Biogeosciences Discuss., 9, 1899–1944, 2012
www.biogeosciences-discuss.net/9/1899/2012/
doi:10.5194/bgd-9-1899-2012

Biogeosciences Discussions

A framework of benchmarking land models

Y. Q. Luo[1], J. Randerson[2], G. Abramowitz[3], C. Bacour[4], E. Blyth[5], N. Carvalhais[6,7], P. Ciais[8], D. Dalmonech[6], J. Fisher[9], R. Fisher[10], P. Friedlingstein[11], K. Hibbard[12], F. Hoffman[13], D. Huntzinger[14], C. D. Jones[15], C. Koven[16], D. Lawrence[10], D. J. Li[1], M. Mahecha[6], S. L. Niu[1], R. Norby[13], S. L. Piao[17], X. Qi[1], P. Peylin[8], I. C. Prentice[18], W. Riley[16], M. Reichstein[6], C. Schwalm[14], Y. P. Wang[19], J. Y. Xia[1], S. Zaehle[6], and X. H. Zhou[20]

[1]Department of Botany and Microbiology, University of Oklahoma, Norman, OK 73019, USA
[2]Department of Earth System Science, University of California, Irvine, CA 92697, USA
[3]Climate Change Research Centre, University of New South Wales, Sydney, Australia
[4]Laboratory of Climate Sciences and the Environment, Joint Unit of CEA-CNRS, Gif-sur-Yvette, France
[5]Centre for Ecology and Hydrology, Wallingford, Oxfordshire, OX10 8BB, UK
[6]Max-Planck-Institute for Biogeochemistry, Jena, Germany
[7]Departamento de Ciências e Engenharia do Ambiente, DCEA, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
[8]Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ, CE Orme des Merisiers, 91191 Gif sur Yvette, France

[9]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA
[10]National Centre for Atmospheric Research, Boulder, CO, USA
[11]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, UK
[12]Pacific Northwest National Laboratory, Richland, WA, USA
[13]Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[14]School of Earth Science and Environmental Sustainability, Northern Arizona University, Flagstaff, AZ 86011, USA
[15]Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK
[16]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[17]Department of Ecology, Peking University, Beijing 100871, China
[18]Department of Biological Sciences, Macquarie University, NSW 2109 Sydney, Australia
[19]CSIRO Marine and Atmospheric Research PMB #1 and Centre for Australian Weather and Climate Research, Aspendale, Victoria 3195, Australia
[20]Research Institute of the Changing Global Environment, Fudan University, 220 Handan Road, Shanghai 200433, China

Received: 11 January 2012 – Accepted: 23 January 2012 – Published: 17 February 2012

Correspondence to: Y. Luo (yluo@ou.edu)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Land models, which have been developed by the modeling community in the past two decades to predict future states of ecosystems and climate, have to be critically evaluated for their performance skills of simulating ecosystem responses and feedback to climate change. Benchmarking is an emerging procedure to measure and evaluate performance of models against a set of defined standards. This paper proposes a benchmarking framework for evaluation of land models. The framework includes (1) targeted aspects of model performance to be evaluated; (2) a set of benchmarks as defined references to test model performance; (3) metrics to measure and compare performance skills among models so as to identify model strengths and deficiencies; and (4) model improvement. Component 4 may or may not be involved in a benchmark analysis but is an ultimate goal of general modeling research. Land models are required to simulate exchange of water, energy, carbon and sometimes other trace gases between the atmosphere and the land-surface, and should be evaluated for their simulations of biophysical processes, biogeochemical cycles, and vegetation dynamics across timescales in response to both weather and climate change. Benchmarks that are used to evaluate models generally consist of direct observations, data-model products, and data-derived patterns and relationships. Metrics of measuring mismatches between models and benchmarks may include (1) a priori thresholds of acceptable model performance and (2) a scoring system to combine data-model mismatches for various processes at different temporal and spatial scales. The benchmark analyses should identify clues of weak model performance for future improvement. Iterations between model evaluation and improvement via benchmarking shall demonstrate progress of land modeling and help establish confidence in land models for their predictions of future states of ecosystems and climate.

1 Introduction

Over the past two 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 (Oleson, 2010; Wang et al., 2010; Zaehle et al., 2010). They also simulate vegetation dynamics (Sitch et al., 2003) and disturbances (Thonicke et al., 2010). When coupled to ESMs, land models now allow simulation of land-atmosphere physical interactions (Bonan, 2008) and carbon-climate feedback (Bonan and Levis, 2010; Friedlingstein et al., 2006). These models are now widely used for policy relevant assessment of climate change and its impact on ecosystems or terrestrial resources, and more recently on allowable anthropogenic CO_2 emissions compatible with a given concentration pathway (Arora et al., 2011). However, there is still very limited knowledge of the performance skills of these land models, especially when embedded in ESMs. Without quantification of the performance skills of land models, their prediction of future states of ecosystems and climate cannot be widely accepted.

Model performance has traditionally been evaluated via comparison with common knowledge, observed data sets, and other models. “Validation” against observed data is traditionally the most common approach to model evaluation (Oreskes, 2003; Rykiel, 1996). However, a land model typically simulates hundreds or thousands of biophysical, biogeochemical, and ecological processes at regional and global scales over hundreds of years. It would be unrealistic to expect validation of so many processes at all spatial and temporal scales independently, even if observations were available. The complex performance behavior of these related processes can only be realistically understood if we holistically assess the land models and their major components. As a consequence, there have been many international model intercomparison projects. For example, the Project for Intercomparison of Land surface Parameterization Schemes (PILPS) focused on simulation of the water and energy balance (Pitman, 2003). The

Carbon Cycle Model Linkage Project (CCMLP) evaluated simulation of the terrestrial carbon cycle (McGuire et al., 2001). The Coupled Carbon Cycle Climate Model Intercomparison Project (C4MIP) compared simulation of the climate-carbon cycle coupling among 11 models (Friedlingstein et al., 2006). Nevertheless, there have been very few, if any, attempts to systematically evaluate land models against data from a range of observation networks and experiments in a comprehensive, objective and transparent manner.

The International Land Model Benchmarking (ILAMB) project (http://www.ilamb.org/) has recently been launched to promote model-data comparison to evaluate and improve the performance of land models. ILAMB aims to (1) develop internationally accepted benchmarks for land model performance, (2) promote the use of these benchmarks by the international community for model comparison, (3) strengthen linkages between experimental, remote sensing, and climate modeling communities, (4) design new model tests, and (5) support the design and development of a new, open source, benchmarking software system for use by the international community. ILAMB has the potential to stimulate observation and experimental communities to design new measurement campaigns to improve models and reduce uncertainties associated with key processes in land models.

This paper was a result of discussion during the second ILAMB workshop held in Irvine, California, USA, on 24–26 January 2011. The workshop participants agreed that the community needs to clearly define terms related to benchmark analysis and specify a general framework of benchmarking to facilitate communication among practitioners in this area of research, as well as with those who are entering into this field of research (e.g. students, post-doctoral fellows, and other scientists). This paper first defines benchmark analysis and presents a framework for its interpretation, which consists of four major components. We then examine each of the four components: targeted aspects of land models to be evaluated; defined benchmarks against which model performance skills can be effectively evaluated; metrics to measure model performances, and; approaches to identify model deficiencies for future improvement.

BGD
9, 1899–1944, 2012
A framework of benchmarking land models
Y. Q. Luo et al.

Specifically, we highlight benchmarks to evaluate biophysical processes, hydrological, biogeochemical cycles and vegetation dynamics. To identify model deficiencies, we also discuss a variety of metrics to evaluate performance of different models and approaches.

2 Benchmark analysis and its general framework

In a general sense, benchmark analysis is a standardized evaluation of one system's performance against defined references (i.e. benchmarks) that can be used to diagnose the system's strengths and deficiencies for future improvement. Benchmark analyses have been widely applied in economics, meteorology, computer sciences, business, and engineering. In business, for example, benchmark analysis provides a systematic approach to improving production efficiency and profitability through identifying, understanding, and adapting the successful business practices and processes used by other companies in terms of quality, time and cost (Fifer, 1988). In engineering, benchmark analysis is used to measure efficiency, productivity, and quality against a reference or benchmark performance of a stanardized instrument (Jamasb and Pollitt, 2003). In meteorology, benchmark analysis facilitates testing the accuracy, efficiency, and efficacy of meteorological model formulations and assumptions against measurements (Bryan and Fritsch, 2002). In computer sciences, benchmark analysis is used to examine the performance of a processor, code structure, features of processor architecture, and optimization of compiler against a number of standard tests to gain insight into how the processor or code can be improved to handle various applications (Simon and McGalliard, 2009; Ghosh and Sonakiya, 1998).

Benchmark analysis is urgently needed to evaluate land models against observations and experimental manipulations as it allows us to identify uncertainties in predictions as well as guiding the priorities for model development (Blyth et al., 2011). Several smaller-scale land model evaluation studies have been attempted. For example, the Carbon-LAnd Model Intercomparison Project (C-LAMP) was conducted to evaluate two

BGD
9, 1899–1944, 2012

biogeochemistry models that are integrated within the Community Land Model (CLM) – Carnegie-Ames-Stanford Approach′ (CASA′) and carbon-nitrogen (CN) against nine different classes of observations (Randerson et al., 2009). The Joint UK Land Environment Simulator (JULES) was evaluated for its performances against surface energy flux measurements from 10 flux network (FLUXNET) sites with a range of climate conditions and biome types (Blyth et al., 2011). Three global models of the coupled carbon-climate system were evaluated against atmospheric CO_2 concentration from a network of stations to quantify each model's ability to reproduce the global growth rate, the seasonal cycle, the El Niño – Southern Oscillation (ENSO) – forced interannual variability of atmospheric CO_2, and the sensitivity to climatic variations (Cadule et al., 2010). The evaluation procedures so far are often carried out in largely "ad-hoc" ways, and done as a matter of personal preference without much coordination among groups.

To effectively evaluate land model performance skills, we need to develop a widely accepted, consistent and comprehensive framework for benchmark analysis. Land models typically simulate thousands of processes related to energy balance, hydrological cycles, biogeochemical cycles, and vegetation dynamics. It is impossible to independently evaluate each of the modeled processes. We have to develop integrative, holistic approaches to understand and assess the complex behavior of these models and major components. Also, a land model is a multidisciplinary product. Evaluation of such a model requires a framework that enables communication among disciplines. In addition, numerous data sets are needed from many research areas to evaluate various aspects of the land models. Organization of those heterogeneous data sets to effectively evaluate land models requires a systems approach with assistance of ecological informatics. Moreover, models simulate long-term and large-scale phenomena. To date, few data sets can match the temporal and spatial scales of global and regional model simulations. We need standardized methods to measure mismatches between models and data given their temporal and spatial characteristics.

A comprehensive benchmarking framework has at least four elements: (1) targeted aspects of model performance to be evaluated; (2) benchmarks as defined references

BGD
9, 1899–1944, 2012

to evaluate model performance; (3) a scoring system of metrics to measure relative performances among models; and (4) diagnostic approaches to identification of model strengths and deficiencies for future improvement (Fig. 1). First, a land model typically simulates biophysical processes, hydrological processes, biogeochemical cycles, and vegetation dynamics. For each of the component processes, the land model has to represent basic system dynamics well (i.e. baseline simulation) and simulate their responses and feedback to climate change and disturbances (i.e. response simulation). Any benchmark analysis has to be clear on what aspects of the land models are evaluated. Second, the most critical component of any benchmark analysis is to define benchmarks. Benchmarks could be composed of direct observations; results from manipulative experiments; derived functional relationships and patterns from observations (e.g. water-use efficiency, phase lags between forcing and predicted ecosystem responses, Bowen ratio), and data model products (i.e. data-based model output). Third, a scoring system is needed to set criteria for a model to pass the benchmark test and measure relative performance among models. Fourth, benchmark analysis should identify needed model improvements and areas where the model is sufficiently robust for accurate simulations. It is challenging to identify model deficiencies in structure and parameters based upon diagnosis of poor performance at various temporal and spatial scales. The four elements of the benchmarking framework are discussed in detail in the following sections.

3 Aspects of land models to be evaluated via benchmarking

Land models typically simulate the surface energy balance, hydrological processes, biogeochemical cycles, and vegetation dynamics. Although individual studies may evaluate one aspect of model performance, a comprehensive framework is required to evaluate all those major components. In addition, unlike models used for weather prediction, the type of land models we are discussing are usually designed to predict longer-term future states of ecosystems and climate. The performance of a model

BGD
9, 1899–1944, 2012

should therefore be evaluated for its baseline simulations over broad spatial and temporal scales, and include evaluation of modeled responses and feedbacks of land processes to global change and disturbances.

Scientists have to establish some level of confidence in land models' baseline simulations before they can be used to study ecosystem responses and feedback to climate change. Baseline simulations of biogeochemical cycles include simulated global totals, spatial distributions, and temporal dynamics of gross primary production, net primary production, vegetation and soil carbon content, ecosystem respiration, litter production, litter mass, net ecosystem production at some reference climatic conditions, and land use and cover patterns. The reference climate conditions usually are reanalysis climate data of 30–50 yr that are used for model spin-up. The baseline simulations of biophysical processes include global totals, spatial distributions, and temporal dynamics of radiation fluxes (latent and sensible heat fluxes, Bowen ratio), evaporation, transpiration, and runoff. The baseline simulations of vegetation dynamics include preindustrial vegetation pattern or change in vegetation distribution over the last 5000 to 10 000 yr. Most baseline simulations are verified against common knowledge and evaluated against benchmarks, for example, for their representation of diurnal and seasonal variations (Fig. 2).

To reliably predict future states of ecosystems under a changed environment, land models have to realistically simulate responses of land processes to disturbances and global change. Natural and anthropogenic disturbances can significantly alter biogeochemical processes, biophysical properties, and vegetation dynamics. Several land models have incorporated algorithms to simulate individual events of fire and land use changes (Thonicke et al., 2010; Prentice et al. 2011). Natural disturbances occur at different frequencies with varying severity on diverse spatial scales in different regions and thus can be characterized by disturbance regimes. Climate change can regulate and, in turn, be affected by disturbance regimes. How to simulate and benchmark the responses and feedback of disturbance regimes to climate change still remains a great challenge.

BGD
9, 1899–1944, 2012

Major global change factors include rising atmospheric CO_2 concentration, increasing land use, surface air temperature, altered precipitation amounts and patterns, and nitrogen (N) deposition. Most land models use the Farquhar leaf photosynthesis model (Farquhar et al., 1980) and its variants to simulate instantaneous increases in carbon influx in response to increasing [CO_2] but there is much greater variation in the extent to which current models account for long-term acclimation of photosynthetic and respiratory parameters. Almost all land models simulate ecosystem responses to climate warming primarily via the kinetic sensitivity of photosynthesis and respiration to temperature and have not fully considered warming-induced changes in phenology and the length of growing seasons, nutrient availability, ecosystem water dynamics and species composition (Luo, 2007). Precipitation changes in its frequency, intensity, amount, and spatial distributions as predicted by climate models. Each of those changes has different effects on ecosystems (Knapp et al., 2008), which are usually represented by response functions that are either directly linked to precipitation or indirectly through soil moisture dynamics in land models. A few global land models have been designed to simulate ecosystem responses to nitrogen deposition (Thornton et al., 2007; Wang et al., 2010; Zaehle et al., 2010), mainly via its simulation of plant growth, but not many indirect effects of nitrogen on ecosystem structure and function or long-term changes in nitrogen capital have been in included (Lu et al., 2011b; Yang et al., 2011).

Feedbacks occur among land processes themselves and between ecosystems and the atmosphere. For example, soil nitrogen availability influences leaf area expansion, plant growth, and ecosystem carbon cycle. Carbon sequestration in plant biomass and soil feeds back not only to short-term mineral nitrogen availability but potentially also stimulates long-term accumulation of nitrogen capital in ecosystems (Luo et al., 2006). Nitrogen availability may also influence albedo (Ollinger et al., 2008) and thus land surface energy and water balances. The latter feed back to climate change. There are numerous feedback processes within land models and in their coupling with climate models. However, it is not straightforward to disentangle these processes and or therefore to evaluate feedback mechanisms in benchmark analysis.

BGD

9, 1899–1944, 2012

4 Benchmarks as defined references

A comprehensive benchmarking framework has a set of defined benchmarks, against which land models are evaluated (Table 1). Different benchmarks are chosen to evaluate different aspects of land models performance. The subsections below discuss what are available and of relevance for the various land model processes of interest.

4.1 Types of benchmarks

The benchmarks could include direct observations or ground-based measurements (Mittelmann and Preussner, 2006), results from manipulative experiments, data-model products, and derived functional relationships or patterns from data (Table 1). Direct observations and experimental results reflect recorded states of ecosystems when the measurements were made and are generally accepted to be the most reliable benchmarks model performance. Direct measurements include atmospheric CO_2 concentration, biomass, species composition, streamflow, snow cover and soil water content. Comparisons with models need to recognize that most direct measurements have had some levels of processing, up-scaling, allometry, and assumptions to generate the final estimates. For example, biomass data of trees are usually derived from allometric equations being applied to actual measured diameter at breast height and tree height (Chave et al., 2005). Values of normalized difference vegetation index (NDVI) are derived from remotely sensed measurements of light reflectance in the red and near infrared wavelength regions (Carlson and Ripley, 1997).

Direct measurements are usually made at specific points of time and space. Evaluating land model performance over the globe and hundreds of years needs benchmarks with extensive spatiotemporal representations of many processes (Sitch et al., 2008). Data-model products with well-quantified errors, which are generated according to some functional relationships to extend data's spatial and temporal scales via interpolation and extrapolation, can become useful for benchmarking. For example, the estimates of a global dataset of gross primary production and latent heat fluxes

BGD

9, 1899–1944, 2012

from eddy flux towers have been used to improve a global land surface model (Bonan et al., 2011). ET that is derived from remote sensing measurements of various energy components together with the energy balance equation (Fisher et al., 2008; Mu et al., 2007; Vinukollu et al., 2011) offers broad spatial and long temporal data sets benchmark analysis.

Land models can also be evaluated on their simulated patterns or relationships instead of absolute values of particular variables against benchmarks. This approach is particularly effective when uncertainties in data due to both random and systematic errors are unknown. For example, the south-north increase in the amplitude of the seasonal cycle in atmospheric CO_2 (Prentice et al., 2000) and latitudinal gradients in the satellite observed fraction of absorbed radiation (Zaehle et al., 2010) both give information about the geographic distribution of vegetation production. Similarly, the spatial relationship between annual net primary production and annual precipitation in a global network of monitoring stations provides more information about the sensitivity of NPP to climate than a comparison of these data on the basis of vegetation types (Randerson et al., 2009; Fig. 3). Correlations between El Niño related climate anomalies and growth rate of atmospheric CO_2 can be used to examine consistency between the observed and simulated ecosystem responses to climate change (Cadule et al., 2010; Fig. 4).

Model performance is also sometimes evaluated against standardized simulation results of a well-accepted model (Dai et al., 2003), the model ensemble mean (Chen et al., 1997), or statistically based-model results (Abramowitz, 2005). For example, a statistically based artificial neural network has been used to compare the performance of process-based land models (Abramowitz, 2005). Their analysis found that none of the tested land models performed better than the statistical model to reproduce observed carbon fluxes. The statistical model results can be used to define a benchmark level of performance that land models can be targeted to achieve relative to the information contained in the meteorological forcing about the surface fluxes.

BGD
9, 1899–1944, 2012

4.2 Applying benchmarks in land model evaluation

Benchmarks are used to evaluate biophysical processes, biogeochemical cycles, and vegetation dynamics of land models. Exchange of water and energy between land surface and atmosphere exerts major influences on the global and regional climate. In general, the available net radiation at the land surface is partitioned into ground, sensible, and latent heat fluxes, which drive the hydrological cycle via latent heat flux. Benchmarking energy and water balances and partitioning are requires estimates of latent heat flux, surface albedo, runoff, surface temperature, and soil moisture. Examples of global-scale reference data sets are shown in Table 2. Manipulative experiments can also be used to evaluate modeled responses of water and energy to global change (Wu et al., 2011). Data sets from over 100 sites on soil and permafrost data and active layer depths from the Circumpolar Active Layer Monitoring (CALM; http://nsidc.org/data/ggd313.html) program (Brown et al., 2003) are useful for benchmarking high-latitude ecosystems.

Data sets that are often used for benchmarking biogeochemical cycles include atmospheric CO_2 records at the seasonal to decadal scale (Dargaville et al., 2002; Heimann et al., 1998), satellite data at seasonal or longer time scales (Blyth et al., 2010; Maignan et al., 2011; Randerson et al., 2009). Other available datasets for biogeochemical cycle benchmarking include global GPP, NPP, soil respiration, ecosystem respiration, plant biomass, litter pool, litter decomposition rates, and soil carbon data products (Table 3). Recently, better estimates of high-latitude soil carbon stocks have been assembled (Tarnocai et al., 2009). In addition, global change experiments offer the potential to benchmark biogeochemical cycle responses to elevated CO_2, warming, precipitation, and nitrogen fertilization or deposition (Table 3). Data sets of methane emissions at various sites have been used to test a methane model (Riley et al., 2011). Preference is always given, where possible, for longer time series data sets, as they offer the potential to detect how the land surface responds to low frequency modes of climate variation (e.g. Piao et al., 2011 on NDVI greening and browning in boreal areas). Data

BGD
9, 1899–1944, 2012

sets on nutrient cycling and state variable at site, regional, and global scales and can be used to benchmark global carbon-nitrogen models (Wang et al., 2010; Zaehle et al., 2010).

Free-air CO_2 enrichment (FACE) experiments are a good example of manipulative experiments that have provided useful benchmarks for land surface models. They provided integrative measures of ecosystem response to future concentrations of atmospheric CO_2 (e.g. NPP, N uptake, stand transpiration) over multiple years, as well as detailed descriptions of contributory processes (e.g. photosynthesis, fine-root production, stomatal conductance) (Norby and Zak, 2011). The LPJ model (Hickler et al., 2008) matched the NPP response to elevated CO_2 observed in four FACE experiments in temperate forests (Norby et al., 2005), which provided more confidence in predictions of response in other biomes. The average response of the 11 models in the C_4MIP project (Friedlingstein et al., 2006) was consistent with the FACE results, although individual models varied widely. Furthermore, the general agreement may have been spurious: the models did not include feedbacks through the N cycle (Friedlingstein et al., 2006), and the experiments may not have been run long enough for N feedbacks to downregulate NPP (Norby et al., 2010).

Vegetation dynamics are usually represented by the combination of 7–17 plant functional types (PFT) in land models. The composition and abundance of PFTs can either be prescribed as time-invariant fields or can evolve with time as results of vegetation dynamics or land use change. Although different land models have their own set of PFTs, pre-industrial vegetation types are very important for benchmarking model performance (Table 4). In addition, it is also critical to have datasets of vegetation responses to disturbance and global change. There are some limited data available for vegetation response to warming, N deposition, fire, and land use and change (Table 4).

Although extensive data sets are available for benchmarking land models, equifinality remains a major issue in model evaluation (Tang and Zhuang, 2008; Luo et al., 2009). That is, the available data streams are insufficient to constrain model parameterization (Weng and Luo, 2011; Wang et al., 2001; Carvalhais et al., 2010) or

BGD
9, 1899–1944, 2012

to distinguish between different modeling structures (Frank et al., 1998). The need to comprehensively represent processes often leads to increasing model complexity, posing under-constraint and over-parameterization problems (Oreskes, 2003). Increases in the number, type, and location of observations used in model calibration and evaluation would ideally mitigate the equifinality issue and better constrain parameterization. Therefore, effective benchmarks should draw upon a broad set of independent observations spanning multiple temporal and spatial scales to identify processes and dynamics for system characterization (Randerson et al., 2009; Wang and Barrett, 2003; Zhou and Luo, 2008).

5 Benchmarking metrics

When land models are evaluated against benchmark data sets, the choice of which measure of performance to use, and the spatial and temporal scale at which the measure applies can significantly affect the nature of results. Defining standard metrics is a key step in any benchmarking framework. There are many quantitative measures (e.g. continental scale daily RMSE, global mean annual deviation from observed values, and global monthly correlations) of mismatches between modeled and observed individual variables (Janssen and Heuberger, 1995; Smith and Rose, 1995). To rank model performance, the measures, or metrics, of model performances for individual variables may be normalized and combined via a scoring system to provide a synthetic skill score, often on a scale from zero (least skillful) to one (most skillful),.

To meet minimal requirements, the research community may decide upon a priori threshold levels of performance level before a benchmark analysis is conducted. Such a threshold may be justified according to criteria of why a model below the threshold is not acceptable. Such thresholds may be viewed as a necessary, but not sufficient condition for a fully functioning model because complex models may perform well on particular metrics due as a result of compensating errors (that is, getting the right answers for the wrong reasons).

BGD
9, 1899–1944, 2012

A comprehensive benchmarking study usually scrutinizes model performance from multiple perspectives. Thus, a suite of metrics across several variables is needed to quantitatively measure model performance at the relevant spatial and temporal scales at which the model operates (Abramowitz et al., 2008; Cadule et al., 2010; Randerson et al., 2009; Taylor, 2001). Several strategies have been pursued to localize and quantify data-model disagreement in both time and frequency domains and at different spatial scales (Cadule et al., 2010; Mahecha et al., 2010; Wang et al., 2011). Model-data disagreements can be evaluated separately for low frequency variations (including nonlinear trends), the phasing and amplitude of seasonality, and short-term stochastic variability (Mahecha et al., 2010; Wang et al., 2010). Much less has been done to measure the model performance against observed ecosystem responses and feedbacks from global change experiments that manipulate global change factors, such as elevated CO_2, climate warming, altered precipitation, and nitrogen deposition.

The ranking of land models should reflect the different purposes that land models have been built for. For instance, land surface models operating within mesoscale meteorology or weather forecast models must be particularly robust at simulating energy and moisture fluxes, while land models coupled to Earth system models must be good at capturing ecosystem responses to changes in atmospheric composition and climate over decadal to centennial time scales. Thus metrics that measure disagreements between simulated and observed energy and water fluxes should be weighed more in a mesoscale meteorological study than in a decadal to centennial climate change study.

Data uncertainty is another important factor for developing appropriate metrics to measure the performance of land models. Different data sets inherently have different levels of uncertainty, and indeed different levels of ability to quantify uncertainty. For example, even at the plot-scale plant biomass estimated from an allometrical relationship together with diameter at breast height usually has much smaller observational errors than measured soil respiration (Luo et al., 2003). With some global scale remotely sensed products the time and frequency of overpass, atmospheric transparency, as well as the models used to translate irradiances into, for example, soil moisture content

BGD
9, 1899–1944, 2012

(as well as uncertainties in the parameters associated with these models) can make uncertainty estimation very difficult and temporally variable. When benchmarks of multiple variables are used, individual variables are commonly normalized by their standard deviations to make them effectively comparable. The C-LAMP system (Randerson et al., 2009) gave metrics for model performance that depended on a qualitative assessment of the importance of the process being tested and the uncertainty in the reference data set. They used those combined metrics to rank the models and cautioned that the assessments were in some sense subjective. Schwalm et al. (2010) used Taylor skill, bias, and observational uncertainty to measure performance of 22 terrestrial ecosystem models against observations from 44 FLUXNET sites (Fig. 5)

There are many techniques that have been explored by the data assimilation research community to combine metrics of measuring mismatches of modeled variables with multiple observations for different processes with different data uncertainties at various temporal and spatial issues (Trudinger et al., 2007). Some of these techniques may be very useful for benchmark analysis. An essential procedure for data assimilation is to define a metric (e.g. cost function) that describes data-model mismatches using multiple observations (Table 5). Luo et al. (2003) used standard deviations of individual observations as weights for model mismatches with data sets whose absolute values differed by several orders of magnitude. That weighing method has been successfully used in regional data assimilation with spatially distributed data (Zhou and Luo, 2008). Other weighting functions used in multiple-variable metrics include a simple sum of mismatches between modeled and observed variables, the standard deviation of residuals after a preliminary run of the calculation, the average value of observations, a linear function of the observation values (Trudinger et al., 2007). Choices of weights used in multiple-variable metrics significantly alter the outcome of parameter estimation (Trudinger et al., 2007; Weng and Luo, 2011; Xu et al., 2006) and are expected to have a similar influence on evaluation of model performance skills in the benchmark analysis.

BGD
9, 1899–1944, 2012

6 The role of benchmarking in model improvement

One of the ultimate goals of a benchmark analysis is to provide clues for diagnosing systematic model errors and thereby aid model improvement, although it need not be an essential part of benchmarking activities. The clues for model improvement usually come from identified poor performances of a land model in its simulations of processes, functions, and/or structures of ecosystems at different temporal and spatial scales. Model improvement is usually implemented through changes in model structures, parameterization, initial values, or input variables.

The average physiological properties of plant functional types are traditionally conceived as model "parameters". Parameter error may therefore arise when the values chosen for model parameters do not correspond to true underlying values. Thus, model benchmarking against plant trait data sets might be useful in assessing whether model parameters fall within realistic ranges. Such data sets include the GLOPNET leaf trait data set (Reich et al., 2007; Wright et al., 2005), and the TRY dataset (Kattge et al., 2009). For example, the TRY data set provides probability density functions of photosynthetic capacity based on 723 data points for observed carboxylation capacity (Vcmax) and 1966 data points of observed leaf nitrogen. Implementing these new, higher, values of observationally constrained Vcmax in the CLM4.0 model resulted in a significant over-estimates of canopy photosynthesis, compared to estimates of photosynthesis scaled from FLUXNET observations (Bonan et al., 2011). The scale of the over-prediction of GPP ($\sim$500 g C m^{-2} yr^{-1}, between 30° and 60° latitude) identified some fundamental issues in the formulation of the canopy model in CLM4.0.

Model structure error arises when key causal dependencies in the system being modeled are missing or represented incorrectly in the model. Based on biogeochemical principles of carbon-nitrogen coupling, for example, Hungate et al. (2003) conducted a plausibility analysis to illustrate that carbon sequestration may be considerably overestimated without the inclusion of nitrogen processes (Fig. 6). Without the carbon-nitrogen feedback, models fail to capture the experimentally observed positive

BGD
9, 1899–1944, 2012

responses of NPP to warming in cool climates (Zaehle and Friend, 2010). Generally, model structure errors are likely to reveal themselves through sufficiently comprehensive benchmarking and usually cannot be resolved by tuning or optimizing parameter values (Abramowitz, 2005; Abramowitz et al., 2006, 2007). Nevertheless, over-parameterizations of related processes may mask structural model deficiencies. A poor representation of the seasonal cycle of heterotrophic respiration in high latitudes by the Hadley Centre model (Cadule et al., 2010) was caused by soil temperature becoming much too low in the winter. Simply improving the seasonal cycle by adjusting the temperature function of respiration would have given the right answer for the wrong reason and materially affected the sensitivity to future changes. By understanding the processes (too little insulation of soil temperatures by the snow pack) enabled tackling the error without changing the long-term sensitivity. The C-LAMP benchmark analysis of CLM-CASA' and CN against atmospheric CO_2 measurements, eddy-flux data, MODIS observations, and TRANSCOM results suggested the need to improve model representation of seasonal and interannual variability of carbon cycle (Fig. 2).

7 Discussion and conclusion

This paper proposed a four-component framework for benchmarking land models. The components are: identification of aspects of models to be evaluated; collation of benchmarks as standardized references to test models; a scoring system to measure model performance skills and to evaluate model strengths and deficiencies; and; a collection of ways that can utilize the first three components to generate model improvement. We now consider a few caveats and concerns.

The first issue is on model predictions vs. performance skills. While an increase in performance gained through benchmark analysis will likely lead to an increase in predictive ability of a model for short-range predictions, it might not be sufficient to guarantee improved long-term projections of ecosystem responses to climate change for at least three reasons. First, observations on past ecosystem dynamics cannot fully

BGD
9, 1899–1944, 2012
A framework of benchmarking land models
Y. Q. Luo et al.

constrain model responses to future climate conditions that have never been observed. Nevertheless, "comparing models and observations over a wide range of conditions increases the chance of capturing important nonlinearities and complex or contingent responses that may control future behavior" (Luo et al., 2011). Second, future states of land ecosystems are determined not only by internal processes but also by external forces. The latter dominates long-term land dynamics so that predictions are clearly bounded by scenario-based, what-if analysis. Embedding land models within Earth system models, however, can help assess feedbacks between internal processes of land ecosystems and various scenarios of climate and land use changes. Third, land ecosystems are more at dynamic disequilibrium than equilibrium states under directional climate change (Luo and Weng, 2011). Dynamic disequilibrium states of biogeochemical cycles can be defined by initial values, changes in element influxes, and altered residence times (Weng and Luo, 2011). Future disequilibrium states of land ecosystems can be better predicted if the benchmark analysis is designed to evaluate key model components that determine their predictive behavior.

The second issue is about the feasibility of a community-wide benchmarking system. Land model benchmarking has reached a critical juncture, with several recent parallel efforts to evaluate different aspects of model performance. One future direction that may minimize duplication of effort is to develop a community-wide benchmarking system supported by multiple modeling and experimental teams. For a community-wide system to function well, it will need to be built using open source software and using only freely available observations with a traceable lineage. The software system that can be used to diagnose impacts of model development, guide synthesis efforts, identify gaps in existing observations needed for model validation, and reduce the human capital costs of making future model-data comparisons (Randerson et al., 2009). This is the approach being taken by the International Land Model Benchmarking Project (ILAMB) that will initially develop benchmarks for CMIP5 models participating in the IPCC 5th Assessment Report. An expectation of the first ILAMB benchmark is that it will be modified and expanded for use in future model intercomparison projects. Ultimately, a robust benchmarking system, when combined with information on model feedback strengths, may reduce uncertainties associated with emissions estimates required for greenhouse gas stabilization over the 21st century or other future climate projections (Qu and Hall, 2007). Such an open source, community-wide platform for model-data intercomparison also speeds up model development and strengthens ties between modeling and measurement communities. Important next steps include the design and analysis of land use change simulations (in both uncoupled and coupled modes), and the entrainment of additional ecological and Earth system observations.

Thirdly, a comprehensive benchmarking framework needs to stimulate communication to broader audience. For the broad science community and the public, it provides a means to show that the representation of the key biological, chemical, and physical processes regulating biosphere-atmosphere exchange is improving. Within the Earth system science community, benchmarking enables model developers from different disciplines to quantitatively diagnose the impacts of new parameterizations and structures on land model performance. It also has the potential to strengthen ties between experimental and modeling communities and allow for more effective syntheses. Benchmarking would lead to closer scrutiny of key observational data sets, and provide information about where model uncertainty was high – thus guiding future data collection efforts. In parallel, synthesis effort such as the IPCC may be able to draw upon benchmarking analyses to identify whether feedback mechanisms that arise in various models are broadly consistent with available contemporary observations.

Lastly, benchmark analysis shares objectives and procedures with data assimilation in many ways (Table 5). Data assimilation is a formal approach to infuse data into models for improving parameterization and adjusting model structures (Peng et al., 2011; Raupach et al., 2005; Wang et al., 2009; Luo et al., 2011). Data assimilation projects a misfit between model and observed quantities in the space of parameters, and quantifies the level of constraints on each parameter with associated uncertainties. It provides quantitative information, instead of performance criteria that should be met in comparing model output with data, to decide that a model has a satisfactory behavior or not. But data assimilation is computationally very costly and, as a consequence, cannot be easily implemented to directly improve the comprehensive, global-scale land models. Combination of benchmarking and data assimilation may facilitate land model improvement. Benchmarking can be used to pinpoint model deficiencies, which can become targeted aspects of model to be improved via data assimilation.

Acknowledgements. ILAMB is sponsored by the Analysis, Integration and Modeling of the Earth System (AIMES) project of the International Geosphere-Biosphere Programme (IGBP). The ILAMB project has received support from NASA's Carbon Cycle and Ecosystems Program and US Dept. of Energy's Office of Biological and Environmental Research. Preparation of the manuscript by YL was financially supported by US National Science Foundation (NSF) grant DEB 0444518, DEB 0743778, DEB 0840964, DBI 0850290, and EPS 0919466. CDJ was supported by the Joint UK DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). SZ and DD were supported by the European Community's Seventh Framework Programme under grant agreement no. 238366 (Greencycles II).

References

Abramowitz, G.: Towards a benchmark for land surface models, Geophys. Res. Lett., 32, L22702, doi:22710.21029/22005gl024419, 2005.

Abramowitz, G., Gupta, H., Pitman, A., Wang, Y. P., Leuning, R., Cleugh, H., and Hsu, K. L.: Neural error regression diagnosis (NERD): A tool for model bias identification and prognostic data assimilation, J. Hydrometeorol., 7, 160–177, 2006.

Abramowitz, G., Pitman, A., Gupta, H., Kowalczyk, E., and Wang, Y.: Systematic Bias in Land Surface Models, J. Hydrometeorol., 8, 989–1001, 2007.

Abramowitz, G., Leuning, R., Clark, M., and Pitman, A.: Evaluating the Performance of Land Surface Models, J. Climate, 21, 5468–5481, 2008.

Arora, V. K., Scinocca, J. F., Boer, G. J., Christian, J. R., Denman, K. L., Flato, G. M., Kharin, V. V., Lee, W. G., and Merryfield, W. J.: Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases, Geophys. Res. Lett., 38, L05805, doi:05810.01029/02010gl046270, 2011.

Batjes, N. H.: Carbon and nitrogen stocks in the soils of Central and Eastern Europe, Soil Use Manage., 18, 324–329, 2002.
Bell, J. E., Sherry, R., and Luo, Y.: Changes in soil water dynamics due to variation in precipitation and temperature: An ecohydrological analysis in a tallgrass prairie, Water Resour. Res., 46, W03523, doi:10.1029/2009WR007908, 2010.
Blyth, E., Gash, J., Lloyd, A., Pryor, M., Weedon, G. P., and Shuttleworth, J.: Evaluating the JULES Land Surface Model Energy Fluxes Using FLUXNET Data, J. Hydrometeorol., 11, 509–519, 2010.
Blyth, E., Clark, D. B., Ellis, R., Huntingford, C., Los, S., Pryor, M., Best, M., and Sitch, S.: 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, Geosci. Model Dev., 4, 255–269, 2011.
Bonan, G. B.: Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests, Science, 320, 1444–1449, 2008.
Bonan, G. B. and Levis, S.: Quantifying carbon-nitrogen feedbacks in the Community Land Model (CLM4), Geophys. Res. Lett., 37, L07401, doi:07410.01029/02010gl042430, 2010.
Bonan, G. B., Lawrence, P. J., Oleson, K. W., Levis, S., Jung, M., Reichstein, M., Lawrence, D. M., and Swenson, S. C.: Improving canopy processes in the Community Land Model version 4 (CLM4) using global flux fields empirically inferred from FLUXNET data, J. Geophys. Res., 116, G02014, doi:02010.01029/02010jg001593, 2011.
Bond-Lamberty, B. P. and Thomson, A. M.: A Global Database of Soil Respiration Data, Version 1.0. Data set, available at: http://daac.ornl.gov, Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA, 2010.
Boyero, L., Pearson, R. G., Gessner, M. O., Barmuta, L. A., Ferreira, V., Graca, M. A. S., Dudgeon, D., Boulton, A. J., Callisto, M., Chauvet, E., Helson, J. E., Bruder, A., Albarino, R. J., Yule, C. M., Arunachalam, M., Davies, J. N., Figueroa, R., Flecker, A. S., Rarnirez, A., Death, R. G., Iwata, T., Mathooko, J. M., Mathuriau, C., Goncalves, J. F., Moretti, M. S., Jinggut, T., Lamothe, S., M'Erimba, C., Ratnarajah, L., Schindler, M. H., Castela, J., Buria, L. M., Cornejo, A., Villanueva, V. D., and West, D. C.: A global experiment suggests climate warming will not accelerate litter decomposition in streams but might reduce carbon sequestration, Ecol. Lett., 14, 289–294, 2011.
Brown, J., Hinkel, K., and Nelson, F.: Circumpolar Active Layer Monitoring (CALM) Program Network: Description and data, in: International Permafrost Association Standing Commit-

BGD
9, 1899–1944, 2012

A framework of benchmarking land models

Y. Q. Luo et al.

tee on Data Information and Communication (comp.), 2003, Circumpolar Active-Layer Permafrost System, Version 2.0, edited by: Parsons, M. and Zhang, T., National Snow and Ice Data Center/World Data Center for Glaciology, Boulder, CO, 2003.
Bryan, G. H. and Fritsch, J. M.: A benchmark simulation for moist nonhydrostatic numerical models, Mon. Weather Rev., 130, 2917–2928, 2002.
Cadule, P., Friedlingstein, P., Bopp, L., Sitch, S., Jones, C. D., Ciais, P., Piao, S. L., and Peylin, P.: Benchmarking coupled climate-carbon models against long-term atmospheric CO_2 measurements, Global Biogeochem. Cy., 24, Gb2016, doi:2010.1029/2009gb003556, 2010.
Carlson, T. N. and Ripley, D. A.: On the relation between NDVI, fractional vegetation cover, and leaf area index, Remote Sens. Environ., 62, 241–252, 1997.
Carvalhais, N., Reichstein, M., Ciais, P., Collatz, G. J., Mahecha, M. D., Montagnani, L., Papale, D., Rambal, S., and Seixas, J.: Identification of vegetation and soil carbon pools out of equilibrium in a process model via eddy covariance and biometric constraints, Glob. Change Biol., 16, 2813–2829, 2010.
Chave, J., Andalo, C., Brown, S., Cairns, M. A., Chambers, J. Q., Eamus, D., Folster, H., Fromard, F., Higuchi, N., Kira, T., Lescure, J. P., Nelson, B. W., Ogawa, H., Puig, H., Riera, B., and Yamakura, T.: Tree allometry and improved estimation of carbon stocks and balance in tropical forests, Oecologia, 145, 87–99, 2005.
Chen, T. H., Henderson-Sellers, A., Milly, P. C. D., Pitman, A. J., Beljaars, A. C. M., Polcher, J., Abramopoulos, F., Boone, A., Chang, S., Chen, F., Dai, Y., Desborough, C. E., Dickinson, R. E., Dümenil, L., Ek, M., Garratt, J. R., Gedney, N., Gusev, Y. M., Kim, J., Koster, R., Kowalczyk, E. A., Laval, K., Lean, J., Lettenmaier, D., Liang, X., Mahfouf, J.-F., Mengelkamp, H.-T., Mitchell, K., Nasonova, O. N., Noilhan, J., Robock, A., Rosenzweig, C., Schaake, J., Schlosser, C. A., Schulz, J.-P., Shao, Y., Shmakin, A. B., Verseghy, D. L., Wetzel, P., Wood, E. F., Xue, Y., Yang, Z.-L., and Zeng, Q.: Cabauw Experimental Results from the Project for Intercomparison of Land-Surface Parameterization Schemes, J. Climate, 10, 1194–1215, 1997.
Dai, A., Qian T., Trenberth K. E., and Milliman J. D.: Changes in continental freshwater discharge from 1948–2004, J. Climate, 22, 2773–2791, 2009.
Dai, Y., Zeng, X., Dickinson, R. E., Baker, I., Bonan, G. B., Bosilovich, M. G., Denning, A. S., Dirmeyer, P. A., Houser, P. R., Niu, G., Oleson, K. W., Schlosser, C. A., and Yang, Z.-L.: The Common Land Model, B. Am. Meteorol. Soc., 84, 1013–1023, 2003.
Dargaville, R. J., Heimann, M., McGuire, A. D., Prentice, I. C., Kicklighter, D. W., Joos, F., Clein,

BGD
9, 1899–1944, 2012

A framework of benchmarking land models

Y. Q. Luo et al.

J. S., Esser, G., Foley, J., Kaplan, J., Meier, R. A., Melillo, J. M., Moore, B., III, Ramankutty, N., Reichenau, T., Schloss, A., Sitch, S., Tian, H., Williams, L. J., and Wittenberg, U.: Evaluation of terrestrial carbon cycle models with atmospheric CO_2 measurements: Results from transient simulations considering increasing CO_2, climate, and land-use effects, Global Biogeochem. Cy., 16, 1092, doi:1010.1029/2001gb001426, 2002.

Dorigo, W. A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., van Oevelen, P., Robock, A., and Jackson, T.: The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements, Hydrol. Earth Syst. Sci., 15, 1675–1698, doi:10.5194/hess-15-1675-2011, 2011.

FAO/IIASA/ISRIC/ISSCAS/JRC.: Harmonized World Soil Database (version 1.1), FAO, Rome, Italy and IIASA, Laxenburg, Austria, 2009.

Farquhar, G. D., Caemmerer, S., and Berry, J. A.: A biochemical model of photosynthetic CO_2 assimilation in leaves of C_3 species, Planta, 149, 78–90, 1980.

Fifer, R. M.: Benchmarking beating the competition, A practical guide to benchmarking, Kaiser Associates, Vienna, Virginia, 1988.

Fisher, J. B., Tu, K. P., and Baldocchi, D. D.: Global estimates of the land-atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites, Remote Sens. Environ., 112, 901–919, 2008.

Frank, E., Wang, Y., Inglis, S., Holmes, G., and Witten, I. H.: Technical note: Using model trees for classification, Mach. Learn., 32, 63–76, 1998.

Frankenberg, C., Fisher, J. B., Worden, J., Badgley, G., Saatchi, S. S., Lee, J.-E., Toon, G. C., Butz, A., Jung, M., Kuze, A., and Yokota, T.: New global observations of the terrestrial carbon cycle from GOSAT: Patterns of plant fluorescence with gross primary productivity, Geophys. Res. Lett., 38, L17706, doi:10.1029/2011GL048738, 2011.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K. G., Schnur, R., Strassmann, K., Weaver, A. J., Yoshikawa, C., and Zeng, N.: Climate-Carbon Cycle Feedback Analysis: Results from the C4MIP Model Intercomparison, J. Climate, 19, 3337–3353, 2006.

Ghosh, A. K. and Sonakiya, S.: Performance analysis of acousto-optic digital signal processors using the describing function approach, IETE J. Res., 44, 3–12, 1998.

Gobron, N., Aussedat, O., Pinty, B., Taberner, M., and Verstraete, M. M.: Medium Resolution

BGD
9, 1899–1944, 2012

Imaging Spectrometer (MERIS) – An optimized FAPAR Algorithm – Theoretical Basis Document, Revision 3.0 (EUR Report No. 21386 EN), Institute for Environment and Sustainability, 2004.

Heimann, M., Esser, G., Haxeltine, A., Kaduk, J., Kicklighter, D. W., Knorr, W., Kohlmaier, G. H., McGuire, A. D., Melillo, J., Moore, B., Otto, R. D., Prentice, I. C., Sauf, W., Schloss, A., Sitch, S., Wittenberg, U., and Wurth, G.: Evaluation of terrestrial Carbon Cycle models through simulations of the seasonal cycle of atmospheric CO_2: First results of a model intercomparison study, Global Biogeochem. Cy., 12, 1–24, 1998.

Hickler, T., Smith, B., Prentice, I. C., Mjöfors, K., Miller, P., Arneth, A., and Sykes, M. T.: CO_2 fertilization in temperate FACE experiments not representative of boreal and tropical forests, Glob. Change Biol., 14, 1531–1542, 2008.

Hungate, B. A., Dukes, J. S., Shaw, M. R., Luo, Y. Q., and Field, C. B.: Nitrogen and climate change, Science, 302, 1512–1513, 2003.

Hurtt, G. C., Frolking, S., Fearon, M. G., Moore, B., Shevliakova, E., Malyshev, S., Pacala, S. W., and Houghton, R. A.: The underpinnings of land-use history: three centuries of global gridded land-use transitions, wood-harvest activity, and resulting secondary lands, Glob. Change Biol., 12, 1208–1229, 2006.

Jamasb, T. and Pollitt, M.: International benchmarking and regulation: an application to European electricity distribution utilities, Energ. Policy, 31, 1609–1622, 2003.

Janssen, P. H. M. and Heuberger, P. S. C.: Calibration of process-oriented models, Ecol. Model., 83, 55–66, 1995.

Janssens, I. A., Dieleman, W., Luyssaert, S., Subke, J. A., Reichstein, M., Ceulemans, R., Ciais, P., Dolman, A. J., Grace, J., Matteucci, G., Papale, D., Piao, S. L., Schulze, E. D., Tang, J., and Law, B. E.: Reduction of forest soil respiration in response to nitrogen deposition, Nat. Geosci., 3, 315–322, 2010.

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

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W.,

BGD
9, 1899–1944, 2012

A framework of benchmarking land models

Y. Q. Luo et al.

Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, J. Geophys. Res., 116, G00J07, doi:10.1029/2010JG001566, 2011.
Kattge, J., Knorr, W., Raddatz, T., and Wirth, C.: Quantifying photosynthetic capacity and its relationship to leaf nitrogen content for global-scale terrestrial biosphere models, Glob. Change Biol., 15, 976–991, 2009.
Knapp, A. K., Beier, C., Briske, D. D., Classen, A. T., Luo, Y., Reichstein, M., Smith, M. D., Smith, S. D., Bell, J. E., Fay, P. A., Heisler, J. L., Leavitt, S. W., Sherry, R., Smith, B., and Weng, E.: Consequences of More Extreme Precipitation Regimes for Terrestrial Ecosystems, Bioscience, 58, 811–821, 2008.
Kurz, W. A., Dymond, C. C., Stinson, G., Rampley, G. J., Neilson, E. T., Carroll, A. L., Ebata, T., and Safranyik, L.: Mountain pine beetle and forest carbon feedback to climate change, Nature, 452, 987–990, 2008a.
Kurz, W. A., Stinson, G., Rampley, G. J., Dymond, C. C., and Neilson, E. T.: Risk of natural disturbances makes future contribution of Canada's forests to the global carbon cycle highly uncerain, P. Natl. Acad. Sci. USA, 105, 1551–1555, 2008b.
Lefsky, M. A.: A global forest canopy height map from the Moderate Resolution Imaging Spectroradiometer and the Geoscience Laser Altimeter System, Geophys. Res. Lett., 37, L15401, doi:15410.11029/12010gl043622, 2010.
Liu, L. and Greaver, T. L.: A global perspective on belowground carbon dynamics under nitrogen enrichment, Ecol. Lett., 13, 819–828, 2010.
Lu, M., Yang, Y., Luo, Y., Fang, C., Zhou, X., Chen, J., Yang, X., and Li, B.: Responses of ecosystem nitrogen cycle to nitrogen addition: a meta-analysis, New Phytol., 189, 1040–1050, 2011a.
Lu, M., Zhou, X., Luo, Y., Yang, Y., Fang, C., Chen, J., and Li, B.: Minor stimulation of soil carbon storage by nitrogen addition: A meta-analysis, Agr., Ecosys. Environ., 140, 234–244, 2011b.
Luo, Y.: Terrestrial carbon-cycle feedback to climate warming, in: Annual Review of Ecology Evolution and Systematics, Annu. Rev. Ecol. Evol. S., 38, 683–712, 2007.
Luo, Y. and Weng, E.: Dynamic disequilibrium of the terrestrial carbon cycle under global change, Trends Ecol. Evol., 26, 96–104, 2011.

BGD
9, 1899–1944, 2012

A framework of benchmarking land models

Y. Q. Luo et al.

Luo, Y., Weng, E., Wu, X., Gao, C., Zhou, X., and Zhang, L.: Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models, Ecol. Appl., 19, 571–574, 2009.
Luo, Y., Ogle, K., Tucker, C., Fei, S., Gao, C., LaDeau, S., Clark, J. S., and Schimel, D. S.: Ecological forecasting and data assimilation in a data-rich era, Ecol. Appl., 21, 1429–1442, 2011.
Luo, Y. Q., White, L. W., Canadell, J. G., DeLucia, E. H., Ellsworth, D. S., Finzi, A. C., Lichter, J., and Schlesinger, W. H.: Sustainability of terrestrial carbon sequestration: A case study in Duke Forest with inversion approach, Global Biogeochem. Cy., 17, 1021, doi:1010.1029/2002gb001923, 2003.
Luo, Y. Q., Hui, D. F., and Zhang, D. Q.: Elevated CO_2 stimulates net accumulations of carbon and nitrogen in land ecosystems: A meta-analysis, Ecology, 87, 53–63, 2006.
Mahecha, M. D., Reichstein, M., Jung, M., Seneviratne, S. I., Zaehle, S., Beer, C., Braakhekke, M. C., Carvalhais, N., Lange, H., Le Maire, G., and Moors, E.: Comparing observations and process-based simulations of biosphere-atmosphere exchanges on multiple timescales, J. Geophys. Res., 115, G02003, doi:02010.01029/02009jg001016, 2010.
Maignan, F., Bréon, F.-M., Chevallier, F., Viovy, N., Ciais, P., Garrec, C., Trules, J., and Mancip, M.: Evaluation of a Global Vegetation Model using time series of satellite vegetation indices, Geosci. Model Dev., 4, 1103–1114, doi:10.5194/gmd-4-1103-2011, 2011.
Matthews, E.: Global litter production, pools, and turnover times: Estimates from measurement data and regression models, J. Geophys. Res.-Atmos., 102, 18771–18800, 1997.
McGuire, A. D., Sitch, S., Clein, J. S., Dargaville, R., Esser, G., Foley, J., Heimann, M., Joos, F., Kaplan, J., Kicklighter, D. W., Meier, R. A., Melillo, J. M., Moore, B., Prentice, I. C., Ramankutty, N., Reichenau, T., Schloss, A., Tian, H., Williams, L. J., and Wittenberg, U.: Carbon balance of the terrestrial biosphere in the twentieth century: Analyses of CO_2, climate and land use effects with four process-based ecosystem models, Global Biogeochem. Cy., 15, 183–206, 2001.
Mittelmann, H. D. and Preussner, A.: A server for automated performance analysis of benchmarking data, Optim. Method. Softw., 21, 105–120, 2006.
Moody, E. G., King, M. D., Platnick, S., Schaaf, C. B., and Gao, F.: Spatially complete global spectral surface albedos: Value-added datasets derived from terra MODIS land products, IEEE T. Geosci. Remote, 43, 144–158, 2005.
Moody, E. G., King, M. D., Schaaf, C. B., and Platnick, S.: MODIS-Derived Spatially Complete Surface Albedo Products: Spatial and Temporal Pixel Distribution and Zonal Averages, J.

Appl. Meteorol. Climatol., 47, 2879–2894, 2008.
Morgan, J. A., Pataki, D. E., Korner, C., Clark, H., Del Grosso, S. J., Grunzweig, J. M., Knapp, A. K., Mosier, A. R., Newton, P. C. D., Niklaus, P. A., Nippert, J. B., Nowak, R. S., Parton, W. J., Polley, H. W., and Shaw, M. R.: Water relations in grassland and desert ecosystems exposed to elevated atmospheric CO_2, Oecologia, 140, 11–25, 2004.
Mu, Q., Heinsch, F. A., Zhao, M., and Running, S. W.: Development of a global evapotranspiration algorithm based on MODIS and global meteorology data, Remote Sens. Environ., 111, 519–536, 2007.
Mu, Q., Zhao, M., and Running, S. W.: Improvements to a MODIS global terrestrial evapotranspiration algorithm, Remote Sens. Environ., 115, 1781–1800, 2011.
Norby, R. J. and Iversen, C. M.: Nitrogen uptake, distribution, turnover, and efficiency of use in a CO_2-enriched sweetgum forest, Ecology, 87, 5–14, 2006.
Norby, R. J. and Zak, D. R.: Ecological lessons from free-air CO_2 enrichment (FACE) experiments, Annu. Rev. Ecol. Evol. S., 42, 181–203, 2011.
Norby, R. J., DeLucia, E. H., Gielen, B., Calfapietra, C., Giardina, C. P., King, J. S., Ledford, J., McCarthy, H. R., Moore, D. J. P., Ceulemans, R., De Angelis, P., Finzi, A. C., Karnosky, D. F., Kubiske, M. E., Lukac, M., Pregitzer, K. S., Scarascia-Mugnozza, G. E., Schlesinger, W. H., and Oren, R.: Forest response to elevated CO_2 is conserved across a broad range of productivity, P. Natl. Acad. Sci., 102, 18052–18056, 2005.
Norby, R. J., Warren, J. M., Iversen, C. M., Medlyn, B. E., and McMurtrie, R. E.: CO_2 enhancement of forest productivity constrained by limited nitrogen availability, P. Natl. Acad. Sci., 107, 19368–19373, 2010.
Notaro, M., Liu, Z. Y., Gallimore, R., Vavrus, S. J., Kutzbach, J. E., Prentice, I. C., and Jacob, R. L.: Simulated and observed preindustrial to modern vegetation and climate changes, J.Climate, 18, 3650–3671, 2005.
Oleson, K. W.: Technical description of version 4.0 of the Community Land Model (CLM), NCAR Technical Note NCAR/TN-478+STR, National Center for Atmospheric Research, Boulder, CO, 2010.
Ollinger, S. V., Richardson, A. D., Martin, M. E., Hollinger, D. Y., Frolking, S. E., Reich, P. B., Plourde, L. C., Katul, G. G., Munger, J. W., Oren, R., Smith, M.-L., Paw U, K. T., Bolstad, P. V., Cook, B. D., Day, M. C., Martin, T. A., Monson, R. K., Schmid, H. P.: Canopy nitrogen, carbon assimilation, and albedo in temperate and boreal forests: Functional relations and potential climate feedbacks, P. Natl. Acad. Sci. USA, 105, 19336–19341, 2008.

Olson, J. S., Watts, J. A., and Allison, L. J.: Carbon in Live Vegetation of Major World Ecosystems, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 152, 1983.
Oreskes, N.: The role of quantitative models in science, in: Models in Ecosystem Science, edited by: Canham, C. D., Cole, J. J., and Lauenroth, W. K., Princeton University Press, Princeton, 13–31, 2003.
Owe, M., De Jeu, R. A. M., and Holmes, T. R. H.: Multi-Sensor Historical Climatology of Satellite-Derived Global Land Surface Moisture, J. Geophys. Res., 113, F01002, doi:1029/2007JF000769, 2008.
Peng, C., Guiot, J., Wu, H., Jiang, H., and Luo, Y.: Integrating models with data in ecology and palaeoecology: advances towards a model-data fusion approach, Ecol. Lett., 14, 522–536, 2011.
Piao, S., Wang, X., Ciais, P., Zhu, B., Wang, T., and Liu, J.: Changes in satellite-derived vegetation growth trend in temperate and boreal Eurasia from 1982 to 2006, Glob. Change Biol., 17, 3228–3239, 2011.
Pitman, A. J.: The evolution of, and revolution in, land surface schemes designed for climate models, Int. J. Climatol., 23, 479–510, 2003.
Post, W. M., Emanuel, W. R., Zinke, P. J., and Stangenberger, A. G.: Soil carbon pools and world life zones, Nature, 298, 156–159, 1982.
Prentice, I. C., Jolly, D., and BIOME 6000 participants: Mid-Holocene and glacial-maximum vegetation geography of the northern continents and Africa, J. Biogeogr., 27, 507–519, 2000.
Prentice, I. C., Kelley, D. I., Foster, P. N., Friedlingstein, P., Harrison, S. P., and Bartlein, P. J.: Modeling fire and the terrestrial carbon balance, Global Biogeochem. Cy., 25, GB3005, doi:10.1029/2010GB003906, 2011.
Prince, S. D. and Zheng, D.: ISLSCP II Global Primary Production Data Initiative Gridded NPP Datam, in: ISLSCP Initiative II Collection, edited by: Hall, F. G., Collatz, G., Meeson, B., Los, S., de Colstoun, E. B., and Landis, D., Data set, available at: http://daac.ornl.gov/, from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA, 2011.
Qu, X. and Hall, A.: What controls the strength of snow-albedo feedback?, J. Climate, 20, 3971–3981, 2007.
Randerson, J. T., Hoffman, F. M., Thornton, P. E., Mahowald, N. M., Lindsay, K., Lee, Y.-H., Nevison, C. D., Doney, S. C., Bonan, G., Stoeckli, R., Covey, C., Running, S. W., and Fung, I.

Y.: Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models, Glob. Change Biol., 15, 2462–2484, 2009.
Raupach, M. R., Rayner, P. J., Barrett, D. J., DeFries, R. S., Heimann, M., Ojima, D. S., Quegan, S., and Schmullius, C. C.: Model-data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications, Glob. Change Biol., 11, 378–397, 2005.
Reich, P. B., Wright, I. J., and Lusk, C. H.: Predicting leaf physiology from simple plant and climate attributes: A global GLOPNET analysis, Ecol. Appl., 17, 1982–1988, 2007.
Riley, W. J., Subin, Z. M., Lawrence, D. M., Swenson, S. C., Torn, M. S., Meng, L., Mahowald, N. M., and Hess, P.: Barriers to predicting changes in global terrestrial methane fluxes: analyses using CLM4Me, a methane biogeochemistry model integrated in CESM, Biogeosciences, 8, 1925–1953, doi:10.5194/bg-8-1925-2011, 2011.
Rodell, M., Chao, B. F., Au, A. Y., Kimball, J. S., and McDonald, K. C.: Global biomass variation and its geodynamic effects: 1982–1998, Earth Interact., 9, 1–19, 2005.
Rustad, L. E., Campbell, J. L., Marion, G. M., Norby, R. J., Mitchell, M. J., Hartley, A. E., Cornelissen, J. H. C., Gurevitch, J., and Gcte, N.: A Meta-Analysis of the Response of Soil Respiration, Net Nitrogen Mineralization, and Aboveground Plant Growth to Experimental Ecosystem Warming, Oecologia, 126, 543–562, 2001.
Rykiel, E. J.: Testing ecological models: The meaning of validation, Ecol. Model., 90, 229–244, 1996.
Saatchi, S. S., Houghton, R. A., Alvala, R. C. D. S., Soares, J. V., and Yu, Y.: Distribution of aboveground live biomass in the Amazon basin, Glob. Change Biol., 13, 816–837, 2007.
Schwalm, C. R., Williams, C. A., Schaefer, K., Anderson, R., Arain, M. A., Baker, I., Barr, A., Black, T. A., Chen, G., Chen, J. M., Ciais, P., Davis, K. J., Desai, A., Dietze, M., Dragoni, D., Fischer, M. L., Flanagan, L. B., Grant, R., Gu, L., Hollinger, D., Izaurralde, R. C., Kucharik, C., Lafleur, P., Law, B. E., Li, L., Li, Z., Liu, S., Lokupitiya, E., Luo, Y., Ma, S., Margolis, H., Matamala, R., McCaughey, H., Monson, R. K., Oechel, W. C., Peng, C., Poulter, B., Price, D. T., Riciutto, D. M., Riley, W., Sahoo, A. K., Sprintsin, M., Sun, J., Tian, H., Tonitto, C., Verbeeck, H., Verma, S.: A model data intercomparison of CO_2 exchange across North America: Results from the North American Carbon Program site synthesis, J. Geophysi. Res., 115, G00H05, doi:10.1029/2009JG001229, 2010.
Sherry, R. A., Zhou, X., Gu, S., Arnone III, J. A., Schimel, D. S., Verburg, P. S., Wallace, L. L., and Luo, Y.: Divergence of reproductive phenology under climate warming, P. Natl. Acad.

BGD
9, 1899–1944, 2012
A framework of benchmarking land models
Y. Q. Luo et al.

Sci. USA, 104, 198–202, 2007.
Simard, M., Pinto, N., Fisher, J. B., and Baccini, A.: Mapping forest canopy height globally with spaceborne LiDAR, J. Geophys. Res.-Atmos. Biogeo., 116, G04021, doi:10.1029/2011JG001708, 2011.
Simon, T. A. and McGalliard, J.: Observation and analysis of the multicore performance impact on scientific applications, Concurr. Comp. Prac. E., 21, 2213–2231, 2009.
Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J. O., Levis, S., Lucht, W., Sykes, M. T., Thonicke, K., and Venevsky, S.: Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model, Glob. Change Biol., 9, 161–185, 2003.
Sitch, S., Huntingford, C., Gedney, N., Levy, P. E., Lomas, M., Piao, S. L., Betts, R., Ciais, P., Cox, P., Friedlingstein, P., Jones, C. D., Prentice, I. C., and Woodward, F. I.: Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five Dynamic Global Vegetation Models (DGVMs), Glob. Change Biol., 14, 2015–2039, 2008.
Smith, E. P. and Rose, K. A.: Model goodness-of-fit analysis using regression and related techniques, Ecol. Model., 77, 49–64, 1995.
Tang, J. and Zhuang, Q.: Equifinality in parameterization of process-based biogeochemistry models: A significant uncertainty source to the estimation of regional carbon dynamics, J. Geophys. Res.-Biogeo., 113, G04010, doi:04010.01029/02008jg000757, 2008.
Tarnocai, C., Canadell, J. G., Schuur, E. A. G., Kuhry, P., Mazhitova, G., and Zimov, S.: Soil organic carbon pools in the northern circumpolar permafrost region, Global Biogeochem. Cy., 23, Gb2023, doi:2010.1029/2008gb003327, 2009.
Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res.-Atmos., 106, 7183–7192, 2001.
Thomas, R. Q., Canham, C. D., Weathers, K. C., and Goodale, C. L.: Increased tree carbon storage in response to nitrogen deposition in the US, Nat. Geosci., 3, 13–17, 2010.
Thonicke, K., Venevsky, S., Sitch, S., and Cramer, W.: The role of fire disturbance for global vegetation dynamics: coupling fire into a Dynamic Global Vegetation Model, Global Ecol. Biogeogr., 10, 661–677, 2001.
Thonicke, K., Spessa, A., Prentice, I. C., Harrison, S. P., Dong, L., and Carmona-Moreno, C.: The influence of vegetation, fire spread and fire behaviour on biomass burning and trace gas emissions: results from a process-based model, Biogeosciences, 7, 1991–2011, doi:10.5194/bg-7-1991-2010, 2010.

Thornton, P. E., Lamarque, J.-F., Rosenbloom, N. A., and Mahowald, N. M.: Influence of carbon-nitrogen cycle coupling on land model response to CO_2 fertilization and climate variability, Global Biogeochem. Cy., 21, GB4018, doi:4010.1029/2006gb002868, 2007.

Trudinger, C. M., Raupach, M. R., Rayner, P. J., Kattge, J., Liu, Q., Pak, B., Reichstein, M., Renzullo, L., Richardson, A. D., Roxburgh, S. H., Styles, J., Wang, Y. P., Briggs, P., Barrett, D., and Nikolova, S.: OptIC project: An intercomparison of optimization techniques for parameter estimation in terrestrial biogeochemical models, J. Geophys. Res.-Biogeo., 112, G02027, doi:02010.01029/02006jg000367, 2007.

van der Werf, G. R., Randerson, J. T., Collatz, G. J., Giglio, L., Kasibhatla, P. S., Arellano, A. F., Olsen, S. C., and Kasischke, E. S.: Continental-scale partitioning of fire emissions during the 1997 to 2001 El Niño/La Niña period, Science, 303, 73–76, 2004.

van der Werf, G. R., Randerson, J. T., Giglio, L., Collatz, G. J., Kasibhatla, P. S., and Arellano Jr., A. F.: Interannual variability in global biomass burning emissions from 1997 to 2004, Atmos. Chem. Phys., 6, 3423–3441, doi:10.5194/acp-6-3423-2006, 2006.

Vinukollu, R. K., Wood, E. F., Ferguson, C. R., and Fisher, J. B.: Global estimates of evapotranspiration for climate studies using multi-sensor remote sensing data: Evaluation of three process-based approaches, Remote Sens. Environ., 115, 801–823, 2011.

Wan, S. Q., Hui, D. F., and Luo, Y. Q.: Fire effects on nitrogen pools and dynamics in terrestrial ecosystems: A meta-analysis, Ecol. Appl., 11, 1349–1365, 2001.

Wang, A., Price, D. T., and Arora, V.: Estimating changes in global vegetation cover (1850–2100) for use in climate models, Global Biogeochem. Cy., 20, GB3028, doi:10.1029/2005GB002514, 2006.

Wang, Y. P. and Barrett, D. J.: Estimating regional terrestrial carbon fluxes for the Australian continent using a multiple-constraint approach – Part 1: Using remotely sensed data and ecological observations of net primary production, Tellus B, 55, 270–289, 2003.

Wang, Y.-P., Leuning, R., Cleugh, H. A., and Coppin, P. A.: Parameter estimation in surface exchange models using nonlinear inversion: how many parameters can we estimate and which measurements are most useful?, Glob. Change Biol., 7, 495–510, 2001.

Wang, Y.-P., Trudinger, C. M., and Enting, I. G.: A review of applications of model-data fusion to studies of terrestrial carbon fluxes at different scales, Agr. Forest Meteorol., 149, 1829–1842, 2009.

Wang, Y. P., Law, R. M., and Pak, B.: A global model of carbon, nitrogen and phosphorus cycles for the terrestrial biosphere, Biogeosciences, 7, 2261–2282, doi:10.5194/bg-7-2261-

2010, 2010.
Wang, Y. P., Kowalczyk, E., Leuning, R., Abramowitz, G., Raupach, M. R., Pak, B., van Gorsel, E., and Luhar, A.: Diagnosing errors in a land surface model (CABLE) in the time and frequency domains, J. Geophys. Res.-Biogeo., 116, G01034, doi:01010.01029/02010jg001385, 2011.
Weng, E. and Luo, Y.: Relative information contributions of model vs. data to short- and long-term forecasts of forest carbon dynamics, Ecol. Appl., 21, 1490–1505, 2011.
Woodhouse, I. H.: Predicting backscatter-biomass and height-biomass trends using a macroecology model, IEEE T. Geosci. Remote, 44, 871–877, 2006.
Wright, I. J., Reich, P. B., Cornelissen, J. H. C., Falster, D. S., Groom, P. K., Hikosaka, K., Lee, W., Lusk, C. H., Niinemets, U., Oleksyn, J., Osada, N., Poorter, H., Warton, D. I., and Westoby, M.: Modulation of leaf economic traits and trait relationships by climate, Global Ecol. Biogeogr., 14, 411–421, 2005.
Wu, Z., Dijkstra, P., Koch, G. W., Penuelas, J., and Hungate, B. A.: Responses of terrestrial ecosystems to temperature and precipitation change: a meta-analysis of experimental manipulation, Glob. Change Biol., 17, 927–942, 2011.
Xu, T., White, L., Hui, D. F., and Luo, Y. Q.: Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction, Global Biogeochem. Cy., 20, Gb2007, doi:2010.1029/2005gb002468, 2006.
Yang, Y., Luo, Y., and Finzi, A. C.: Carbon and nitrogen dynamics during forest stand development: a global synthesis, New Phytol., 190, 977–989, 2011.
Yuan, H., Dai, Y., Xiao, Z., Ji, D., and Shangguan, W.: Reprocessing the MODIS Leaf Area Index Products for Land Surface and Climate Modelling, Remote Sens. Environ., 115, 1171–1187, 2011.
Zaehle, S. and Friend, A. D.: Carbon and nitrogen cycle dynamics in the O-CN land surface model – Part 1: Model description, site-scale evaluation, and sensitivity to parameter estimates, Global Biogeochem. Cy., 24, Gb1005, doi:1010.1029/2009gb003521, 2010.
Zaehle, S., Friedlingstein, P., and Friend, A. D.: Terrestrial nitrogen feedbacks may accelerate future climate change, Geophys. Res. Lett., 37, L01401, doi:01410.01029/02009gl041345, 2010.
Zhou, T. and Luo, Y.: Spatial patterns of ecosystem carbon residence time and NPP-driven carbon uptake in the conterminous United States, Global Biogeochem. Cy., 22, GB3032, doi:3010.1029/2007gb002939, 2008.

Zinke, P. J., Stangenberger, A. G., Post, W. M., Emanuel, W. R., and Olson, J. S.: Worldwide Organic Soil Carbon and Nitrogen Data, NDP-018, Oak Ridge National Laboratory, Oak Ridge, Tennessee USA, 146, 1986.

BGD
9, 1899–1944, 2012

A framework of benchmarking land models

Y. Q. Luo et al.

Table 1. Types of benchmarks to be used for evaluating model performance.

Type	Description	Example	Pros	Cons
Direct observations	Data from instrument readings with some processing	Temperature, soil respiration	Records of systems states	Limited spatial and temporal coverage
Experimental results	Data at two or more levels of treatments	Biomass, soil moisture	Effects of climate changes	Step changes in treatments, site-idiosyncrasy
Data-model products	Interpolation and extrapolation of data according to some functions	Global distribution of GPP calculated from satellite or flux data	Extended spatial and temporal coverage with estimated errors	Artifacts induced by the functions, especially outside the observation ranges
Functional relationships or patterns	Derived or emerged from data	NPP vs. precipitation, Soil respiration vs. temperature	Evaluation of environmental scalars and response functions	Not absolute values of the variables

BGD
9, 1899–1944, 2012

Table 2. Sample benchmarks to be used to evaluate biophysical processes.

Variable/factor	Benchmark				Evaluation
	Data set	Temporal frequency	Spatial coverage	Reference	
Baseline states and fluxes					
Latent heat flux (ET)	Gridded map	8-day to yearly	Global	Fisher et al. (2008) Jung et al. (2010) Mu et al. (2011)	Heat flux and ET
Surface albedo	Gridded map	16-days to yearly	Global	Moody et al. (2005, 2008)	Energy-water partitioning
Runoff	Gridded map	Monthly to yearly	Global	Dai et al. (2009)	Water cycle
Surface and soil temperature	Gridded map	Monthly to yearly	Global	FLUXNET, CRU, GISS, and NCDC	Energy balance
Soil moisture	Gridded map	Monthly to yearly	Global	Owe et al. (2008); Dorigo et al. (2011)	Water cycle
Snow cover	Gridded map	Monthly to yearly	Global	AVHRR, MODIS, GlobSow	Energy partitioning
Snow depth/SWE	Gridded map	Monthly to yearly	Regional -NA	CMC	Water cycle
Responses of state and rate variables to disturbances and global change					
Elevated CO_2	Response ratio	Weekly-yearly	Site	Morgan et al. (2004)	Water cycle
Warming	Response ratio	Weekly-yearly	Site	Bell et al. (2010)	Soil water dynamics

Table 3. Sample benchmarks to be used to evaluate biogeochemical cycles.

Variable/factor	Benchmark				Evaluation
	Data set	Temporal frequency	Spatial coverage	Reference	
Baseline states and fluxes					
GPP	Gridded map	Monthly to yearly	Global	Jung et al. (2011) Frankenberg et al. (2011)	Carbon influx
NPP	Gridded map	Yearly	Global	Prince et al. (2011)	Carbon influx
Soil respiration	Gridded map	Yearly	Global	Bond-Lamberty and Thomson (2010)	Carbon efflux
Ecosystem respiration	Gridded map	Yearly	Global	Jung et al. (2011)	Carbon efflux
Plant biomass	Gridded map		Global	Olson et al. (1983); Rodell et al. (2005); Saatchi et al. (2007); Woodhouse (2006)	Carbon pool
Litter pool	Gridded map		Global	Matthews (1997)	Carbon pool
Litter decay rate			Various sites	Boyero et al. (2011)	Rate process
Soil carbon	Gridded map		Global	Batjes (2002); Post et al. (1982); Zinke et al. (1986); FAO (2009)	Carbon pool
FAPAR	Gridded map	Monthly to yearly	Regional to Global	Gobron et al. (2004); Yuan et al. (2011)	Carbon influx
Responses of state and rate variables to disturbances and global change					
Elevated CO_2	Response ratio		Various regions	Luo et al. (2006); Norby and Iversen (2006)	Responses of carbon and nitrogen processes
Warming	Response ratio		Various regions	Rustad et al. (2001); Wu et al. (2011)	Responses of carbon processes
N deposition	Response ratio		Various regions	Janssens et al. (2010); Liu and Greaver (2010); Lu et al. (2011a); Thomas et al. (2010); Lu et al. (2011b)	Carbon and nitrogen cycles
Fire		Monthly to yearly		Wan et al. (2001); van der Werf et al. (2004, 2006)	Carbon cycle Nitrogen cycle
Insect outbreak		Yearly		Kurz et al. (2008a, b)	Carbon cycle

BGD
9, 1899–1944, 2012

Table 4. Sample benchmarks to be used to evaluate biogeographical processes (vegetation dynamics).

Variable/factor	Benchmark				Evaluation
	Data set	Temporal frequency	Spatial coverage	Reference	
Baseline states and fluxes					
Pre-industrial vegetation types	Vegetation map	Once	Global	Notaro et al. (2005)	Initial values of vegetation
Canopy height	Gridded map	Once	Global	Lefsky (2010); Simard et al. (2011)	Vegetation dynamics
Responses of state and rate variables to disturbances and global change					
Warming	Response ratio	Yearly	Site	Sherry et al. (2007)	Phenology
N deposition	Response ratio	Yearly	Various regions	Thomas et al. (2010)	
Fire	Burned area, vegetation change	Seasonal and Yearly	Global	Thonicke et al. (2001), GFED3	Vegetation
Land use and change	Changes in global vegetation cover	Yearly	Global	Wang et al. (2006) MODIS PFT fraction	Plant functional type
Wood harvest	Biomass removal	Annual mean	Global	Hurtt et al. (2006)	Land use change

BGD
9, 1899–1944, 2012
A framework of benchmarking land models
Y. Q. Luo et al.

Table 5. Comparison of evaluation procedures between benchmark analysis and data assimilation.

Procedure	Benchmark analysis	Data assimilation
Targets	Model aspects to be evaluated	Parameters to be estimated or model structures to be chosen
References	Benchmarks	Multiple data sets
Criteria	Scoring systems	Cost functions
Outcomes	Suggesting model improvement	Estimates of parameter and/or selections of model structures

BGD
9, 1899–1944, 2012

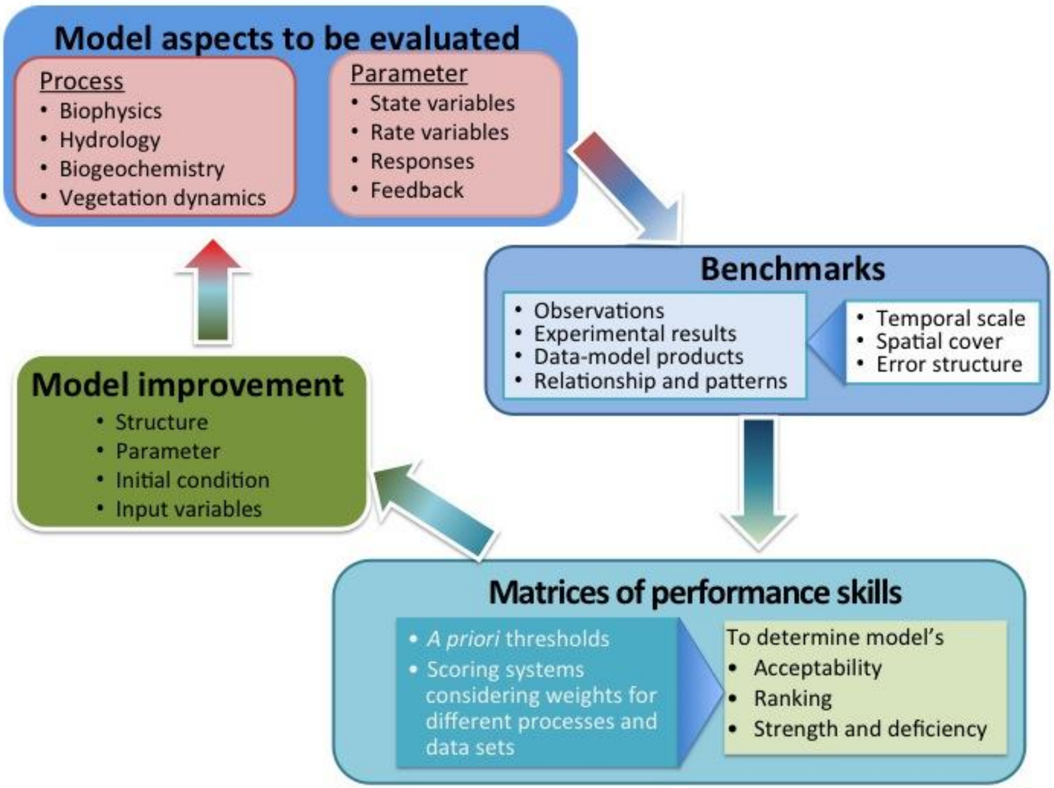

Fig. 1. Schematic diagram of the benchmarking framework for evaluating land models. The framework include four major components: (1) defining model aspects to be evaluated, (2) selecting benchmarks as standardized references to test models, (3) developing a scoring system to measure model performance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

BGD
9, 1899–1944, 2012

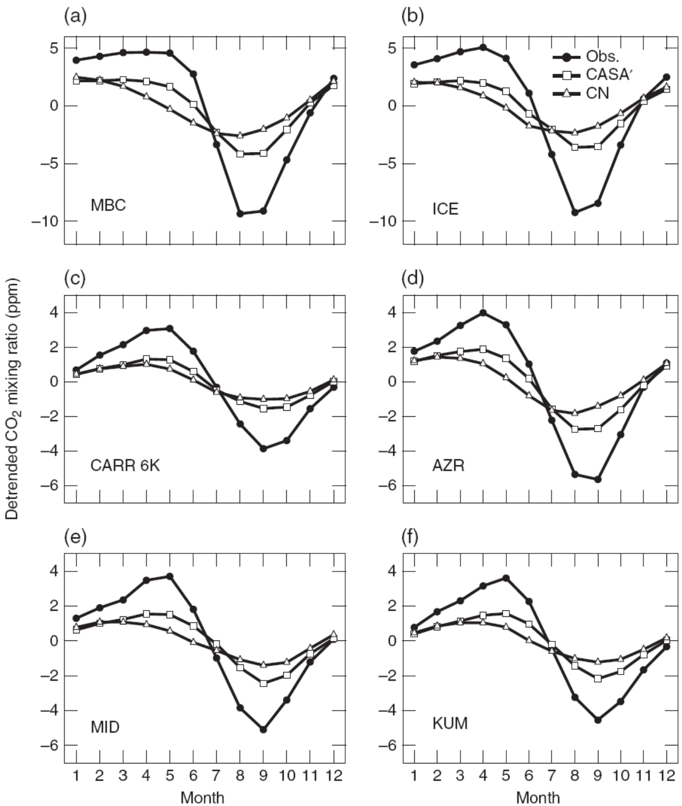

Fig. 2. Benchmark analysis of Community Land Model (CLM) CASA' and CN versions against the seasonal cycle observations from NOAA. The annual cycle of CO_2 is regulated by plant phenology, photosynthesis, allocation, and decomposition processes. A well functioning model has to match the observations, but it is possible to get the right answer for the wrong reasons. Thus, multiple orthogonal constraints and parallel use of functional relationships are needed for benchmark analysis (adopted from Randerson et al., 2009).

BGD
9, 1899–1944, 2012

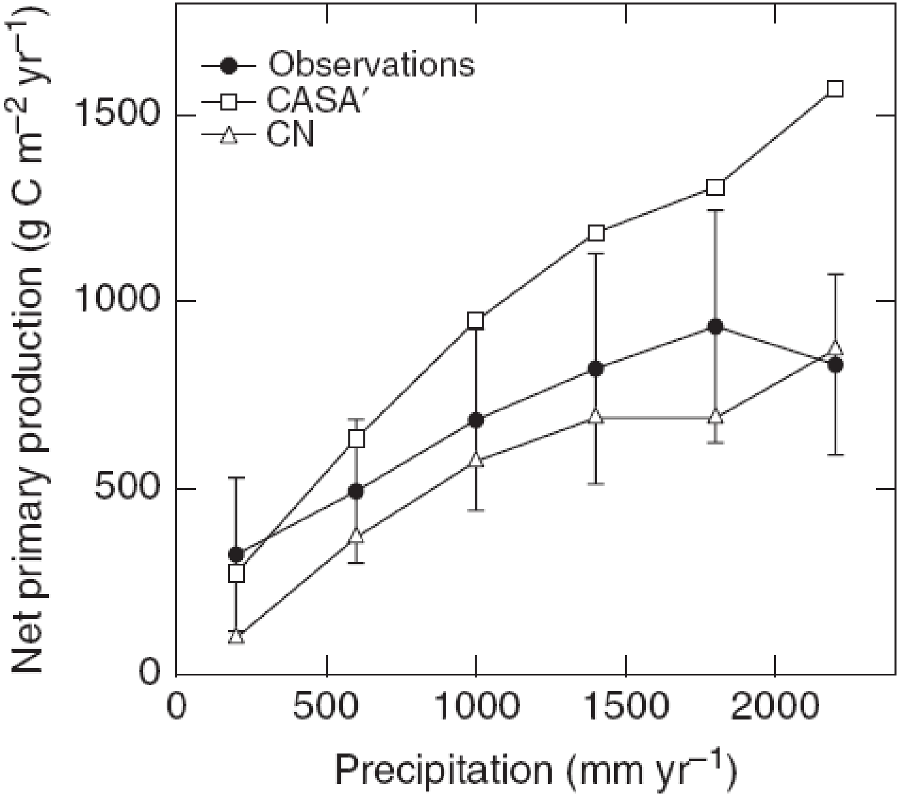

Fig. 3. Functional relationship between net primary production with precipitation used in a benchmark analysis for coupled models that account for possible biases in model climate (adopted from Randerson et al., 2009).

BGD
9, 1899–1944, 2012

A framework of benchmarking land models

Y. Q. Luo et al.

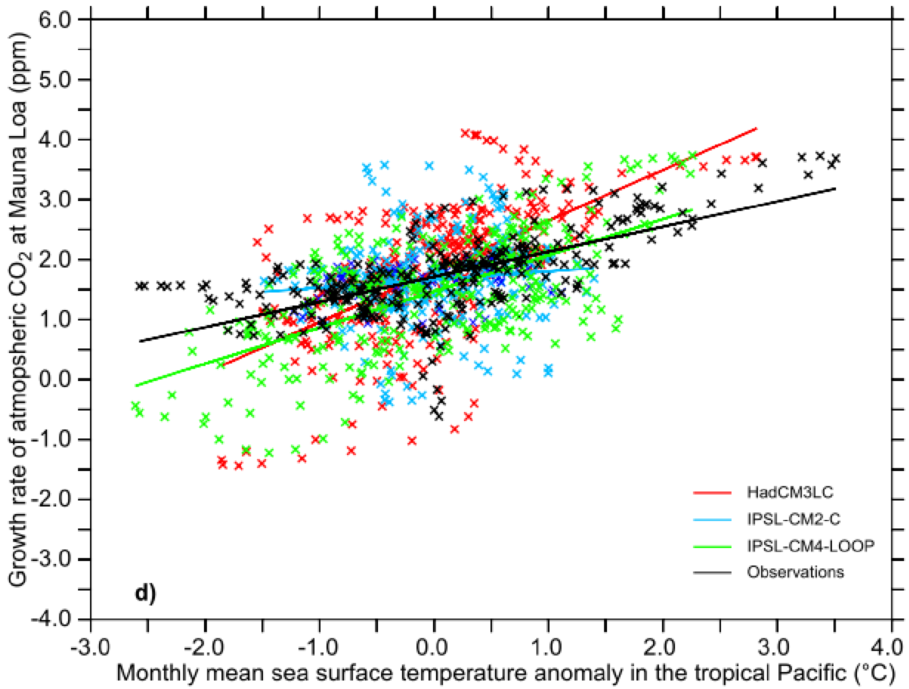

Fig. 4. CO_2-temperature relationships used in a benchmark analysis to show the positive and negative anomalies of atmospheric CO_2 growth rate as a function of anomalies of Eastern Tropical Pacific SST (adopted from Cadule et al., 2010).

BGD
9, 1899–1944, 2012

A framework of benchmarking land models

Y. Q. Luo et al.

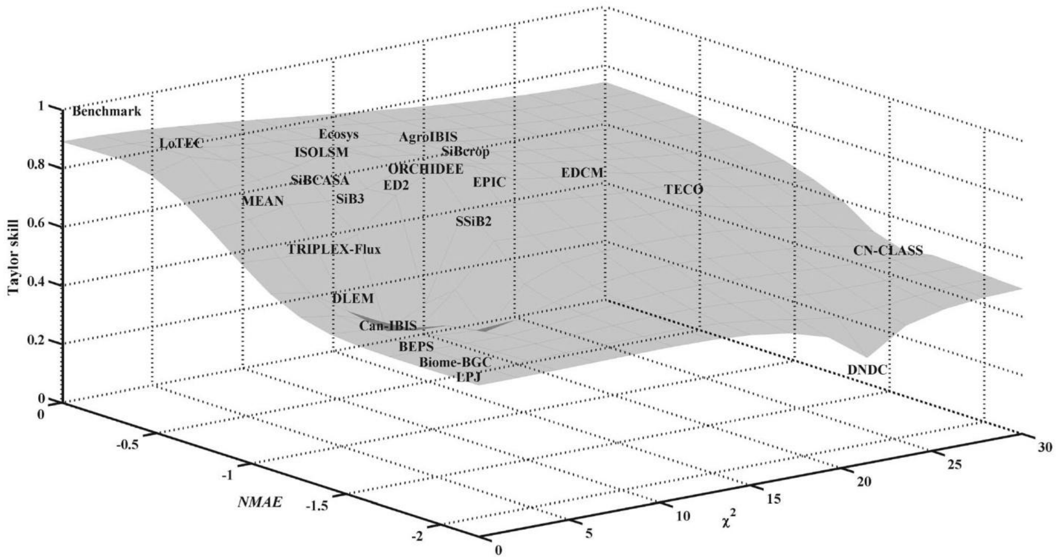

Fig. 5. Model skill metrics for 22 terrestrial ecosystem models. Skill metrics are Taylor skill (S), normalized mean absolute error (NMAE), and reduced chi-squared statistic (χ^2). χ^2 is the distance between simulated and observed values denominated in multiples of observational uncertainty. Better model-data agreement corresponds to the upper left corner. Benchmark represents perfect model-data agreement: $S = 1$, NMAE $= 0$, and $\chi^2 = 1$. Gray interpolated surface added and model names jittered to improve readability. Model names are described in Schwalm et al. (2010).

BGD
9, 1899–1944, 2012

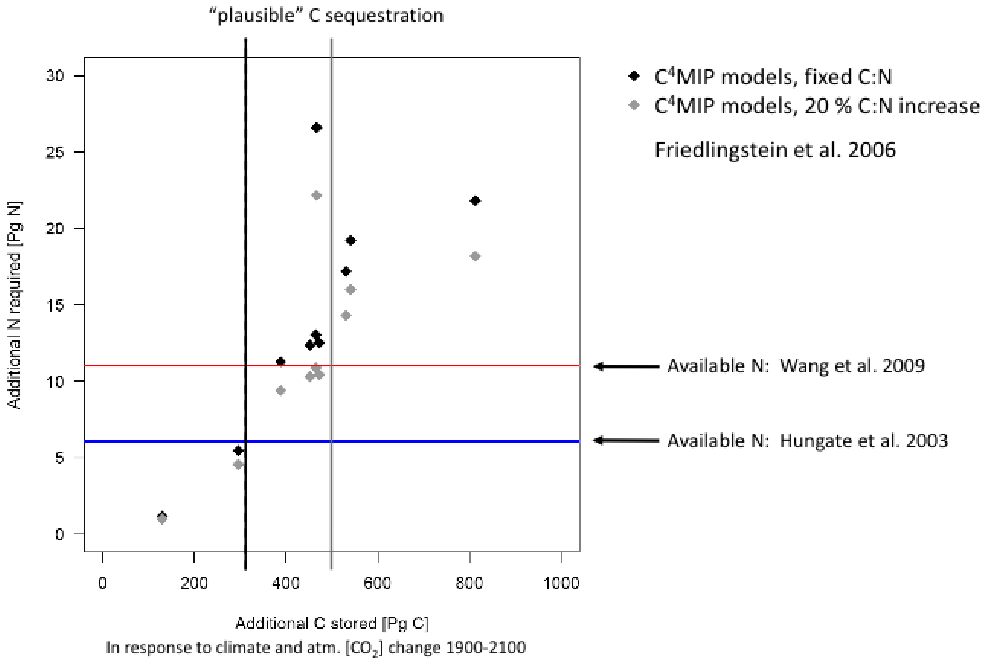

Fig. 6. Nitrogen constraints of carbon sequestration. The original analysis by Hungate et al. (2003) was based on some biogeochemical principles to reveal major deficiencies in global biogeochemical models. The analysis may not be considered as a typical benchmark analysis but played a role in stimulating global modeling groups to incorporate nitrogen processes into their models. However, relative performance skills of land models as measured by the benchmark analysis vary with additional considerations of data sets as illustrated in analysis on flexibility of C:N ratio by Wang et al. (2009). Moreover, nitrogen capital in terrestrial ecosystem is considerably dynamic in response to rising atmospheric CO_2 concentration (Luo et al., 2006), rendering less limitation of ecosystem carbon sequestration.

BGD
9, 1899–1944, 2012

A framework of benchmarking land models

Y. Q. Luo et al.