CHAPTER NINETEEN

Benchmark Analysis

Yiqi Luo
Cornell University, Ithaca, USA

Forrest M. Hoffman
Oak Ridge National Laboratory, Oak Ridge, USA

CONTENTS

Tremendous progress has been achieved in the development of land models and their inclusion in Earth system models (ESMs). However, we still have very limited knowledge on the performance skills of these land models. This chapter introduces benchmark analysis, which is a procedure to measure performance of models against a set of defined standards. The benchmark analysis includes: (1) defining targeted aspects of model performance to be evaluated; (2) testing model performance in comparison with a set of benchmarks; (3) measuring model performance skill through quantitative metrics; and (4) evaluating model performance and offering suggestions for future model improvement.

INTRODUCTION

Over the past decades, tremendous progress has been achieved in the development of land models and their inclusion in Earth system models (ESMs). State-of-the-art land models now account for biophysical processes (exchanges of water and energy) and biogeochemical cycles of carbon, nitrogen, and trace gases. They also simulate vegetation dynamics and disturbances. When coupled as components in ESMs, land models now allow simulation of land-atmosphere biophysical interactions and climate-carbon feedbacks. These models are now widely used for policy-relevant assessment of climate change and its impact on ecosystems or terrestrial resources. However, there is still very limited knowledge of the performance skills of these land models, especially when embedded in ESMs. Quantifying the performance skills of land models would promote confidence in their predictions of future states of ecosystems and climate, and identify those models whose predictions are more likely to be accurate, where ensemble members diverge.

Model performance has traditionally been evaluated via comparison with observed data sets. 'Validation' by plotting model data side-by-side with observed data, or computing mismatch metrics such as root-mean-square-error, is traditionally the most common approach to model evaluation (Oreskes, 2003; Rykiel, 1996; see also

DOI: 10.1201/9780429155659-24

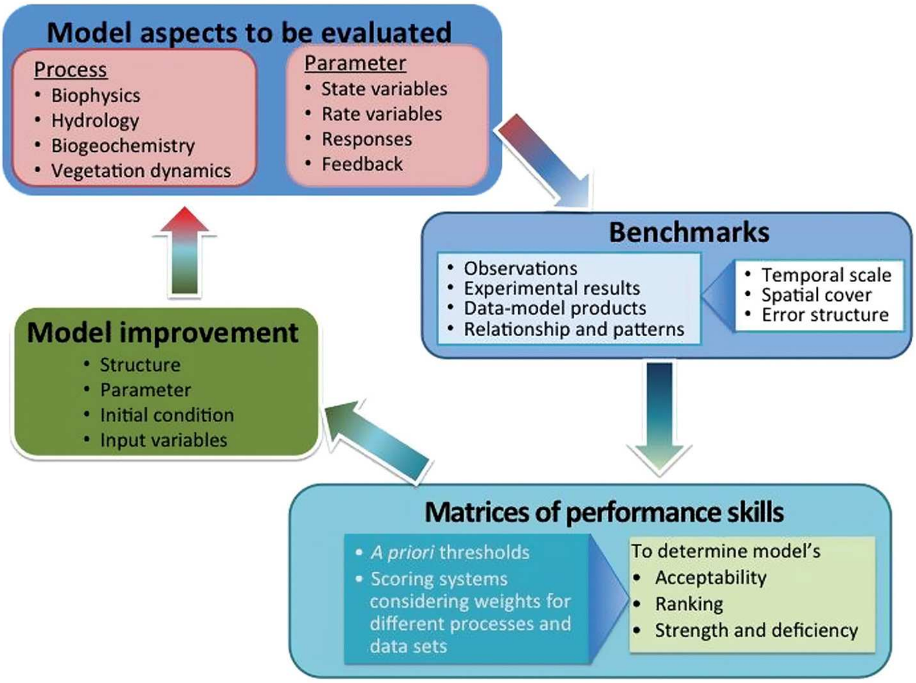

Figure 19.1. Schematic diagram of the benchmarking framework for evaluating land models. The framework includes four major components: (1) defining model aspects to be evaluated, (2) selecting benchmarks as standardized references to test models, (3) developing a scoring system to measure model performance skills, and (4) stimulating model improvement. (Adopted from Luo et al., 2012).

Chapter 2). However, a land model typically simulates hundreds of biophysical, biogeochemical, and ecological processes at regional and global scales over hundreds of years. It would be unrealistic to undertake validation of so many processes at all spatial and temporal scales, even if observations were available. The complex behavior of these interacting processes can be realistically understood only if we holistically assess land models and their major components. Benchmark analysis is an approach that has been recently developed to evaluate the performance of land models.

Benchmark analysis is a standardized evaluation of one system's performance against defined reference data (i.e., benchmarks) that can be used to diagnose the system's strengths and deficiencies for future improvement (Luo et al., 2012). Benchmark analysis has been recently applied to evaluate land models against observations (Collier et al., 2018). A benchmark analysis has four elements: (1) targeted aspects of model performance to be evaluated; (2) benchmarks as defined reference data to evaluate model performance; (3) a scoring system of metrics to measure relative performance among models; and (4) evaluated performance of models and future improvement (Figure 19.1).

ASPECTS OF LAND MODELS TO BE EVALUATED

Land models typically simulate many processes. Although individual studies may assess only a few aspects of model performance, a comprehensive benchmark analysis is required to evaluate all these major components when land models are integrated with ESMs. The performance of a model should be evaluated for its baseline simulations over broad spatial and temporal scales, and modeled responses of land processes to global change.

The baseline state for biogeochemical cycles includes simulated global totals, spatial distributions, and temporal dynamics of gross primary production, net primary production, vegetation and soil carbon stocks, ecosystem respiration, litter production, litter mass, and net ecosystem production. For example, the International Land Model Benchmarking (ILAMB) project evaluated biomass, burned area, gross primary productivity, leaf area index, global net ecosystem carbon balance, net ecosystem exchange, ecosystem respiration, and soil carbon (Collier et al., 2018).

To reliably predict future states of ecosystems under a changing environment, land models have to realistically simulate responses of land processes to disturbances and global change. Major global change factors include rising atmospheric CO_2 concentration, increasing land use and surface air temperature, altered precipitation amounts and patterns, and changing nitrogen (N) deposition. The direct effects of these global change factors are relatively easily benchmarked since we have direct knowledge of how ecosystems respond to rising atmospheric CO_2 concentration, increasing temperature, altered precipitation, and changing nitrogen deposition. However, indirect effects of these factors on ecosystem carbon processes are not well understood, although many field experiments have been conducted. Thus, it is more difficult to benchmark model performance in predicting future states of ecosystems.

REFERENCE DATA SETS AS BENCHMARKS

A comprehensive benchmarking analysis usually uses a set of benchmarks, against which land model performance can be evaluated (Table 19.1). Benchmarks could consist of direct observations, results from manipulative experiments, data-model products, or data-derived functional relationships. Direct observations and experimental results are generally accepted to be the most reliable benchmarks for model performance and are typically referred to as reference data. Reference data that are often used for benchmarking biogeochemical cycle models include global data products of gross primary production (GPP), net primary production (NPP), soil respiration, ecosystem respiration, plant biomass, and soil carbon. When they are used in a benchmarking analysis, reference data sets are usually assessed and weighted for their degree of certainty, scale appropriateness, and overall importance of the constraint or process to model predictions (Collier et al., 2018). The ILAMB project evaluates eight variables using a variety of reference data as listed in Table 19.1.

Land models can also be evaluated on their simulated variable-to-variable relationships in comparison with relationships in observations. For example, model representations of the relationships that GPP exhibits with precipitation,

TABLE 19.1

Reference data sets used to measure ecosystem and carbon cycle performance

Variables	Reference data sets	Description
Biomass	Tropical (Saatchi et al., 2011)	forest carbon stocks in tropical regions across three continents
	NBCD2000 (Kellndorfer et al., 2013)	aboveground biomass and carbon baseline data in north America
	USForest (Blackard et al., 2008)	U.S. forest biomass
Burned area	GFED4S (Giglio et al., 2010)	variability and long-term trends in burned area
GPP	Fluxnet (Lasslop et al., 2010)	net ecosystem exchange, photosynthesis, and respiration
Leaf area index	AVHRR (Myneni et al., 1997)	global land cover, LAI and FPAR
	MODIS (De Kauwe et al., 2011)	leaf area index product for a region of mixed coniferous forest
Global NECB	GCP (Le Quéré et al., 2016)	global carbon budget 2016
Net ecosystem exchange	Fluxnet (Lasslop et al., 2010)	net ecosystem exchange, photosynthesis, and respiration
Ecosystem respiration	Fluxnet (Lasslop et al., 2010)	net ecosystem exchange, photosynthesis, and respiration
Soil carbon	HWSD (Todd-Brown et al., 2013)	Harmonized World Soil Data
	NCSCDV22 (Hugelius et al., 2013)	organic carbon storage to 3m depth in soils of the northern circumpolar permafrost region.

NECB = net ecosystem carbon balance

evapotranspiration, and temperature are often assessed. Such variable-to-variable relationships are quantified over a time period from reference data sets and used as benchmarks for the relationships diagnosed in models. This approach is particularly effective to understand the consistency between the observed and simulated sensitivity of ecosystem responses to climate change.

BENCHMARKING METRICS

A comprehensive benchmarking study usually uses a suite of metrics across several variables to holistically assess model performance at the relevant spatial and temporal scales. Many statistical measures are available to quantify mismatches between multiple modeled and observed variables. Five metrics were developed for ILAMB to evaluate model performance. The five metrics are to measure bias, root-mean-square-error (RMSE), phase shift, interannual variability, and spatial distributions (Collier et al., 2018).

The bias measures differences between the mean value of the reference data and that of the model over the same time period and the same spatial area. For example, the bias of gross primary productivity between the reference data and the model (e.g., Community Land Model version 4.5, CLM4.5) is calculated between their respective means in each grid cell where both reference data and modeled values are available. To account for the bias due to the variability at any given spatial location, the bias is nondimensionalized as a relative error to measure the bias score.

RMSE is computed as the square root of the mean square error between modeled values and the reference data over a time period. The RMSE is normalized by the centralized RMSE of the reference data set to get a relative error as a score. By scoring the centralized RMSE, the bias is removed from the RMSE, allowing the RMSE score to be focused on an orthogonal aspect of model performance.

The phase shift is evaluated for 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 at each spatial cell across the time period of the reference data set. The phase shift is calculated as the difference between the reference and model data sets by subtracting their respective maximum values in days.

The interannual variability in model simulations is evaluated by removing the annual cycle from both the reference data and the model. A score is then computed as a function of their differences over space.

The spatial distribution of any time-averaged variable is evaluated by computing the standard deviation of modeled values over space normalized by the standard deviation of the reference data. The spatial correlation is also calculated for the period mean values of reference data and modeled values. A score is assigned by the penalty for large deviation of the normalized standard deviations and the spatial correlation from a value of 1.

The overall score for a given variable and data product is a weighted sum of the five metrics, producing a single scalar score for each variable for every model or model version. Readers who are interested in details of these metrics may study the paper by Collier et al. (2018).

PERFORMANCE OF THREE CLM VERSIONS AND FUTURE IMPROVEMENTS

The metrics for bias, RMSE, seasonal cycle phase, spatial distribution, interannual variability, and variable-to-variable assessments were applied to evaluate three CLM versions (CLM4 *vs.* CLM4.5 *vs.* CLM5) under two forcing data sets (GSWP3v1 *vs.* CRUNCEPv7) (Lawrence et al., 2019). The quality of the simulations across model generations was found to be generally improving. CLM5 outperforms CLM4 for the majority of assessed variables (Figure 19.2). The improvements from CLM4.5 to CLM5 were relatively subtle in that several variables show improvement (e.g., biomass, burned area, LAI, net ecosystem carbon balance, net ecosystem exchange, and ecosystem respiration) but others show degradation (e.g., soil carbon).

The functional relationships were also assessed between two variables (e.g., precipitation *vs.* GPP or LAI) (Figure 19.3). CLM5 performed better than CLM4 or CLM4.5 for the relationships between GPP and climate variables. However, the relationship between GPP and surface air temperature slightly degraded from CLM4.5 to CLM5.

The ILAMB benchmark analysis provides some insights into model development. An improvement or degradation trend between two CLM versions can result from a mix of scores for individual metrics. The degradation in the simulations of

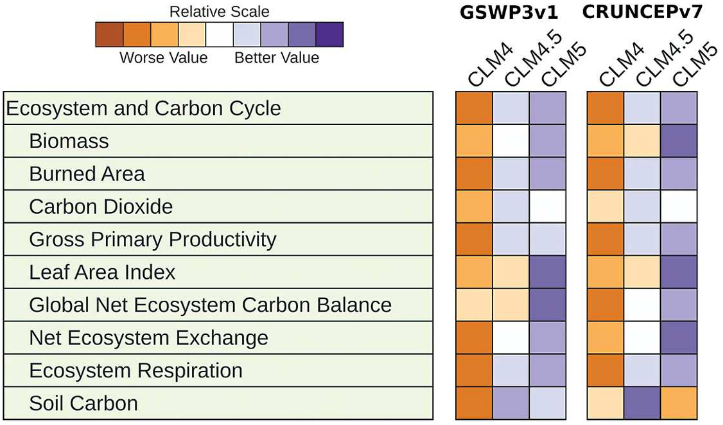

Figure 19.2. Evaluation of performance of CLM4, CLM4.5, and CLM5 under two sets of forcing, GSWP3v1 and CRUNCEPv7. A stoplight color scheme is used to indicate aggregate performance for each model by variable. (Adopted from Lawrence et al., 2019).

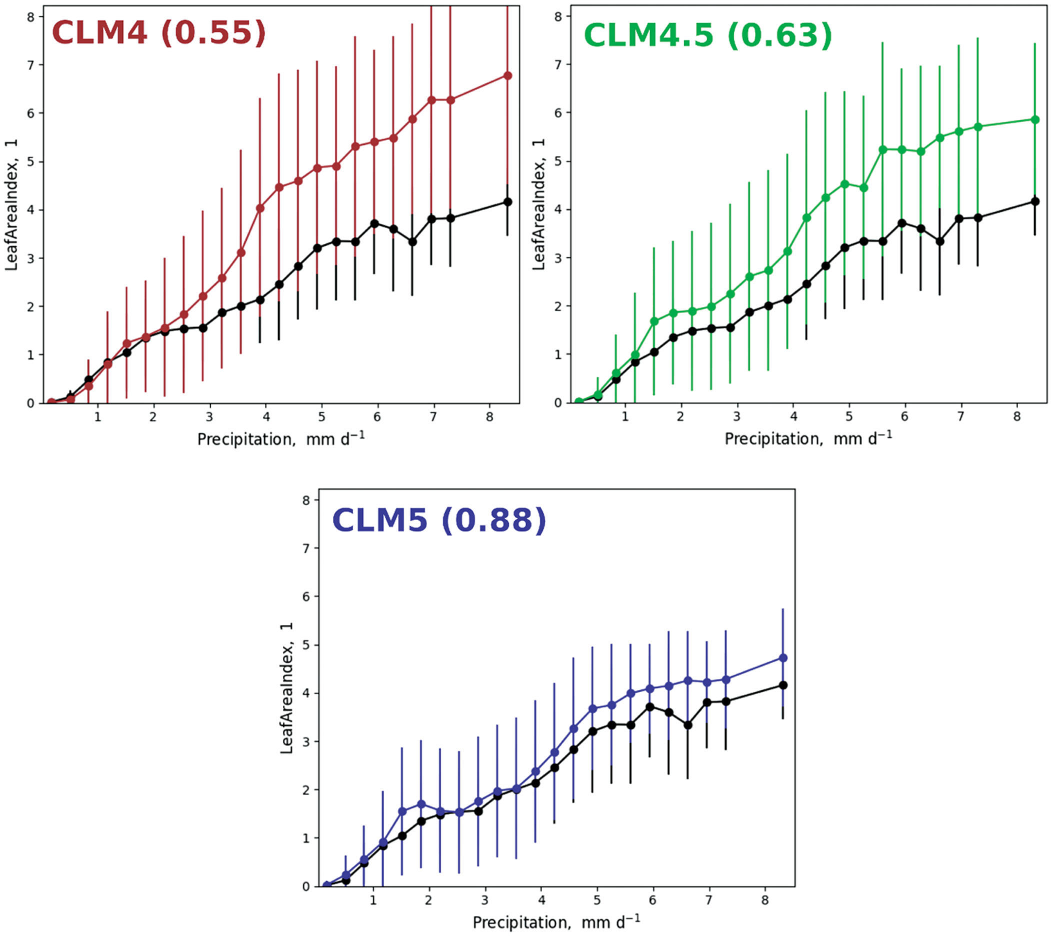

Figure 19.3. Variable-to-variable comparison between annual precipitation and LAI for CLM4, CLM4.5, and CLM5 under the GSWP3v1 forcing. The black line is the observationally derived relationship. Error bars indicate the ±1 standard deviation of LAI for all grid cells that lie within that precipitation bin. Values in parentheses are the scores for that comparison. (Adopted from Lawrence et al., 2019).

soil carbon stocks from CLM4.5 to CLM5 is partially due to high uncertainty in the observational estimates. Another metric evaluates the models against apparent soil carbon turnover time, showing an improvement from CLM4.5 to CLM5. The disagreement between two metrics of soil carbon may suggest the need for future improvement of observationally constrained estimates.

Model performance depends on three elements: model structure, parameterization, and forcing (see Chapters 21 and 33). The model structure that simulates soil carbon dynamics in CLM is primarily based on first-order kinetics. Although this model structure has been questioned, almost all data sets from studies of litter decomposition and soil incubation suggest the structure may be highly reliable (Chapter 1). Model parameterization is likely the main cause of the model-data mismatch. Chapter 37 discusses methods to improve model parameterizations of CLM5 to improve model performance.

CONCLUSIONS

A four-component benchmark analysis was outlined: (1) identification of aspects of models to be evaluated; (2) selection of benchmarks as standardized references to evaluate models; (3) a scoring system to measure model performance skills; and (4) evaluation of model performance to inform model improvement. The International Land Model Benchmarking (ILAMB) project has developed an open-source model benchmarking software package to score model performance. ILAMB has developed a suite of reference data sets as benchmarks, five metrics plus variable-to-variable relationships as the scoring system to evaluate models or model versions. The ILAMB package has been applied to perform comprehensive model assessment across a wide range of land variables. Such benchmark analysis offers insights into strengths and weaknesses of different models or model versions for identifying future improvements.

SUGGESTED READING

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

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

QUIZZES

1. What are the similarities and differences between model validation and benchmark analysis?
2. How does benchmark analysis evaluate model performance?
3. What variables in carbon cycle models would you choose to be evaluated by a benchmark analysis?
4. What data sets do you think would be important to be used as benchmarks to evaluate models?
5. What five metrics does the ILAMB package use to score model performance?