right)^2 \, dt} \tag{24}$$

$$iav_{\text{mod}}\left(\mathbf{x}\right) = \sqrt{\frac{1}{t_f - t_0}\int_{t_0}^{t_f} \left(v_{\text{mod}}\left(t,\mathbf{x}\right) - c_{\text{mod}}\left(t,\mathbf{x}\right)\right)^2 \, dt} \tag{25}$$

$$\varepsilon_{\text{iav}}(\mathbf{x}) = \left(iav_{\text{mod}}\left(\mathbf{x}\right) - iav_{\text{ref}}\left(\mathbf{x}\right)\right) / iav_{\text{ref}}\left(\mathbf{x}\right) \tag{26}$$

and then computing a score as a function of space,

$$s_{\text{iav}}(\mathbf{x}) = e^{-\varepsilon_{\text{iav}}(\mathbf{x})}. \tag{27}$$

The scalar score is then obtained by

$$S_{\text{iav}} = \overline{\overline{s_{\text{iav}}}}(\mathbf{x}), \tag{28}$$

where mass weighting is used when necessary. We include plots of the variability and the score in the standard output and show them here in Figure 7. Note that while here we have shown the interannual variability of the GBAF product for illustration, in the default ILAMB configuration, the interannual variability is currently omitted for the GBAF products because its representativeness is considered to be poor (see Figure 10 of; Kumar et al., 2016).

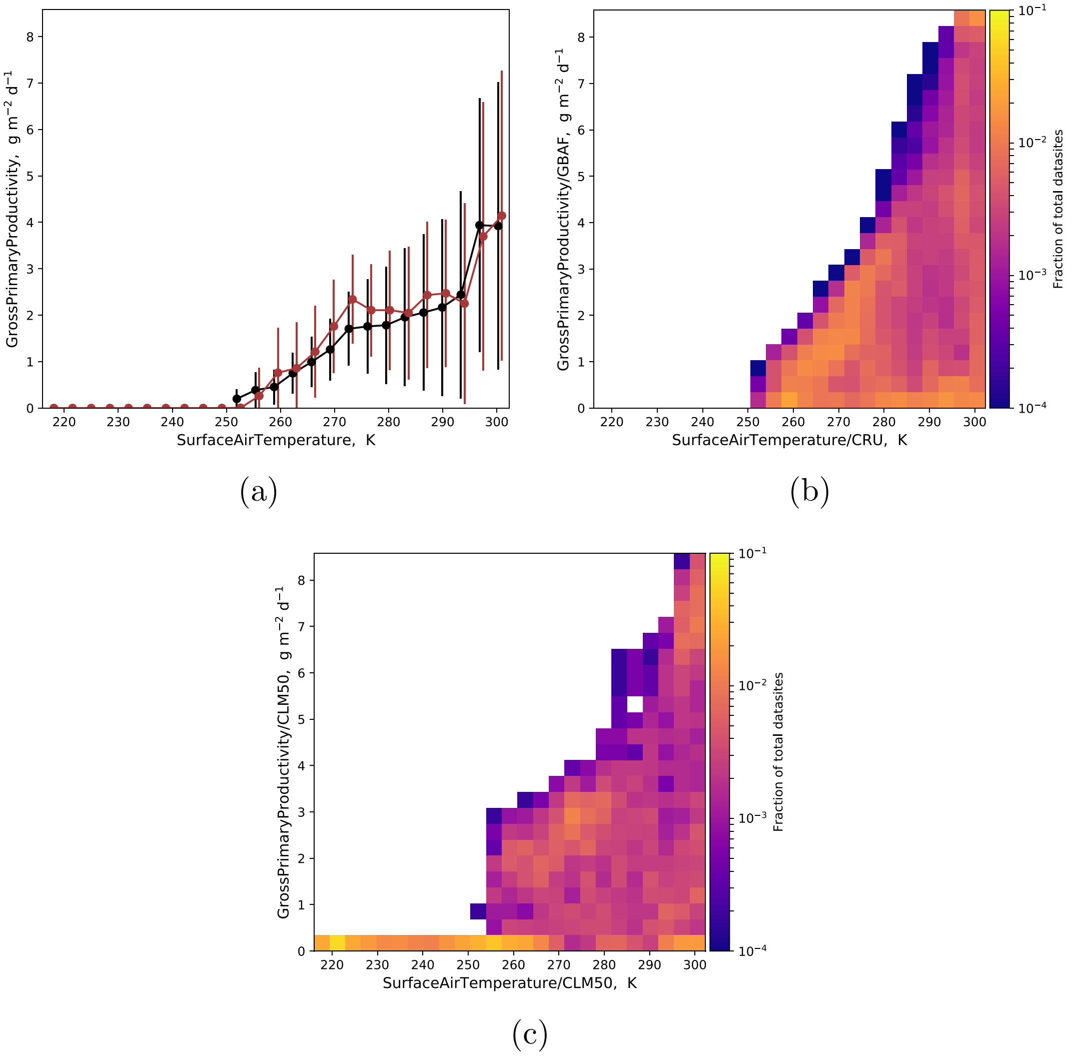


Figure 9. Variable-to-variable relationship plots which are a part of the standard output from the International Land Model Benchmarking methodology. (a) Functional responses, the reference $f_{\rm ref}(u)$ in black, and the model $f_{\rm mod}(u)$ in maroon. Data points reflect the mean for each independent value and the error bars reflect the standard deviation range. (b) Reference distribution, $d_{\rm ref}(u)$; (c) model distribution, $d_{\rm mod}(u)$.

2.2.5. Spatial Distribution

We score the spatial distribution of the time averaged variable by generating a Taylor (Taylor, 2001) diagram. We do this by computing the normalized standard deviation,

$$\sigma = \frac{\text{stdev}\ \left(\overline{v_{\rm mod}}(\mathbf{x})\right)}{\text{stdev}\ \left(\overline{v_{\rm ref}}(\mathbf{x})\right)} \tag{29}$$

and the spatial correlation R of the period mean values $\overline{v_{\rm ref}}(\mathbf{x})$ and $\overline{v_{\rm mod}}(\mathbf{x})$, and then assigning a score by the following relationship

$$S_{\rm dist} = \frac{2(1+R)}{(\sigma + \frac{1}{\sigma})^2}, \tag{30}$$

where the main idea is that we penalize the score when R and σ deviate from a value of 1. We include the Taylor plot in the standard output and represent it here in Figure 8.

2.2.6. Overall Score

The overall score for a given variable and data product is a composite of the suite of metrics defined above. We use a weighted sum,

$$S_{\text{overall}} = \frac{S_{\text{bias}} + 2S_{\text{rmse}} + S_{\text{phase}} + S_{\text{iav}} + S_{\text{dist}}}{1+2+1+1+1}, \tag{31}$$

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

2.3. Relationship Analysis

As models are frequently calibrated using the mean state scalar measures described in section 2.2, a higher score does not necessarily reflect a more process-oriented model. In order to assess the representation of mechanistic processes in models, we also evaluate variable-to-variable relationships. For example, we look at how well models represent the relationship that GPP has with precipitation, evapotranspiration, and temperature. For the purposes of this section, we represent a generic dependent variable as v, as before, and score its relationship with an independent variable u. We then quantify the variable-to-variable relationship of the time period mean, $\bar{u}(\mathbf{x})$ on $\bar{v}(\mathbf{x})$, derived from the combination of reference data sets to the relationship diagnosed in models. We use the mean values over the reference time period to establish relationships as they represent a logical starting point. In the future, we plan to extend the relationship analysis to include seasonal and interannual variability.

2.3.1. Functional Response

We estimate a functional response by a 1-D histogram, binned in terms of the independent variable $\bar{u}(\mathbf{x})$ with a number of bins, initially set to $n_{\text{bins}} = 25$. Then in each bin, we compute the mean value of the corresponding dependent variable, $\bar{v}(\mathbf{x})$ to approximate the functional dependence of u on v. We represent this binning with the operator $\mathcal{F}$ that operates on the dependent and independent variables. We use it to compute functions from both the reference and model data sets.

$$f_{\text{ref}}(u) = \mathcal{F}(\overline{v_{\text{ref}}}(\mathbf{x}), \overline{u_{\text{ref}}}(\mathbf{x})) \tag{32}$$

$$f_{\text{mod}}(u) = \mathcal{F}(\overline{v_{\text{mod}}}(\mathbf{x}), \overline{u_{\text{mod}}}(\mathbf{x})), \tag{33}$$

where both curves are plotted in Figure 9a for the case of GPP compared to surface air temperature. These response curves are then scored by computing a relative error based on the RMSE,

$$\varepsilon^{u}_{\text{func}} = \sqrt{\frac{\int \left(f_{\text{ref}}(u) - f_{\text{mod}}(u)\right)^2 \, \mathrm{d}u}{\int f_{\text{ref}}(u)^2 \, \mathrm{d}u}}, \tag{34}$$

where the integrals are approximated by the midpoint rule over the bins of the independent variable $\bar{u}(\mathbf{x})$. Then we use equation (9) to map this relative error to a score by

$$S^{u}_{\text{func}} = e^{-\varepsilon^{u}_{\text{func}}}. \tag{35}$$

The superscript u reinforces that this score represents functional performance with respect to a given independent variable u. The ILAMB implementation allows for any number of independent variables to be studied. In terms of our sample, ILAMB scores the functional relationship of GPP with respect to each independent variable separately (precipitation, evapotranspiration, temperature, etc.) and then computes the mean of these scores for the overall relationship score.

2.3.2. Hellinger Distance

In addition to the one-dimensional histograms, we also build normalized two-dimensional histograms ($n_{\text{bins}} =$ 25 in both dimensions) from the time averaged data $\bar{v}(\mathbf{x})$ and $\bar{u}(\mathbf{x})$, represented here by the operator $\mathcal{D}$. We represent these distributions by

$$d_{\text{ref}}(u) = \mathcal{D}(\overline{v_{\text{ref}}}(\mathbf{x}), \overline{u_{\text{ref}}}(\mathbf{x})), \tag{36}$$

$$d_{\text{mod}}(u) = \mathcal{D}(\overline{v_{\text{mod}}}(\mathbf{x}), \overline{u_{\text{mod}}}(\mathbf{x})), \tag{37}$$

Table 2
References and Weighting of Data Sets Used to Measure the Ecosystem and Carbon Cycle

Variable/Data Set	Certainty	Scale	Process
Biomass			5
Tropical (Saatchi et al., 2011)	4	4	
NBCD2000 (Kellndorfer et al., 2013)	4	2	
USForest (Blackard et al., 2008)	4	2	
Burned area			4
GFED4S (Giglio et al., 2010)	4	5	
Gross primary productivity			5
Fluxnet (Lasslop et al., 2010)	3	3	
GBAF (Jung et al., 2010)	3	5	
Leaf area index			3
AVHRR (Myneni et al., 1997)	3	5	
MODIS (De Kauwe et al., 2011)	3	5	
Global net ecosystem carbon balance			5
GCP (Le Quéré et al., 2016)	4	5	
Hoffman (Hoffman et al., 2014)	4	5	
Net ecosystem exchange			5
Fluxnet (Lasslop et al., 2010)	3	3	
GBAF (Jung et al., 2010)	2	2	
Ecosystem respiration			4
Fluxnet (Lasslop et al., 2010)	2	3	
GBAF (Jung et al., 2010)	2	2	
Soil carbon			5
HWSD (Todd-Brown et al., 2013)	3	5	
NCSCDV22 (Hugelius et al., 2013)	3	4	

Note. Weights are chosen using the rubric in Table 1 and reflect a focus on understanding the carbon cycle.

as depicted in Figures 9b and 9c. If we represent individual elements from these distributions $d_{\text{ref}}(u) = (p_1, \dots, p_{n^2_{\text{bins}}})$ and $d_{\text{mod}}(u) = (q_1, \dots, q_{n^2_{\text{bins}}})$, we can compute the so-called Hellinger distance (Law et al., 2015)

$$S^u_{\text{dist}} = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n^2_{\text{bins}}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2} \tag{38}$$

as a measure of how similar two distributions are to each other. While there are other choices, such as the Kullback-Leibler divergence, which are more commonly employed (Dirmeyer et al., 2014), the Hellinger distance comes with the added benefit of being already normalized [0,1] and, thus, further normalization is not necessary to use this directly as a score.

However, we only report the Hellinger distance as a scalar and do not include it in the scoring of the relationships. This is due to the fact that a bias in an independent variable can cause a density shift in the 2-D distribution that would cause the score to unreasonably decrease. In terms of our example, a bias in precipitation (e.g., arising from a coupled model) could result in a poor relationship score with GPP, even if there is no underlying deficiency in the land-model-simulated precipitation versus GPP relationship.

3. Data Sets

In this section we explain how we utilize the methodology presented in section 2 to evaluate model performance with respect to a collection of data sets (Tables 2– 5) assembled by the ILAMB community. Errors in measurements, lack of measured or reported uncertainties, and inconsistencies in measurement methodology or instrumentation leading to ambiguous confidence in derived or synthesized data products all

Table 3
References and Weighting of Data Sets Used to Measure the Hydrology Cycle

Variable/Data Set	Certainty	Scale	Process
Evapotranspiration			5
GLEAM (Miralles et al., 2011)	3	5	
MODIS (De Kauwe et al., 2011)	3	5	
Evaporative fraction			5
GBAF (Jung et al., 2010)	3	3	
Latent heat			5
Fluxnet (Lasslop et al., 2010)	3	1	
GBAF (Jung et al., 2010)	3	3	
Runoff			5
Dai (Dai & Trenberth, 2002)	3	5	
Sensible heat			2
Fluxnet (Lasslop et al., 2010)	3	3	
GBAF (Jung et al., 2010)	3	5	
Terrestrial water storage anomaly			5
GRACE (Swenson & Wahr, 2006)	5	5	

Note. Weights are chosen using the rubric in Table 1 and reflect a focus on understanding the carbon cycle.

represent challenges in using observational data for benchmarking. In addition, the spatial and temporal coverage of different data products can vary substantially.

To account for the lack of quantitative uncertainties and scale mismatches between observations and models, and to bring a quantitative objectivity to model-data comparison, we developed a three-element rubric for weighting data sets as represented in Table 1. The first weight is based on a qualitative estimate of the certainty we have in a particular data set. This weight encompasses both our certainty in the process used to obtain the observational information as well as the presence of quantitative uncertainty in the measurements themselves. A second weight for each data set reflects its spatial and temporal coverage. The data sets employed in ILAMB are diverse and include site level data, reanalysis data products, and remotely sensed data. As our aim is to provide insight in land model performance on global and decadal scales, we give more weight to global products, which are time series that extend for several years. The weights are combined multiplicatively to assign a total weight for each data set. Then we normalize the weight by the sum of the weights of all the data sets for a given variable. For example, from Table 2 we see that there are two data sets used to benchmark GPP: Fluxnet and GBAF. For the Fluxnet product, we assign a certainty weight of 3 because while the collection is discussed in the published literature, there is no quantitative uncertainty provided. We assign a scale weight of 3 because the collection of sites covers multiple years of a substantial region of the globe yet has sparse coverage over important regions such as the tropics. The GBAF product is assigned a certainty weight of 3 for the same reason and a scale weight of 5 as it provides global coverage spanning multiple years. Then the total weight for the GPP variable which the GBAF data set carries is

$$w_{\mathrm{GBAF}}^{\mathrm{GPP}} = \frac{3 \cdot 5}{3 \cdot 3 + 3 \cdot 5} \approx 63\%.$$

We use these weights to blend the overall score (equation (31)) from each data set for each variable. In this way ILAMB remains flexible to adding data sets as they are developed, allowing more weight to be given to those that the community believes are more credible and that are more comparable in scale to models.

A third weight reflects how useful the measured variable is in the focus of a model intercomparison project (MIP). Here, as an example, we show weighting for an analysis of model performance in representing the carbon cycle. We use these weights to blend the overall scores from each variable into a complete score across all variables for a given model. This allows ILAMB to include comparisons that are important for a complete understanding of the carbon cycle without necessarily allowing them to heavily influence the overall score. For example, the radiation and energy cycle data sets in Table 4 are all weighted comparatively low because, while they help one understand the carbon cycle, they are not as influential in the overall behavior.

Table 4
References and Weighting of Data Sets Used to Measure the Radiation and Energy Cycle

Variable/Data Set	Certainty	Scale	Process
Albedo			1
CERES (Kato et al., 2013)	4	5	
GEWEX.SRB (Stackhouse et al., 2011)	4	5	
MODIS (De Kauwe et al., 2011)	4	5	
Surface upward SW radiation			1
CERES (Kato et al., 2013)	4	4	
GEWEX.SRB (Stackhouse et al., 2011)	4	5	
WRMC.BSRN (König-Langlo et al., 2013)	4	3	
Surface net SW radiation			1
CERES (Kato et al., 2013)	4	5	
GEWEX.SRB (Stackhouse et al., 2011)	4	5	
WRMC.BSRN (König-Langlo et al., 2013)	4	3	
Surface upward LW radiation			1
CERES (Kato et al., 2013)	4	5	
GEWEX.SRB (Stackhouse et al., 2011)	4	5	
WRMC.BSRN (König-Langlo et al., 2013)	4	3	
Surface net LW radiation			1
CERES (Kato et al., 2013)	4	5	
GEWEX.SRB (Stackhouse et al., 2011)	4	5	
WRMC.BSRN (König-Langlo et al., 2013)	4	3	
Surface net radiation			2
CERES (Kato et al., 2013)	4	5	
Fluxnet (Lasslop et al., 2010)	4	3	
GEWEX.SRB (Stackhouse et al., 2011)	4	5	
WRMC.BSRN (König-Langlo et al., 2013)	4	3	

Note. Weights are chosen using the rubric in Table 1 and reflect a focus on understanding the carbon cycle.

We emphasize that this rubric is particular to our overarching goal of understanding the carbon cycle on global and decadal scales. However, the implementation is flexible and allows for an arbitrary weighting scheme to be developed that suits the needs of the user, community, or MIP that it serves.

The references and weights for each data set that we have selected may be found in Tables 2–5. Each table represents a different aspect of the model: the ecosystem and carbon cycle in Table 2, the hydrological cycle in Table 3, the radiation and energy cycle in Table 4, and the forcings in Table 5. For the majority of these data sets, we make a direct comparison of the observed quantity to model outputs, or algebraic combinations of model outputs using the methodology described in section 2. However, there are a few special cases which require specific handling which we describe in the next section.

3.1. Special Cases

In general, a consistent methodology is applied to compare model output with each data set. This consistency across variables and data sets is a strength of the ILAMB methodology. However, this is not always possible, and here we enumerate a few exceptions and how they are handled.

3.1.1. Evaporative Fraction

To test the partitioning of surface energy, we compare the evaporative fraction derived from the GBAF (Jung et al., 2010) data product to that of the models. The evaporative fraction is an algebraic expression in terms of the latent heat $L_e(t, \mathbf{x})$ and the sensible heat $S_h(t, \mathbf{x})$, given as

$$ef(t, \mathbf{x}) = \frac{L_e(t, \mathbf{x})}{L_e(t, \mathbf{x}) + S_h(t, \mathbf{x})}. \tag{39}$$

Table 5
References and Weighting of Data Sets Used to Measure the Forcings

Variable/Data Set	Certainty	Scale	Process
Surface air temperature			2
CRU (Harris et al., 2014)	5	5	
Fluxnet (Lasslop et al., 2010)	3	3	
Precipitation			2
CMAP (Xie & Arkin, 1997)	4	5	
Fluxnet (Lasslop et al., 2010)	3	3	
GPCC (Schneider et al., 2014)	4	5	
GPCP2 (Adler et al., 2012)	4	5	
Surface relative humidity			3
ERA (Dee et al., 2011)	2	5	
Surface downward SW radiation			2
CERES (Kato et al., 2013)	4	5	
Fluxnet (Lasslop et al., 2010)	4	3	
GEWEX.SRB (Stackhouse et al., 2011)	4	5	
WRMC.BSRN (König-Langlo et al., 2013)	4	3	
Surface downward LW radiation			1
CERES (Kato et al., 2013)	4	5	
GEWEX.SRB (Stackhouse et al., 2011)	4	5	
WRMC.BSRN (König-Langlo et al., 2013)	4	3	

Note. Weights are chosen using the rubric in Table 1 and reflect a focus on understanding the carbon cycle.

The expression can cause nonsensical results because in winter, the sensible heat flux can be negative, leading to a change of sign in the evaporative fraction. The expression can also lead to large evaporative fraction values since the magnitudes of both the latent and sensible heat can become small. For this reason, we apply a mask to ef, L_e, and S_h only considering values for which $S_h > 0$, $L_e > 0$, and $S_h + L_e > \phi$, where $\phi = 20$ (W/m^2) is a surface energy threshold.

Equation (39) is used to study how models partition the surface energy throughout the relevant season. Thus, we use that expression when computing the RMSE or seasonal cycle. However, when comparing period mean values and the bias, equation (39) leads to a combination of averaging methods. For this reason, when computing the mean evaporative fraction over time and the bias, we use a ratio of means in place of the mean of the ratio,

$$\overline{ef}(\mathbf{x}) = \frac{\overline{L_e}(\mathbf{x})}{\overline{L_e}(\mathbf{x}) + \overline{S_h}(\mathbf{x})}. \tag{40}$$

Beyond this change, the evaporative fraction is evaluated using the methodology defined in section 2.

3.1.2. Albedo

We compare the albedo derived from observational data products (Kato et al., 2013; König-Langlo et al., 2013; Stackhouse et al., 2011) to that of models using the following expression:

$$a\ell(t, \mathbf{x}) = \frac{R_u(t, \mathbf{x})}{R_d(t, \mathbf{x})}. \tag{41}$$

where R_u and R_d are the upward shortwave radiation and downward shortwave radiation, respectively. As with the evaporative fraction in section 3.1.1, the albedo expression can become numerically unstable when R_d approaches 0. Thus, we again apply a mask, ignoring regions where no significant incoming radiation is observed, $R_d < \delta$. Equation (41) is used when comparing the RMSE and seasonal cycle. When the period mean and bias are computed, we compute the period mean average albedo based on the ratio of averages,

$$\overline{a\ell}(\mathbf{x}) = \frac{\overline{R_u}(\mathbf{x})}{\overline{R_d}(\mathbf{x})}. \tag{42}$$

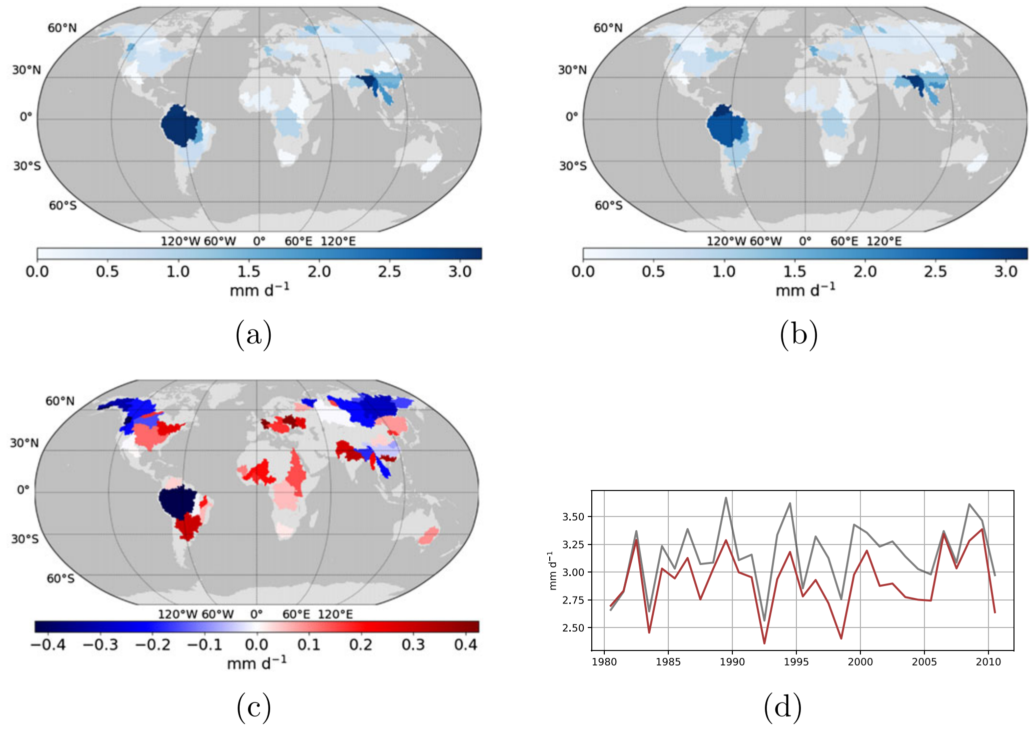


Figure 10. Comparisons of runoff between the reference (Dai & Trenberth, 2002) and the model (CLM4.5) data sets. (a) Reference mean runoff; (b) model mean runoff; (c) mean runoff bias; (d) annual mean runoff for the Amazon river basin where the reference is shown in gray and the model in maroon.

3.1.3. Global Net Ecosystem Carbon Balance

The observational data sets for the global net ecosystem carbon balance (Hoffman et al., 2014; Le Quéré et al., 2016) represent global totals, yet models return this value as fluxes defined over space. To create a model quantity commensurate with the observational data, ILAMB must integrate over the globe using equation (4). As the observational data set is a time series, much of our scoring methodology does not apply. For this discussion we will represent the global rate of carbon as *nbp* (Pg C/year). We compute the accumulation of *nbp*

$$anbp(t) = \int_{t_0}^{t} nbp(t)\,\mathrm{d}t \tag{43}$$

and score the difference in accumulated total at the end of the time period. The precise method differs slightly in each observational data set.

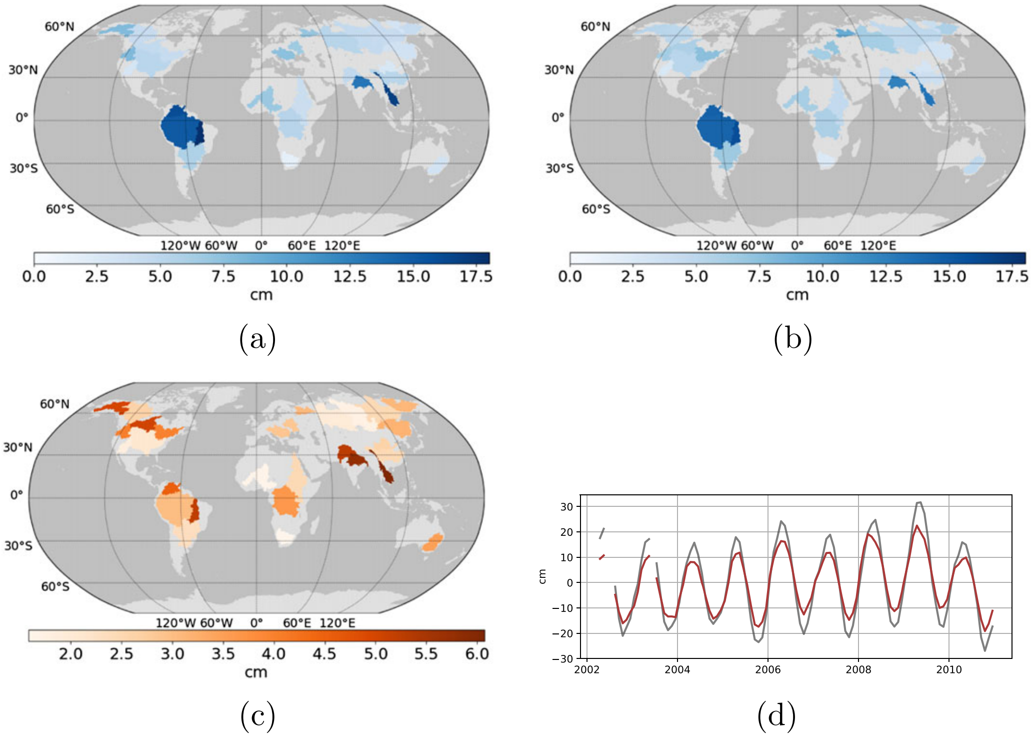
The Global Carbon Project (GCP) data set is derived by taking the land sink (uncertainty of ±0.8 (Pg C/year)) and subtracting the emissions from land use change (uncertainty of ±0.5 (Pg C/year)). This means that the total uncertainty of the accumulated *nbp* at the end of 2010 is $\sqrt{0.5^2 + 0.8^2} \cdot (2010 - 1959) = 48.1$ (Pg C). We use this uncertainty to normalize the difference in accumulation at the end of the time period as a measure of relative error,

$$\varepsilon_{\mathrm{GCP}} = \left| \frac{anbp_{\mathrm{mod}}(2010) - anbp_{\mathrm{ref}}(2010)}{48.1} \right| \tag{44}$$

and then again equation (9) to compute a score of the difference

$$S_{\mathrm{GCP}}^{\mathrm{diff}} = e^{-\alpha_{\mathrm{nbp}}\varepsilon_{\mathrm{GCP}}}, \tag{45}$$

where $\alpha_{\mathrm{nbp}} = 0.287$ and is chosen such that if a model falls within the certainty bounds of the accumulated amount through 2010, the corresponding score is at minimum 0.75. We see this as an important first step in incorporating uncertainty into the comparison methodology. We use the uncertainty to tune the scoring methodology, giving a good score to models that fall inside this uncertainty bound. We also compare the global rates of carbon across the time period in the form of a Taylor score of the time series, $S_{\mathrm{GCP}}^{\mathrm{dist}}$, equation



(a)

(b)

(c)

(d)

Figure 11. Comparisons of the terrestrial water storage anomaly between the reference (Gravity Recovery and Climate Experiment) and the model (CLM4.5) data sets. (a) Reference mean anomaly magnitude; (b) model mean anomaly magnitude; (c) mean anomaly RMSE; (d) annual mean anomaly for the Amazon river basin where the reference is shown in gray and the model in maroon.

(30), where the correlation and standard deviation are taken across the temporal dimension. Then the overall score is

$$S_{\text{GCP}}^{\text{nbp}} = \frac{1}{2}\left(S_{\text{GCP}}^{\text{diff}} + S_{\text{GCP}}^{\text{dist}}\right) \tag{46}$$

In the Hoffman et al. (2014) data set, we only score the accumulated amount at the end of the observed period. We omit providing a Taylor scoring of the rates because there appears to be some smoothing of the rate data inherent in the process of producing this data set. However, this data set explicitly provides a lower and upper bound on uncertainty as a function of time throughout the data set. So we determine the integrated uncertainty at the end of 2010 by accumulating the upper (52.4 Pg C) and lower (−32.1 Pg C) limit of uncertainty, computing the difference, and then halving the value resulting in an uncertainty of 42.3 (Pg C). We then use the same approach to score the difference,

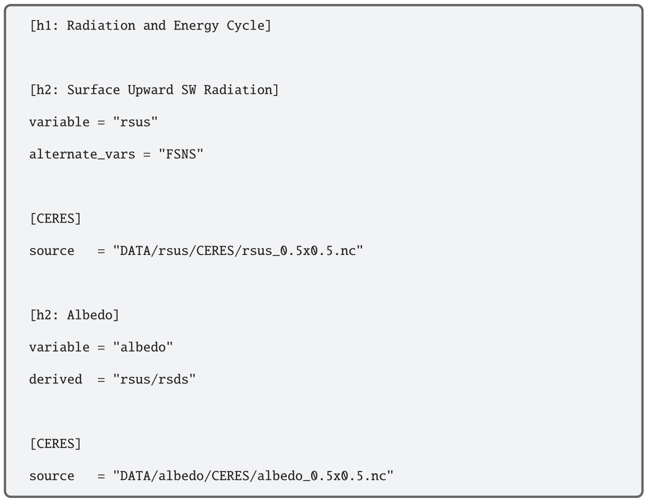
$$\varepsilon_{\text{Hoffman}} = \left|\frac{a_{\text{mod}}(2010) - a_{\text{ref}}(2010)}{42.3}\right| \tag{47}$$

$$S_{\text{Hoffman}}^{\text{nbp}} = e^{-\alpha_{\text{nbp}}\varepsilon_{\text{Hoffman}}} \tag{48}$$

3.1.4. Runoff

We use the Dai and Trenberth (2002) river discharge data set to assess model performance of runoff for the world's 50 largest river basins. First, we compute the mean annual runoff from the model over the time period of the observational data set. Then we take the river discharge data and distribute it over the area of the river basins and compare this to the mean runoff over the same basin. This simple approach was taken to allow us to compare runoff across models even if they do not have a river routing model.

We include plots of the mean runoff of the reference and model over river basins and the bias, represented in Figure 10. We also include regional mean runoff plots for each of the river basins included but only show that



```
[h1: Radiation and Energy Cycle]


[h2: Surface Upward SW Radiation]

variable = "rsus"

alternate_vars = "FSNS"


[CERES]

source   = "DATA/rsus/CERES/rsus_0.5x0.5.nc"


[h2: Albedo]

variable = "albedo"

derived  = "rsus/rsds"


[CERES]

source   = "DATA/albedo/CERES/albedo_0.5x0.5.nc"
```

Figure 12. Sample International Land Model Benchmarking configure file defining comparisons to the surface upward shortwave radiation and albedo variables from the CERES (Kato et al., 2013) product.

of the Amazon river basin in Figure 10d. The model performance is then scored using the bias (section 2.2.1), the interannual variability (section 2.2.4), and the spatial distribution (section 2.2.5) metrics.

3.1.5. Terrestrial Water Storage Anomaly

We use the Gravity Recovery and Climate Experiment (Swenson & Wahr, 2006) data set to assess the terrestrial water storage anomaly in models. However, there are a few challenges in producing a fair comparison. The first of those is that models report only the storage and so the anomaly must be computed. The more serious challenge is that the resolution of this data is quite coarse (300–400 km]), and thus, pointwise comparisons are not appropriate (Swenson, 2013). Instead, we compare mean anomaly values over 30 of the world's largest river basins. In this way the comparison is more fair as it is over large areas and automatically omits dry areas, which are not of interest.

We include plots of the magnitude of the mean anomaly of the reference and model over river basins and the RMSE, represented in Figure 11. We also include regional mean anomaly plots for each of the river basins but show only that of the Amazon river basin in Figure 11d. The model performance is then scored using the RMSE (section 2.2.2) and the interannual variability (section 2.2.4) metrics.

4. Software

We have implemented the methodology described in sections 2 and 3 into a software package that is freely available to the community. We previously developed a prototype implementation (Mu et al., 2015) based on the National Center for Atmospheric Research Command Language. We then moved the algorithm into an open source, openly developed python package (Collier et al., 2016) in an effort to produce a product to which the community can more easily make contributions. The referenced digital object identifier will lead to the software repository, where the source code and documentation can be found. The documentation includes the public interface as well as tutorials that span topics such as installation, basic usage, adding models or benchmark data sets, and formatting benchmark data sets.

The ILAMB package is designed to ingest data sets that follow the Climate and Forecast convention (Eaton et al., 2017). The Climate and Forecast website explains that the "conventions define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities." We have built the ILAMB package to embody this philosophy, making it directly useful to those who adhere to this standard. While model intercomparison efforts, such as CMIP5, have encouraged the use of these conventions among modelers, the observational community has not yet widely put them into practice. Much of the work in adding data sets to the collection is in encoding them to follow this convention.

For the purpose of communicating how the ILAMB package works, consider the configure file shown in Figure 12, which defines a set of observational data sets that will be used to confront models. The *h1* bracket is a heading used to categorize variables, represented by the *h2* heading. This comparison involves the surface upward shortwave radiation and the albedo, both of which are variables belonging to the radiation and energy cycle. Inside each h2 heading, we specify the variable name that will be compared (*rsus* is the netCDF variable name for surface upward shortwave radiation). However, we provide a mechanism for variable synonyms in this case by specifying alternate variable names. If the ILAMB system cannot find the main variable, it will try to find any alternates that the user specifies. This allows the software to encourage the use of standard variable names but accounts for modeling groups wanting to use ILAMB without preprocessing. Also note the *derived* keyword in the albedo section. While the components of albedo are part of standard model output, the albedo is not. The ILAMB package allows for users to specify algebraic relationships in the configure file process. This makes the process automatic and transparent to those who may read this configure file.

The ILAMB package will ingest this configure file and try to build commensurate quantities from model outputs. While observational data sets come in different forms (globally gridded remote sensing products, tower data collections, etc.), the ILAMB system reads the spatial and temporal information found in the file and uses it to trim, subsample, and/or coarsen the model data as appropriate.

5. Discussion

The ILAMB framework is designed to be both powerful and flexible. While we have made choices in the default configuration, described above, focused on global analysis for decadal to centennial scale ESMs, ILAMB allows the user to customize selection of variables, weighting of data sets, and spatial subsetting that make it useful for assessing results from mesoscale weather forecasting or other models. We envision developing a library of sample configuration files, targeting various well-known models and model applications.

As much of the usefulness of ILAMB depends on the quality of the underlying observational data, we recommend that data providers include explicit representations of the underlying spatial grids including the areas over which quantities have been averaged. Observational data sets frequently report mean values in a cell taken over an area which may include land but also portions of lakes, rivers, and oceans. This leads to ambiguity with regard to the contribution of land cover types to the measurement itself and subsequently adds to the uncertainty when comparing values to model output.

5.1. Interpreting the Overall Score

The thrust of this paper is to detail a methodology for computing a single overall score that captures a model's skill in reproducing patterns found in the observed record. However, we do not view the absolute value of the score as particularly meaningful beyond the precise definition described in this paper. In general, no model can achieve a perfect score for any given variable for several reasons.

First, there is measurement error and uncertainty in the observational data that makes a perfect score unlikely or undesirable against even a single data set. This is what motivates some in the community (Abramowitz, 2005; Best et al., 2015) to pose that benchmarking requires an expectation of performance, which is admittedly lacking in our approach. Second, despite that every attempt is made to employ multiple independent data sets of high quality for confrontation with models, these data sets are inconsistent with each other, making a perfect score across all data sets impossible. We do this as comparisons with multiple observational and synthesized data sets for a single variable to offer the user more information about the robustness of model predictions within the limits of observational uncertainty at varying spatial and temporal scales. Third, a lower score with respect to a given variable is not necessarily a sign of a poor model. It may rather highlight the need for better measurement campaigns or improved metrics (i.e., sometimes we learn that our measurements are incomplete or do not acknowledge important uncertainties, or our metrics are inappropriate for a given data set).

The overall score is meant to aid the scientist in discovering when meaningful changes have occurred in the model or across models. The holistic nature of the ILAMB suite of data sets and metrics helps provide a synthesis of model performance that directs the attention of the user to relevant aspects. While we present Figure 1 as the main result of the ILAMB methodology, it is intended to merely indicate variables of particular interest for further consideration. ILAMB output is presented as a hierarchy of interactive web pages that employ JavaScript features to present information to users in a logical and intuitive fashion. From the graphical overview, the user can select individual variables and data sets from the *Results Table* tab to be led to pages that detail the contributing factors to the model's overall score. On this new page, predefined spatial regions can be individually selected, causing the tabular data and diagnostics to be updated automatically to reflect information relevant only to that region. Although all the tabular information, scores, and graphical diagnostics are precomputed and generated when ILAMB is run, the web-based interface is designed to facilitate discovery and understanding of model results. The overall score does not replace the scientist, it guides her/him to the relevant plots and diagnostics.

5.2. How Is ILAMB Used?

The ILAMB package is particularly useful for verification, that is, during model development to confirm that new model code improves performance in a targeted area without degrading performance in another area, and for validation, that is, when comparing performance of one model or model version to that of other models or model versions.

In developing and applying the ILAMB package, we have incorporated a wide variety of representative observational data sets (see Tables 2–5), and we have favored data that have the most open data policies.

In many cases, these data have been averaged or remapped to be more directly comparable with model output. As this collection of data sets grows, maintaining and distributing the latest versions will be challenging and require community collaboration. For tracking the evolving performance of models over the long term, it may be necessary to maintain access to older versions of data as well as the latest version since corrections to observational data sets can significantly impact model performance scores. Various technologies could fill this role, and the Observations for Climate Model Intercomparisons (obs4MIPs; https://www.earthsystemcog.org/projects/obs4mips/) activity shows promise as a potential solution to this challenge. The preferred solution would ideally support versioning and allow for long-lived versions associated with ILAMB releases. In the interim, we have implemented a simple scheme for sharing summarized and remapped data sets through a web server.

The ILAMB package is currently being used by individual model developers and international modeling centers. ILAMB offers developers a quick and easy method for checking the impacts of new model development before committing code changes. For modeling centers, ILAMB provides a systematic assessment of historical simulation experiments and enables tracking of performance of model revisions. ILAMB will also be useful for MIPs as a starting point for evaluating model variability and uncertainty. As a part of such MIPs, investigators may wish to develop custom metrics or incorporate data sets specific to their purposes. ILAMB could be executed automatically as model results are uploaded to a system like the Earth System Grid Federation (https://esgf.llnl.gov/) to give users a *first look* at variation in results and to determine if output should be downloaded for a particular study. ILAMB diagnostics can also be useful for parameter sensitivity studies or for optimization experiments in combination with an automated modeling framework like the Predictive Ecosystem Analyzer (http://pecanproject.org/; Dietze et al., 2014; LeBauer et al., 2013). For the assessments community, the results of a multimodel ILAMB evaluation could be useful for understanding which model results would be appropriate for use in studying impacts and which models may poorly capture processes relevant to the impacts under consideration.

5.3. Future Work

Development of the ILAMB package is ongoing, and the terrestrial modeling and observational communities are being engaged to identify in situ and remote sensing data sets, to define additional evaluation metrics, and to use the package for a wide variety of MIPs (Hoffman et al., 2017). While most effort has been invested in global- and regional-scale model evaluation, new work is focused on improved benchmarking for site level time series, spatial transects, and seasonal and diurnal variability. Future development will include incorporation of experiment-specific model evaluation metrics derived from prior studies, including Free-Air CO_2 Enrichment (Walker et al., 2014, 2015; Zaehle et al., 2014), nutrient addition, rainfall exclusion, and warming experiments (Bouskill et al., 2014; Zhu et al., 2016). Partner activities, like NASA's Permafrost Benchmarking System project and the Arctic-Boreal Vulnerability Experiment, are integrating additional data sets and building metrics for specific regions, study areas, or processes of interest. We are applying the ILAMB methodology and code base to develop a marine biogeochemical model benchmarking tool, called the International Ocean Model Benchmarking package.

Based on previous prototypes and community discussion, we developed the ILAMB model benchmarking package for evaluating the fidelity of land carbon cycle models. The package generates graphical diagnostics and computes a comprehensive set of statistics through model-data comparisons and scores model performance for a wide variety of variables for a suite of observational data sets. Rigorously defined model evaluation metrics and strategies for handling multiple resolutions and land masks are documented above. The ILAMB package is open source and is becoming widely adopted by modeling centers and for informing model intercomparison studies. We are actively seeking community involvement in adding more evaluation metrics and new observational data sets.

Acknowledgments
This manuscript has been authored by UT-Battelle, LLC under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). 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 Climate Modeling (RGCM) 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 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 DE-AC02-05CH11231.

References

Abramowitz, G. (2005). Towards a benchmark for land surface models. *Geophysical Research Letters*, *32*, L22702. https://doi.org/10.1029/2005GL024419

Abramowitz, G. (2012). Towards a public, standardized, diagnostic benchmarking system for land surface models. *Geoscientific Model Development*, *5*(3), 819–827. https://doi.org/10.5194/gmd-5-819-2012

Adler, R. F., Gu, G., & Huffman, G. J. (2012). Estimating climatological bias errors for the Global Precipitation Climatology Project (GPCP). *Journal of Applied Meteorology and Climatology*, *51*(1), 84–99. https://doi.org/10.1175/JAMC-D-11-052.1

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., et al. (2013). Evaluating the land and ocean components of the global carbon cycle in the CMIP5 Earth system models. *Journal of Climate*, *26*(18), 6801–6843. https://doi.org/10.1175/JCLI-D-12-00417.1

Arora, V. K., Boer, G. J., Friedlingstein, P., Eby, M., Jones, C. D., Christian, J. R., et al. (2013). Carbon-concentration and carbon-climate feedbacks in CMIP5 Earth system models. *Journal of Climate*, *26*(15), 5289–5314. https://doi.org/10.1175/JCLI-D-12-00494.1

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., et al. (2015). The plumbing of land surface models: Benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), 1425–1442. https://doi.org/10.1175/JHM-D-14-0158.1

Blackard, J. A., Finco, M. V., Helmer, E. H., Holden, G. R., Hoppus, M. L., Jacobs, D. M., et al. (2008). Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment*, *112*(4), 1658–1677. https://doi.org/10.1016/j.rse.2007.08.021, Remote Sensing Data Assimilation Special Issue.

Blyth, E., Clark, D. B., Ellis, R., Huntingford, C., Los, S., Pryor, M., et al. (2011). A comprehensive set of benchmark tests for a land surface model of simultaneous fluxes of water and carbon at both the global and seasonal scale. *Geoscientific Model Development*, *4*(2), 255–269. https://doi.org/10.5194/gmd-4-255-2011

Bouskill, N. J., Riley, W. J., & Tang, J. (2014). Meta-analysis of high-latitude nitrogen-addition and warming studies implies ecological mechanisms overlooked by land models. *Biogeosciences*, *11*(23), 6969–6983. https://doi.org/10.5194/bg-11-6969-2014

Cadule, P., Friedlingstein, P., Bopp, L., Sitch, S., Jones, C. D., Ciais, P., et al. (2010). Benchmarking coupled climate-carbon models against long-term atmospheric CO_2 measurements. *Global Biogeochemical Cycles*, *24*, GB2016. https://doi.org/10.1029/2009GB003556

Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., et al. (2013). Carbon and other biogeochemical cycles. In T. F. Stocker, et al. (Eds.), *Climate change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 465–570). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Collier, N., Hoffman, F., Mu, M., Randerson, J. T., & Riley, W. J. (2016). International Land Model Benchmarking (ILAMB) package v2, online. https://doi.org/10.18139/ILAMB.v002.00/1251621

Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A., & Totterdell, I. J. (2000). Acceleration of global warming due to carbon–cycle feedbacks in a coupled climate model. *Nature*, *408*(6809), 184–187. https://doi.org/10.1038/35041539

Dai, A., & Trenberth, K. E. (2002). Estimates of freshwater discharge from continents: Latitudinal and seasonal variations. *Journal of Hydrometeorology*, *3*(6), 660–687. https://doi.org/10.1175/1525-7541(2002)003<0660:EOFDFC>2.0.CO;2

Dalmonech, D., & Zaehle, S. (2013). Towards a more objective evaluation of modelled land-carbon trends using atmospheric CO_2 and satellite-based vegetation activity observations. *Biogeosciences*, *10*(6), 4189–4210. https://doi.org/10.5194/bg-10-4189-2013

De Kauwe, M. G., Disney, M. I., Quaife, T., Lewis, P., & Williams, M. (2011). An assessment of the MODIS Collection 5 leaf area index product for a region of mixed coniferous forest. *Remote Sensing of Environment*, *115*(2), 767–780.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal Of The Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Denman, K. L., Brasseur, G., Chidthaisong, A., Ciais, P., Cox, P. M., Dickinson, R. E., et al. (2007). Couplings between changes in the climate system and biogeochemistry. In S. Solomon, et al. (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 499–587). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Dietze, M. C., Serbin, S. P., Davidson, C., Desai, A. R., Feng, X., Kelly, R., et al. (2014). A quantitative assessment of a terrestrial biosphere model's data needs across North American biomes. *Journal of Geophysical Research: Biogeosciences*, *119*, 286–300. https://doi.org/10.1002/2013JG002392

Dirmeyer, P. A., Wei, J., Bosilovich, M. G., & Mocko, D. M. (2014). Comparing evaporative sources of terrestrial precipitation and their extremes in MERRA using relative entropy. *Journal of Hydrometeorology*, *15*(1), 102–116. https://doi.org/10.1175/JHM-D-13-053.1

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., et al. (2017). Netcdf Climate and Forecast (CF) metadata conventions. http://cfconventions.org/

Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., et al. (2016). ESMValTool (v1.0)—A community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *9*(5), 1747–1802. https://doi.org/10.5194/gmd-9-1747-2016

Friedlingstein, P., Bopp, L., Ciais, P., Dufresne, J.-L., Fairhead, L., LeTreut, H., et al. (2001). Positive feedback between future climate change and the carbon cycle. *Geophysical Research Letters*, *28*(8), 1543–1546. https://doi.org/10.1029/2000GL012015

Friedlingstein, P., Cox, P. M., Betts, R. A., Bopp, L., von Bloh, W., Brovkin, V., et al. (2006). Climate–carbon cycle feedback analysis: Results from the C^4MIP model intercomparison. *Journal of Climate*, *19*(14), 3373–3353. https://doi.org/10.1175/JCLI3800.1

Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K., & Knutti, R. (2014). Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *Journal of Climate*, *27*(2), 511–526. https://doi.org/10.1175/JCLI-D-12-00579.1

Fung, I. Y., Doney, S. C., Lindsay, K., & John, J. (2005). Evolution of carbon sinks in a changing climate. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(32), 11,201–11,206. https://doi.org/10.1073/pnas.0504949102

Ghimire, B., Riley, W. J., Koven, C. D., Mu, M., & Randerson, J. T. (2016). Representing leaf and root physiological traits in CLM improves global carbon and nitrogen cycling predictions. *Journal of Advances in Modeling Earth Systems*, *8*, 598–613. https://doi.org/10.1002/2015MS000538

Giglio, L., Randerson, J. T., van der Werf, G. R., Kasibhatla, P. S., Collatz, G. J., Morton, D. C., & DeFries, R. S. (2010). Assessing variability and long-term trends in burned area by merging multiple satellite fire products. *Biogeosciences*, *7*(3), 1171–1186. https://doi.org/10.5194/bg-7-1171-2010

Gleckler, P. J., Doutriaux, C., Durack, P. J., Taylor, K. E., Zhang, Y., Williams, D. N., et al. (2016). A more powerful reality test for climate models. *Eos, Transactions American Geophysical Union*, *97*. https://doi.org/10.1029/2016EO051663

Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research*, *113*, D06104. https://doi.org/10.1029/2007JD008972

Gregory, J. M., Jones, C. D., Cadule, P., & Friedlingstein, P. (2009). Quantifying carbon cycle feedbacks. *Journal of Climate*, *22*(19), 5232–5250. https://doi.org/10.1175/2009JCLI2949.1

Harris, I., Jones, P., Osborn, T., & Lister, D. (2014). Updated high-resolution grids of monthly climatic observations—The CRU TS3.10 dataset. *International Journal of Climatology*, *34*(3), 623–642. https://doi.org/10.1002/joc.3711

Hoffman, F. M., Koven, C. D., Keppel-Aleks, G., Lawrence, D. M., Riley, W. J., Randerson, J. T., et al. (2017). International Land Model Benchmarking (ILAMB) 2016 workshop report (Tech. Rep. DOE/SC-0186). Germantown, Maryland, USA: U.S. Department of Energy, Office of Science. https://doi.org/10.2172/1330803

Hoffman, F. M., Randerson, J. T., Arora, V. K., Bao, Q., Cadule, P., Ji, D., et al. (2014). Causes and implications of persistent atmospheric carbon dioxide biases in Earth System Models. *Journal of Geophysical Research: Biogeosciences*, *119*, 141–162. https://doi.org/10.1002/2013JG002381

Hugelius, G., Bockheim, J. G., Camill, P., Elberling, B., Grosse, G., Harden, J. W., et al. (2013). A new data set for estimating organic carbon storage to 3 m depth in soils of the northern circumpolar permafrost region. *Earth System Science Data*, *5*(2), 393–402. https://doi.org/10.5194/essd-5-393-2013

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., et al. (2010). Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature*, *467*, 951–954. https://doi.org/10.1038/nature09396

Kato, S., Loeb, N. G., Rose, F. G., Doelling, D. R., Rutan, D. A., Caldwell, T. E., et al. (2013). Surface irradiances consistent with CERES-derived top-of-atmosphere shortwave and longwave irradiances. *Journal of Climate*, *26*(9), 2719–2740. https://doi.org/10.1175/JCLI-D-12-00436.1

Kelley, D. I., Prentice, I. C., Harrison, S. P., Wang, H., Simard, M., Fisher, J. B., & Willis, K. O. (2013). A comprehensive benchmarking system for evaluating global vegetation models. *Biogeosciences*, *10*(5), 3313–3340. https://doi.org/10.5194/bg-10-3313-2013

Kellndorfer, J., Walker, W., Kirsch, K., Fiske, G., Bishop, J., Lapoint, L., et al. (2013). NACP aboveground biomass and carbon baseline data, V.2 (NBCD 2000), U.S.A., 2000. https://doi.org/10.3334/ornldaac/1161

König-Langlo, G., Sieger, R., Schmithüsen, H., Bücker, A., Richter, F., & Dutton, E. (2013). The Baseline Surface Radiation Network and its World Radiation Monitoring Centre at the Alfred Wegener Institute (Tech. Rep. WCRP-2). Bremerhaven, Germany: Alfred Wegener Institute. https://doi.org/10013/epic.42596.d001

Kumar, J., Hoffman, F. M., Hargrove, W. W., & Collier, N. (2016). Understanding the representativeness of FLUXNET for upscaling carbon flux from eddy covariance measurements. *Earth System Science Data Discussions*, *2016*, 1–25. https://doi.org/10.5194/essd-2016-36

Kumar, S. V., Peters-Lidard, C. D., Santanello, J., Harrison, K., Liu, Y., & Shaw, M. (2012). Land surface Verification Toolkit (LVT)—A generalized framework for land surface model evaluation. *Geoscientific Model Development*, *5*(3), 869–886. https://doi.org/10.5194/gmd-5-869-2012

Lasslop, G., Reichstein, M., Papale, D., Richardson, A. D., Arneth, A., Barr, A., et al. (2010). Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: Critical issues and global evaluation. *Global Change Biology*, *16*(1), 187–208. https://doi.org/10.1111/j.1365-2486.2009.02041.x

Law, K., Stuart, A., & Zygalakis, K. (2015). *Data Assimilation: A Mathematical Introduction, Texts in Applied Mathematics* (1st ed., Vol. 63, pp. 242). Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-20325-6

Le Quéré, C., Andrew, R. M., Canadell, J. G., Sitch, S., Korsbakken, J. I., Peters, G. P., et al. (2016). Global carbon budget 2016. *Earth System Science Data*, *8*(2), 605–649. https://doi.org/10.5194/essd-8-605-2016

LeBauer, D. S., Wang, D., Richter, K. T., Davidson, C. C., & Dietze, M. C. (2013). Facilitating feedbacks between field measurements and ecosystem models. *Ecology Monographs*, *83*(2), 133–154. https://doi.org/10.1890/12-0137.1

Luo, Y. Q., Randerson, J. T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., et al. (2012). A framework for benchmarking land models. *Biogeosciences*, *9*(10), 3857–3874. https://doi.org/10.5194/bg-9-3857-2012

Mahowald, N. M., Randerson, J. T., Lindsay, K., Muñoz, E., Doney, S. C., Lawrence, P., et al. (2017). Interactions between land use change and carbon cycle feedbacks. *Global Biogeochemical Cycles*, *31*, 96–113. https://doi.org/10.1002/2016GB005374

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., et al. (2007). The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, *88*(9), 1383–1394. https://doi.org/10.1175/BAMS-88-9-1383

Miralles, D., Holmes, T., Jeu, R. D., Gash, J., Meesters, A., & Dolman, A. (2011). Global land-surface evaporation estimated from satellite-based observations. *Hydrology and Earth System Sciences*, *15*, 453–469.

Moore, J. K., Fu, W., Primeau, F., Britten, G. L., Lindsay, K., Long, M., et al. (2018). Sustained climate warming drives declining marine biological productivity. *Science*, *359*(6380), 1139–1143. https://doi.org/10.1126/science.aao6379

Mu, M., Randerson, J. T., Riley, W. J., Koven, C. D., Keppel-Aleks, D., Lawrence, D. M., & Hoffman, F. M. (2015). International Land Model Benchmarking (ILAMB) package v1, online. https://doi.org/10.18139/ILAMB.v001.00/1251597

Myneni, R. B., Nemani, R. R., & Running, S. (1997). Algorithm for the estimation of global land cover, LAI and FPAR based on radiative transfer models. *IEEE Transactions on Geoscience and Remote Sensing*, *35*, 1380–1392.

Oleson, K. W., Lawrence, D. M., Bonan, G. B., Drewniak, B., Huang, M., Koven, C. D., et al. (2013). Technical description of version 4.5 of the Community Land Model (CLM) (*Technical Note NCAR/TN-503+STR*). Boulder, Colorado, USA: National Center for Atmospheric Research.

Piao, S., Sitch, S., Ciais, P., Friedlingstein, P., Peylin, P., Wang, X., et al. (2013). Evaluation of terrestrial carbon cycle models for their response to climate variability and to CO_2 trends. *Global Change Biology*, *19*(7), 2117–2132. https://doi.org/10.1111/gcb.12187

Randerson, J. T., Hoffman, F. M., Thornton, P. E., Mahowald, N. M., Lindsay, K., Lee, Y.-H., et al. (2009). Systematic assessment of terrestrial biogeochemistry in coupled climate–carbon models. *Global Change Biology*, *15*(9), 2462–2484. https://doi.org/10.1111/j.1365-2486.2009.01912.x

Randerson, J. T., Lindsay, K., Munoz, E., Fu, W., Moore, J. K., Hoffman, F. M., et al. (2015). Multicentury changes in ocean and land contributions to the climate–carbon feedback. *Global Biogeochemical Cycles*, *29*, 744–759. https://doi.org/10.1002/2014GB005079

Reichler, T., & Kim, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, *89*(3), 303–311. https://doi.org/10.1175/BAMS-89-3-303

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. A., Salas, W., et al. (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(24), 9899–9904. https://doi.org/10.1073/pnas.1019576108

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., & Rudolf, B. (2014). GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. *Theoretical and Applied Climatology*, *115*(1), 15–40. https://doi.org/10.1007/s00704-013-0860-x

Stackhouse, P. W. Jr., Gupta, S. K., Cox, S. J., Mikovitz, J. C., Zhang, T., & Hinkelman, L. M. (2011). The NASA/GEWEX surface radiation budget release 3.0: 24.5-year dataset. *GEWEX News*, *21*(1), 10–12.

Swenson, S. (2013). GRACE: Gravity Recovery and Climate Experiment: Surface mass, total water storage, and derived variables, *The Climate Data Guide, edited by National Center for Atmospheric Research Staff* . Boulder, CO: National Center for Atmospheric Research.

Swenson, S., & Wahr, J. (2006). Post-processing removal of correlated errors in GRACE data. *Geophysical Research Letters*, *33*, L08402. https://doi.org/10.1029/2005GL025285

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000JD900719

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E. A. G., & Allison, S. D. (2013). Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. *Biogeosciences*, *10*(3), 1717–1736. https://doi.org/10.5194/bg-10-1717-2013

Walker, A. P., Hanson, P. J., De Kauwe, M. G., Medlyn, B. E., Zaehle, S., Asao, S., et al. (2014). Comprehensive ecosystem model–data synthesis using multiple data sets at two temperate forest free-air CO_2 enrichment experiments: Model performance at ambient CO_2 concentration. *Journal of Geophysical Research: Biogeosciences*, *119*, 2169–8961. https://doi.org/10.1002/2013JG002553

Walker, A. P., Zaehle, S., Medlyn, B. E., De Kauwe, M. G., Asao, S., Hickler, T., et al. (2015). Predicting long-term carbon sequestration in response to CO_2 enrichment: How and why do current ecosystem models differ? *Global Biogeochemical Cycles*, *29*, 476–495. https://doi.org/10.1002/2014GB004995

Xie, P., & Arkin, P. A. (1997). Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bulletin of the American Meteorological Society*, *78*(11), 2539–2558. https://doi.org/10.1175/1520-0477(1997)078<2539:GPAYMA>2.0.CO;2

Zaehle, S., Medlyn, B. E., De Kauwe, M. G., Walker, A. P., Dietze, M. C., Hickler, T., et al. (2014). Evaluation of 11 terrestrial carbon–nitrogen cycle models against observations from two temperate free-air CO_2 enrichment studies. *New Phytologist*, *202*(3), 803–822. https://doi.org/10.1111/nph.12697

Zhu, Q., Riley, W. J., Tang, J., & Koven, C. D. (2016). Multiple soil nutrient competition between plants, microbes, and mineral surfaces: Model development, parameterization, and example applications in several tropical forests. *Biogeosciences*, *13*(1), 341–363. https://doi.org/10.5194/bg-13-341-2016