Biogeosciences, 10, 3313–3340, 2013
www.biogeosciences.net/10/3313/2013/
doi:10.5194/bg-10-3313-2013

Biogeosciences Open Access

A comprehensive benchmarking system for evaluating global vegetation models

D. I. Kelley[1], **I. C. Prentice**[1,2], **S. P. Harrison**[1,3], **H. Wang**[1,4], **M. Simard**[5], **J. B. Fisher**[5], **and K. O. Willis**[1]

[1]Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109, Australia
[2]Grantham Institute for Climate Change, and Department of Life Sciences, Imperial College, Silwood Park Campus, Ascot SL5 7PY, UK
[3]Geography & Environmental Sciences, School of Human and Environmental Sciences, Reading University, Whiteknights, Reading, RG6 6AB, UK
[4]State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Science, Xiangshan Nanxincun 20, 100093 Beijing, China
[5]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA

Correspondence to: D. I. Kelley (douglas.kelley@students.mq.edu.au)

Received: 20 July 2012 – Published in Biogeosciences Discuss.: 9 November 2012
Revised: 5 March 2013 – Accepted: 7 March 2013 – Published: 17 May 2013

Abstract. We present a benchmark system for global vegetation models. This system provides a quantitative evaluation of multiple simulated vegetation properties, including primary production; seasonal net ecosystem production; vegetation cover; composition and height; fire regime; and runoff. The benchmarks are derived from remotely sensed gridded datasets and site-based observations. The datasets allow comparisons of annual average conditions and seasonal and inter-annual variability, and they allow the impact of spatial and temporal biases in means and variability to be assessed separately. Specifically designed metrics quantify model performance for each process, and are compared to scores based on the temporal or spatial mean value of the observations and a "random" model produced by bootstrap resampling of the observations. The benchmark system is applied to three models: a simple light-use efficiency and water-balance model (the Simple Diagnostic Biosphere Model: SDBM), the Lund-Potsdam-Jena (LPJ) and Land Processes and eXchanges (LPX) dynamic global vegetation models (DGVMs). In general, the SDBM performs better than either of the DGVMs. It reproduces independent measurements of net primary production (NPP) but underestimates the amplitude of the observed CO_2 seasonal cycle. The two DGVMs show little difference for most benchmarks (including the inter-annual variability in the growth rate and seasonal cycle of atmospheric CO_2), but LPX represents burnt fraction demonstrably more accurately. Benchmarking also identified several weaknesses common to both DGVMs. The benchmarking system provides a quantitative approach for evaluating how adequately processes are represented in a model, identifying errors and biases, tracking improvements in performance through model development, and discriminating among models. Adoption of such a system would do much to improve confidence in terrestrial model predictions of climate change impacts and feedbacks.

1 Introduction

Dynamic global vegetation models (DGVMs) are widely used in the assessment of climate change impacts on ecosystems, and feedbacks through ecosystem processes (Cramer et al., 1999; Scholze et al., 2006; Sitch et al., 2008; Scheiter and Higgins, 2009). However, there are large differences in model projections of the vegetation response to scenarios of future changes in atmospheric CO_2 concentration and climate (Friedlingstein et al., 2006; Denman et al., 2007; Sitch et al., 2008). Assessing the uncertainty around vegetation-model simulations would provide an indicator of confidence in model predictions under different climates. Such a system would serve several functions, including the following: comparing the performance of different models; identifying processes in a particular model that need improvement; and checking that improvements in one part of a model do not compromise performance in another.

Benchmarking is a routine component in the assessment of climate-model performance, including investigation of parameter uncertainties (e.g. Murphy et al., 2004; Piani et al., 2005) and multi-model comparison (Randall et al., 2007; Reichler and Kim, 2008), and is used both to inform model development (e.g. Jackson et al., 2008) and to interpret the reliability of projections of future climate (e.g. Shukla et al., 2006: Hall and Qu, 2006). In recent years, there has been considerable effort spent on the development of standard metrics for climate-model evaluation (Taylor, 2001; Gleckler et al., 2008: Lenderink, 2010; Moise and Delage, 2011; Yokoi et al., 2011). In comparison, there has been little quantitative assessment of DGVM performance under recent conditions. Although most studies describing vegetation-model development provide some assessment of the model's predictive ability by comparison with observational datasets (e.g. Sitch et al., 2003; Woodward and Lomas, 2004; Prentice et al., 2007), such comparisons often focus just on one aspect of the model where recent development has taken place (e.g. Gerten et al., 2004; Arora and Boer, 2005; Zeng et al., 2008; Thonicke et al., 2010; Prentice et al., 2011). It has not been standard practice to track improvements in (or degradation of) general model performance caused by new developments.

A benchmarking system should facilitate more comprehensive model evaluation, and help to make such tracking routine. The land modelling community has recently recognized the need for such a system (e.g. the International Land Model Benchmarking Project, ILAMB: http://www.ilamb.org/), and some recent studies have designed and applied benchmarking systems. Blyth et al. (2009, 2011) compared results of the JULES land-surface model with site-based water and CO_2 flux measurements and satellite vegetation indices, quantifying the difference between model output and observations using root mean squared error (RMSE) as a metric. Beer et al. (2010) used a gridded dataset of gross primary productivity (GPP), derived from up-scaling GPP from the FLUXNET network of eddy covariance towers (Jung et al., 2009, 2010) to assess and compare the Lund-Potsdam-Jena (LPJ), LPJmL, ORCHIDEE, CLM-CN and SDGVM models. Bonan et al. (2011) evaluated latent heat fluxes with the tower-derived gridded GPP dataset (Beer et al., 2010) to evaluate the calibration of the CLM4 model. Cadule et al. (2010) used the model-to-data deviation, normalised standard deviation and Pearson's correlation to quantify the "distance" between simulated and observed CO_2 concentration and applied these to compare three coupled climate–vegetation models that incorporate two DGVMs: TRIFFID and ORCHIDEE. All of these studies focus on a very limited number of simulated processes, and use metrics that are difficult to interpret across processes and models. Randerson et al. (2009) introduced a more systematic framework to assess and compare the performance of two biogeochemical models (CLM-CN and CASA') against net primary production (NPP) and CO_2 concentration data, including the definition of comparison metrics tailored to the benchmark observations and a composite skill score that combined metric scores for each observation into an overall measure of model performance. The Randerson et al. (2009) composite score was a weighted combination of scores across different metrics, where the weights were based on a qualitative and necessarily somewhat subjective assessment of the "importance" and uncertainty of each process (Randerson et al., 2009). Luo et al. (2012) recommended the development of a working benchmarking system for vegetation models that incorporates some of the approaches used in these various studies including a set of standard target datasets for benchmarks, a scoring system; and a way of comparing across model processes in order to evaluate model strengths and weaknesses to guide model development. Luo et al. (2012) reject the idea of a single composite metric because of the subjectivity involved in choices of relative weightings.

Our purpose here is to demonstrate a benchmarking system including multiple observational datasets and transparent metrics of model performance with respect to individual processes. We have tested the system on three vegetation models to demonstrate the system's capabilities in comparing model performance, assigning a level of confidence to the models' predictions of key ecosystem properties, assessing the representation of different model processes and identifying deficiencies in each model.

2 Materials and methods

2.1 Principles

The benchmarking system consists of a collection of datasets, selected to fulfil certain criteria and to allow systematic evaluation of a range of model processes, and metrics, designed with the characteristics of each benchmark dataset in mind. We selected site-based and remotely sensed observational datasets that, as far as possible, fulfil the following requirements:

- They should be global in coverage or, for site-based data, they should sample reasonably well the different biomes on each continent. This criterion excludes "campaign mode" measurements, and datasets assembled only for one continent or region.

- They should be independent of any modelling approach that involves calculation of vegetation properties from the same driving variables as the vegetation models being tested. This criterion allows remotely sensed fraction of absorbed photosynthetically active radiation (fAPAR) products but excludes the MODIS NPP product used by Randerson et al. (2009), or remotely sensed evapotranspiration (e.g. Fisher et al., 2008, 2011; Mu et al., 2011). It allows use of flux measurements and CO_2 inversion products, but excludes, for example, the up-scaled GPP used by Beer et al. (2010).

- They should be available for multiple years and seasonal cycles to allow assessment of modelled seasonal and inter-annual variation, for variables that change on these time scales.
- Datasets should be freely available, so that different modelling groups can evaluate their models against the same benchmarks.

The selected datasets (Table 1) provide information for the following: fAPAR, the fractional coverage of different plant life and leaf forms, GPP and NPP, height of the canopy, fire, as burnt fraction; runoff, as river discharge, and seasonal and inter-annual variation in atmospheric CO_2 concentration (Fig. 1):

- fAPAR is the fundamental link between primary production and available energy (Monteith, 1972). It measures the seasonal cycle, inter-annual variability and trends of vegetation cover. Of all ecosystem properties derived from spectral reflectance measurements, fAPAR is closest to the actual measurements.
- Fractional cover of different life forms and leaf forms provides basic information about vegetation structure and phenology.
- GPP and NPP are the two fundamental measures of primary production.
- Vegetation height is a key variable for characterising vegetation structure, function and biomass.
- Remotely sensed data on fire (as fractional burnt area) have been available for a few years (e.g. Carmona-Moreno et al., 2005; Giglio et al., 2006). The latest dataset (Giglio et al., 2010; van der Werf et al., 2010) is derived from active fire counts and involves empirical (biome-dependent) modelling to translate between active fire counts and burned area. Our criteria exclude the use of the accompanying fire CO_2 emissions product (van der Werf et al., 2010), however, as this depends strongly on the use of a particular biogeochemical model.
- Annual runoff is an indicator of ecosystem function, as it represents the spatial integration of the difference between precipitation and evapotranspiration – the latter primarily representing water use by vegetation. It is a sensitive indicator, because a small proportional error in modelled evapotranspiration translates into a larger proportional error in runoff (Raupach et al., 2009). Runoff is measured independently of meteorological data by gauges in rivers.
- Atmospheric CO_2 concentration is measured at high precision at a globally distributed set of stations in remote locations (distant from urban and transport centres of CO_2 emission). The pattern of the seasonal cycle of atmospheric CO_2 concentration at different locations provides information about the sources and sinks of CO_2 in the land biosphere (Heimann et al., 1998), while the inter-annual variability of the increase in CO_2 provides information about CO_2 uptake at the global scale. Ocean impacts on the seasonal cycle are small (Nevison et al., 2008). For inter-annual variability we use inversion products which selectively remove the ocean contribution (about 20 % of the signal: Le Quéré et al., 2003).

All remotely sensed data were re-gridded to a 0.5° resolution grid and masked to a land mask common to all three models.

Data–model comparison metrics were designed to be easy to implement, intuitive to understand, and comparable across multiple benchmarked processes. Metric scores for comparison of models with these datasets were compared against scores from two null models: one corresponding to the observational mean and the other obtained by randomly resampling the observations.

To demonstrate whether the benchmark system fulfilled the functions of evaluating specific modelled processes and discriminating between models, we applied it to three global models: a simple light-use efficiency and water-balance model introduced by Knorr and Heimann (1995), known as the Simple Diagnostic Biosphere Model (SDBM: Heimann et al., 1998) and two DGVMs. The SDBM is driven by observed precipitation, temperature and remotely sensed observations of fAPAR. The model has two tunable global parameters representing light-use efficiency under well-watered conditions, and the shape of the exponential temperature dependence of heterotrophic respiration. The DGVMs are the Lund-Potsdam-Jena (LPJ) model (version 2.1: Sitch et al., 2003, as modified by Gerten et al., 2004) and the Land surface Processes and eXchanges (LPX) model (Prentice et al., 2011). LPX was developed from LPJ-SPITFIRE (Thonicke et al., 2010), and represents a further refinement of the fire module in LPJ-SPITFIRE.

2.2 Benchmark datasets

2.2.1 fAPAR

fAPAR data (http://oceancolor.gsfc.nasa.gov/SeaWiFS/; Table 1) were derived from the SeaWiFS remotely sensed fAPAR product (Gobron et al., 2006), providing monthly data for 1998–2005. fAPAR varies between 0 and 1, and the average uncertainty for any cell/month is 0.05 with highest uncertainties in forested areas. Reliable fAPAR values cannot be

Table 1. Summary description of the benchmark datasets.

Dataset	Variable	Type	Period	Comparison	Reference
SeaWiFS	Fraction of absorbed photosynthetically active radiation (fAPAR)	Gridded	1998–2005	Annual average, seasonal phase and concentration, inter-annual variability	Gobron et al. (2006)
ISLSCP II vegetation continuous fields	Vegetation fractional cover	Gridded	Snapshot – 1992/1993	Fractional cover of bare ground, herbaceous and tree; comparison of tree cover split into evergreen or deciduous, and broadleaf or needleleaf	DeFries and Hansen (2009)
Combined net primary production	Net primary production (NPP)	Site	Various 1950–2006	Direct comparison with grid cell in which site falls	Luyssaert et al. (2007), Olson et al. (2001)
Luyssaert gross primary production	Gross primary production (GPP)	Site	Various 1950–2006	Direct comparison with grid cell in which site falls	Luyssaert et al. (2007)
Canopy height	Annual average height	Gridded	2005	Direct comparison	Simard et al. (2011)
GFED3	Fractional burnt area	Gridded	1997–2006	Annual average, seasonal phase and concentration, inter-annual variability	Giglio et al. (2010)
River discharge	River discharge (at or near river mouths)	Site	1950–2005 for LPJ and LPX; 1998–2005 for all models	Annual average discharge per river basin, inter-annual variability in global runoff	Dai et al. (2009)
CDIAC atmospheric CO_2 concentration	Atmospheric CO_2 concentration	Site	1998–2005	Seasonal phase and concentration	CDIAC: cdiac.ornl.gov
CO_2 inversions	Atmospheric CO_2 concentration	Site	1980–2006	Inter-annual comparisons	Keeling (2008), Bousquet et al. (2000), Rödenbeck et al. (2003), Baker et al. (2006), Chevalier et al. (2010)

obtained for times when the solar incidence angle is $> 50°$. This limitation mostly affects cells at high latitudes, or with complex topography, during winter. Cells where fAPAR values could not be obtained for any month were excluded from all comparisons. Annual fAPAR, which is the ratio of total annual absorbed to total annual incident PAR, is not the same as the average of the monthly fAPAR. True annual fAPAR was obtained by averaging monthly values weighted by PAR. Monthly PAR values were calculated using Clime Research Unit (CRU) TS3.1 monthly fractional cloud cover (Jones and Harris, 2012) as described in Gallego-Sala et al. (2010). Monthly and annual fAPAR values were used for annual average, inter-annual variability and seasonality comparisons. The monthly fAPAR data are used as a driver for the SDBM, but as a benchmark for the DGVMs.

2.2.2 Vegetation cover

Fractional cover data (Table 1) were obtained from International Satellite Land-Surface Climatology Project (ISLSCP) II vegetation continuous field (VCF) remotely sensed product (Hall et al., 2006; DeFries and Hansen, 2009 and references therein). The VCF product provides separate information on life form, leaf type and leaf phenology at 0.5° resolution for 1992–1993. There are three categories in the life-form dataset: tree (woody vegetation > 5 m tall), herbaceous (grass/herbs and woody vegetation < 5 m), and bare ground cover. Leaf type (needleleaf or broadleaf) and phenology (deciduous or evergreen) is only given for cells that have some tree cover. Tree cover greater than 80 % is not well delineated due to saturation of the satellite signal, whereas tree cover of less than 20 % can be inaccurate due to the influence of soil and understorey on the spectral signature (DeFries et al., 2000).

The 0.5° dataset was derived from a higher resolution (1 km) dataset (DeFries et al., 1999). Evaluation of the 1 km dataset against ground observations shows it reproduces the distribution of the major vegetation types: the minimum correlation is for bare ground at high latitudes ($r^2 = 0.79$) whereas grasslands and forests have an r^2 of 0.93.

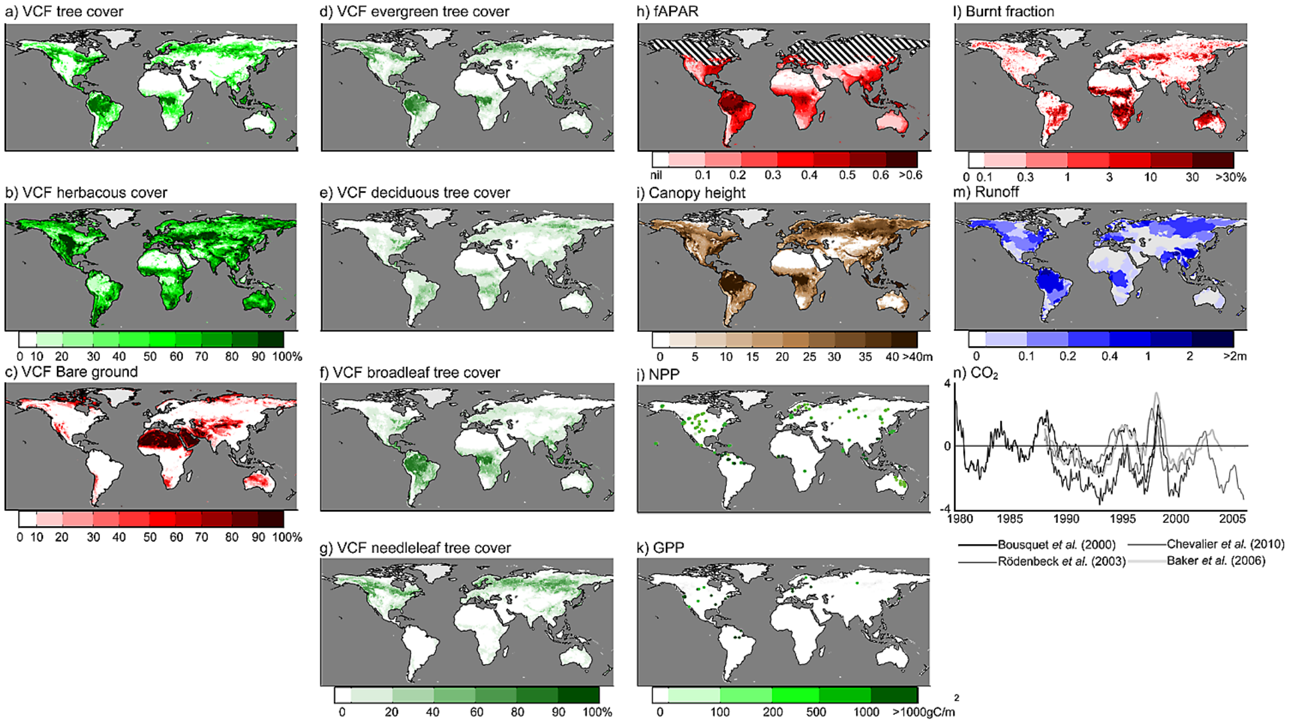

Fig. 1. Illustration of the benchmark datasets: ISLSCP II continuous vegetation fields based on a snapshot for 1992–1993 (DeFries and Hansen, 2009) give the proportions of **(a)** woody vegetation > 5 m in height (tree), **(b)** grass/herb and woody vegetation < 5 m (herbaceous), and **(c)** bare ground; for areas with tree cover, the datasets also give the proportion of **(d)** evergreen, **(e)** deciduous, **(f)** broadleaf and **(g)** needleleaf; **(i)** annual average fAPAR value for 1998–2005 from SeaWiFS (Gobron et al., 2006); **(j)** annual average burnt fraction for 1997–2006 from the GFED3 dataset (Giglio et al., 2010); **(k)** sites with measurements of net primary production, NPP and **(l)** measurements of gross primary production, GPP are both from the Luyssaert et al. (2007) dataset; **(m)** global atmospheric CO_2 concentrations for 1980–2005 based on inversion datasets (Bousquet et al., 2000; Rödenbeck et al., 2003; Baker et al., 2006; Chevalier et al., 2010); **(n)** annual average river runoff from 1950–2005 from the Dai et al. (2009) dataset, displayed over associated GRDC basins (http://www.bafg.de/GRDC); and **(m)** vegetation height based on a snapshot from 2005 (Simard et al., 2011). Hashed area in **(g)** shows areas without comparison data.

2.2.3 NPP

The NPP dataset (Table 1) was created by combining site data from the Luyssaert et al. (2007) and the Ecosystem Model/Data Intercomparison (EMDI: Olson et al., 2001) databases. We exclude sites from managed or disturbed environments; i.e. we do not use class B records from EMDI, and we exclude sites classified as “managed”, “recently burnt”, “recently cut clear”, “fertilized” or “irrigated” in Luyssaert et al. (2007) . The Luyssaert et al. (2007) data used here are all from woody biomes, and all but two of the EMDI data used are from grasslands. The NPP estimates in Luyssaert et al. (2007) were obtained by summing direct measurements of the following: (a) year-round leaf litter collection, (b) stem and branch NPP (from measurements of basal area, scaled using allometric equations), (c) fine root NPP from soil coring, isotopic turnover estimates or upscaling of root length production as observed in mini-rhizotrons, or indirectly via soil respiration, and (d) understorey NPP through destructive harvests. The uncertainty in the NPP estimate is provided for each site, and ranges from 110–656 $g\,C\,m^{-2}$ depending on the latitude, data collection and analysis methods. The NPP estimates in the EMDI database were collected from the published literature, and therefore derived using a similar variety of methodologies as used in the Luyssaert et al. (2007) compilation. The individual studies were divided into 2 classes based on an assessment of data quality. Here, we use only the top class (class A), which represents sites that are geolocated, have basic environmental metadata, and have NPP measurements on both above- and below-ground components. The EMDI database does not include estimates of the uncertainties associated with individual sites.

2.2.4 GPP

GPP data were obtained from the Luyssaert et al. (2007) database, and are estimated from flux tower (eddy covariance) measurements. The sites used here are, again, only representative of woody biomes. The uncertainty of the site-based estimates ranges from 75–677 $g\,C\,m^{-2}$, again depending on latitude, data collection and analysis methods.

2.2.5 Canopy height

The forest canopy height dataset (Table 1; Simard et al., 2011) is derived from Ice, Cloud, and land Elevation Satellite/Geoscience Laser Altimeter System (ICESat/GLAS) estimates of canopy height and its relationship with forest type, MODIS percent tree cover product (MOD44B), elevation and climatology variables (annual mean and seasonality of precipitation and temperature). Only GLAS and MODIS data from 2005 were used. The canopy height product was validated with globally distributed field measurements. Canopy height ranges from 0 to 40 m, and uncertainty is of the order of 6 m (root mean squared error). There are no estimates of the uncertainty for individual grid cells.

2.2.6 Burnt fraction

Burnt fraction data (Table 1) were obtained for each month from 1997–2006 from the third version of the Global Fire Emissions Database (GFED3: Giglio et al., 2010). Burnt fraction was calculated from high-resolution, remotely sensed daily fire activity and vegetation production using statistical modelling. Quantitative uncertainties in the estimates of burnt fraction, provided for each grid cell, are a combination of errors in the higher resolution fire activity data and errors associated with the conversion of these maps to low-resolution burnt area.

2.2.7 River discharge

River discharge (Table 1) was obtained from monthly measurements at station gauges between 1950 and 2005 (Dai et al., 2009). Dai et al. (2009) use a model-based infilling procedure in their analyses, but the dataset used here is based only on the gauge measurements. The basin associated with gauges close to a river mouth was defined using information from the Global Runoff Data Centre (GRDC: http://www.bafg.de/GRDC). Average runoff for the basin was obtained by dividing discharge by total basin area. Although individual gauge measurements may have measurement errors of the order of 10–20 %, the use of spatially integrated discharge values means that the uncertainties are considerably less than this (Dai et al., 2009). Annual average and inter-annual variability comparisons for runoff were made only for years in which there were 12 months of data, to avoid seasonal biases.

2.2.8 CO_2 concentration

CO_2 concentration (Table 1) data were taken from 26 Carbon Dioxide Information Analysis Center (CDIAC: cdiac.ornl.gov) stations (Fig. 3) for seasonal cycle comparisons. For inter-annual comparisons, we used several inversion products (Bousquet et al., 2000; Rödenbeck et al., 2003; Baker et al., 2006; Keeling, 2008; Chevalier et al., 2010), processed as in Prentice et al. (2011). The inversions are designed to isolate the component of variability in the CO_2 growth rate due to land–atmosphere exchanges. The differences between these inversions (maximum difference 3.8 ppm) give a measure of the associated uncertainty.

2.3 Metrics

Many measures with different properties are used in the geosciences literature to compare modelled and observed quantities. These typically fall into three categories: non-normalised metrics; metrics normalised by observational uncertainty; and metrics normalised by observational variance. Non-normalised metrics, which include RMSE (used e.g. by Blyth et al., 2009, 2011) and mean squared error (MSE), cannot be compared directly between different variables as they are in different units. Metrics normalised by observational uncertainty require uncertainty estimates to be given for each site/grid cell in a dataset. Most of the datasets used in this study do not have such estimates, ruling out the use of metrics normalised by observational uncertainty. We therefore use metrics normalised by observational variance, allowing metrics based on both mean deviations (modulus-based) and mean squared deviations as alternative "families".

The mean, variance and standard deviation provide a basic measure of global agreement between model and observation. Our basic normalised metrics for taking the geographic patterning into account in data–model comparisons of annual averages or totals were the normalised mean error (NME) and the normalised mean squared error (NMSE) (for definitions, limits and applications, see Table 2):

$$\mathrm{NME} = \sum\nolimits_i |y_i - x_i| / \sum\nolimits_i |x_i - \bar{x}| , \qquad (1)$$

$$\mathrm{NMSE} = \sum\nolimits_i (y_i - x_i)^2 / \sum\nolimits_i (x_i - \bar{x})^2 , \qquad (2)$$

where y_i is the modelled value of variable x in grid cell (or at site) i, x_i the corresponding observed value, and $\bar{x}$ the mean observed value across all grid cells or sites. NMSE is equal to the one-complement of the Nash–Sutcliffe model efficiency metric (Nash and Sutcliffe, 1970). NMSE thus conveys the same information as the Nash–Sutcliffe metric. As NME and NMSE are normalised by the spatial variability of the observations, these scores provide a description of the spatial error of the model. NME differs from NMSE only in the use of mean deviations, which are less sensitive to extreme values than standard deviations. We prefer NME, but retain NMSE because of its direct relation to a metric established in the literature. Both metrics take the value zero when agreement is perfect, unity when agreement is equal to that expected when the mean value of all observations is substituted for the model, and values > 1 when the model's performance is worse than the null model.

Table 2. Summary description of the benchmark metrics. y_i is the modelled and x_i is the corresponding observed value in cell or site i, and $\bar{x}$ is the mean observed value across all grid cells or sites. ω_i is the modelled phase, and φ_i is the observed phase. q_{ij} is the modelled and p_i observed proportion of item j in cell or site i.

Metric	Equation	Limits	Use in this study
Normalised mean error (NME) Normalised mean squared error (NMSE)	$\mathrm{NME} = \sum_i \lvert y_i - x_i \rvert / \sum_i \lvert x_i - \bar{x} \rvert$ $\mathrm{NMSE} = \sum_i (y_i - x_i)^2 / \sum_i (x_i - \bar{x})^2$	0 – Perfect agreement 1 – Model performs as well as observational mean 2 – complete disagreement for step 3 Infinity – complete disagreement for step 1 and 2	For burnt fraction and fAPAR: annual averages, phase concentration, inter-annual variability. For runoff: annual averages, inter-annual variability For CO_2: phase concentration For NPP, GPP and height: annual averages
Mean phase difference (MPD)	$\mathrm{MPD} = (1/\pi) \arccos\left[\cos(\omega_i - \phi_i)/n\right]$	0 – in phase 1 – 6 months out (out of phase)	Assessing difference in seasonality for fAPAR, burnt fraction and CO_2
Manhattan metric (MM) Squared chord distance (SCD)	$\mathrm{MM} = \sum_{ij} \left\lvert q_{ij} - p_{ij} \right\rvert / n$ $\mathrm{SCD} = \sum_{ij} \left(\sqrt{q_{ij}} - \sqrt{p_{ij}}\right)^2 / n$	0 – Perfect agreement 2 – Perfect disagreement	Vegetation cover comparisons for life forms, tree, grassland, bare ground, evergreen vs. deciduous tree and broadleaf vs. needleleaf tree.

Table 3. Mean, absolute variance (as defined in Eq. 3) and standard deviation (SD) of the annual average values of observations. The variance for most variables is from the long-term mean of the gridded or site data, whereas CO_2 is the variance of the inter-annual differences.

Variable	Measure	Mean	Variance	SD
Fraction of photosynthetically active radiation (fAPAR)	Annual average fAPAR	0.18	0.18	0.20
Vegetation cover	Tree cover	0.22	0.22	0.26
	Herb cover	0.52	0.25	0.29
	Bare ground	0.20	0.24	0.30
	Evergreen	0.44	0.33	0.37
	Needleleaf	0.59	0.41	0.43
Net primary production (NPP)	Annual average NPP	688	242	325
Gross primary production (GPP)	Annual average GPP	1540	642	820
Canopy height	Annual average canopy height	18.3	11.8	13.7
Burnt fraction	Annual average burnt fraction	0.028	0.043	0.094
runoff	Annual average 1950–2005	307	12	15
	Annual average 1998–2005	331	8.4	10.6
Atmospheric CO_2 concentration	Bousquet	N/A	0.93	1.10
	Rödenbeck	N/A	0.89	1.13
	Baker	N/A	0.86	1.09
	Chevalier	N/A	0.86	1.06
	Average (all inversions)	N/A	0.919	1.11

2.3.1 Annual average

Annual average comparisons were made using the mean, mean deviation (Eq. 3) and standard deviation of simulated and observed values (Table 3). NME and NMSE comparisons were conducted in three stages: (1) x_i and y_i take modelled and observed values; (2) x_i and y_i become the difference between observed or modelled values and their respective means ($x_i \to x_i - \bar{x}$); and (3) x_i and y_i from step 2 are divided by either the mean deviation or standard deviation ($x_i \to x_i / d(x)$):

$$\text{for NME}, d_{\text{NME}}(x) = \sum\nolimits_i |x_i - \bar{x}|/n; \tag{3}$$

$$\text{for NMSE}, d_{\text{NMSE}}(x) = \sqrt{\sum\nolimits_i (x_i - \bar{x})^2/n}. \tag{4}$$

Stage 2 removes the influence of the mean, and stage 3 removes the influence of the variability, on the measure. The NMSE at stage 3 is related to the correlation coefficient (Barnston et al., 1992). Van Oijen et al. (2011) showed that MSE can be decomposed into three elements similar to stage 1, 2 and 3 here, but as MSE is not normalised the decomposition is not directly applicable for this study.

2.3.2 Inter-annual variability

Inter-annual variability comparisons were made by calculating global values for each year of the model output and observations, and comparing them using Eqs. (1) and (2), but with y_i now being the global sum of modelled values for year i, and x_i the corresponding observed value. Only stage 2 and 3 comparisons were made, as the stage 1 provides no extra information from the annual-average comparisons. Stage 3 comparison measures whether a model has the correct timing or phasing of inter-annual peaks and troughs. For inter-annual CO_2 concentration, the observational data were detrended to remove the effect of anthropogenic emissions.

2.3.3 Seasonality

The seasonal expression of change can be characterised in terms of the length and timing of the season, as well as the magnitude of differentiation between seasons. For example, in simulating the fire regime at a particular place, the length of the fire season and the time that fires occur are as important as correctly predicting the area burnt. Seasonality comparisons were conducted in two parts: seasonal concentration (which is inversely related to season length) and phase (expressing the timing of the season). Each simulated or observed month was represented by a vector in the complex plane, the length of the vector corresponding to the magnitude of the variable for each month and the directions of the vector corresponding to the time of year:

$$\theta_t = 2\pi\,(t-1)/12, \tag{5}$$

where θ_t is the direction corresponding to month t, with month 1 (January) arbitrarily set to an angle of zero. A mean vector L was calculated by averaging the real and imaginary parts of the 12 vectors, x_t.

$$L_x = \sum\nolimits_t x_t \cos(\theta_t) \text{ and } L_y = \sum\nolimits_t x_t \sin(\theta_t) \tag{6}$$

The length of the mean vector divided by the annual value stands for seasonal concentration, C; its direction stands for phase, P:

$$C = \frac{\sqrt{L_x^2 + L_y^2}}{\sum_t x_t}; \tag{7}$$

$$P = \arctan\left(L_x / L_y\right). \tag{8}$$

Thus, if the variable is concentrated all in one month, seasonal concentration is equal to 1 and the phase corresponds to that month. If the variable is evenly spread over all months, then concentration is equal to zero and phase is undefined. If either modelled or observed values have zero values for all months in a given cell or site, then that cell/site is not included in the comparisons. Concentration comparisons use Eqs. (1) and (2) and steps 1, 2 and 3. Modelled and observed phase are compared using mean phase difference (MPD):

$$\text{MPD} = (1/\pi) \arccos[\cos(\omega_i - \phi_i)/n], \tag{9}$$

where ω_i is the modelled phase, and φ_i is the observed phase. The measure can be interpreted as the average timing error, as a proportion of the maximum error (6 months). For seasonal CO_2 concentrations, where the data are monthly deviations from the mean CO_2, we compared the seasonal amplitude instead of seasonal concentration by comparing the simulated and observational sum of the absolute CO_2 deviation for each month using Eqs. (1) and (2).

2.3.4 Relative abundance

Relative abundance was compared using the Manhattan metric (MM) and squared chord distance (SCD) (Gavin et al., 2003; Cha, 2007):

$$\text{MM} = \sum\nolimits_{ij} \left|q_{ij} - p_{ij}\right|/n; \tag{10}$$

$$\text{SCD} = \sum\nolimits_{ij} \left(\sqrt{q_{ij}} - \sqrt{p_{ij}},\right)^2/n \tag{11}$$

where q_{ij} is the modelled abundance (proportion) of item j in grid cell i, p_i the observed abundance of item j in grid cell i, and n the number of grid cells or sites. So in the case of comparing life forms, items j would be trees; herbaceous; and bare ground. The sum of items in each cell must be equal to one for these metrics to be meaningful. They both take the value of 0 for perfect agreement, and 2 for complete disagreement.

2.3.5 Null models

To facilitate interpretation of the scores, we compared each benchmark dataset to a dataset of the same size, filled with the mean of the observations (Table 4). We also compared each benchmark dataset with "randomized" datasets (Table 4). This was done using a bootstrapping procedure (Efron, 1979; Efron and Tibshirani, 1993), whereby we constructed a dataset of the same dimensions as the benchmark set, filled by randomly resampling the cells or sites in the original dataset with replacement. We created 1000 randomized datasets to estimate a probability density function of their scores (Fig. 2). Models are described as better/worse than randomized resampling if they were less/more than two standard deviations from the mean randomized score.

As NME and MM are the sum of the absolute spatial variation between the model and observations, the comparison of scores obtained by two different models shows the relative magnitude of their biases with respect to the observations, or how much "better" one model is than another. If a model has an NME score of 0.5, for example, its match to the observations is 50 % better than the mean of the data score of 1.0. Similarly, when this model is compared to a model with an NME score of 0.75, it can be described as 33 % better than the second model as its average spatial error is 0.5/0.75 = 67 % the size. Conversely, the second model would need to reduce its errors/improve by 33 % in order to provide as good a match to observations as the first.

2.4 Models

2.4.1 SDBM

The SDBM simulates NPP and heterotrophic respiration (R_h) as described in Knorr and Heimann (1995) while the embedded water-balance calculation models evapotranspiration and therefore implicitly runoff. NPP is obtained from a simple relationship:

$$\mathrm{NPP} = \varepsilon \cdot \mathrm{fapar} \cdot \mathrm{Ipar} \cdot \alpha, \tag{12}$$

where ε is light-use efficiency, set at 1 g C MJ^{-1}; *Ipar* is incident PAR; and α is the ratio of actual to equilibrium evapotranspiration, calculated as in Prentice et al. (1993) and Gallego-Sala et al. (2010). R_h was calculated as a function of temperature and water availability and for each cell is assumed to be equal to NPP each year (i.e. assuming the respiring pool of soil carbon is in equilibrium):

$$R_\mathrm{h} = \beta \cdot Q_{10}^{T/10} \cdot \alpha, \tag{13}$$

where Q_{10} is the slope of the relationship between ln(R_h) and temperature (expressed in units of proportional increase per 10 K warming) and takes the value of 1.5; and T is temperature (°C). β is calculated by equating annual R_h and annual NPP:

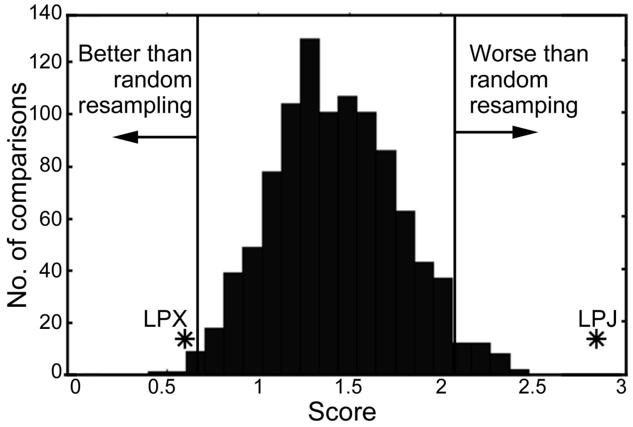

Fig. 2. Results of bootstrap resampling of inter-annual variability in global burnt fraction (1997–2005) from the GFED3 dataset. The asterisks labelled LPX and LPJ show the scores achieved by the LPX and LPJ models respectively. The limits for better than and worse than random resampling are set at two standard deviations away from the mean bootstrapping value (vertical lines).

$$\beta = \frac{\sum_t \mathrm{NPP}_t}{\sum_t Q_{10}^{T_t/10} \cdot \alpha_t}. \tag{14}$$

GPP was assumed to be twice simulated NPP (Poorter et al., 1990). Runoff was assumed to be the difference between observed precipitation and evapotranspiration. Groundwater exchanges are disregarded. The free parameters ε and Q_{10} were assigned values of 1.0 and 1.5 respectively, following Knorr and Heimann (1995) who obtained these values by tuning to observed seasonal cycles of $\mathrm{CO_2}$.

2.4.2 LPJ

LPJ (version 2.1: Gerten et al., 2004) simulates the dynamics of terrestrial vegetation via a representation of biogeochemical processes, with different properties prescribed for a small set of plant function types (PFTs). Each PFT is described by its life form (trees or herbaceous), leaf type (needleleaf or broadleaf) and phenology (evergreen or deciduous). A minimal set of bioclimatic limits constrain the global distribution of the PFTs. Nested time steps allow different processes to be simulated at different temporal resolution: photosynthesis, respiration and water balance are calculated on a daily time step while carbon allocation and PFT composition are updated on an annual time step. A weather generator converts monthly data of precipitation and fractional rain days to a daily time series of precipitation amounts. Fire is calculated annually and is based upon a simple empirical model which calculates the probability of fire based on daily moisture content of the uppermost soil layer as a proxy for fuel moisture (Thonicke et al., 2001). Assuming ignitions are always available, burnt fraction and its associated carbon fluxes

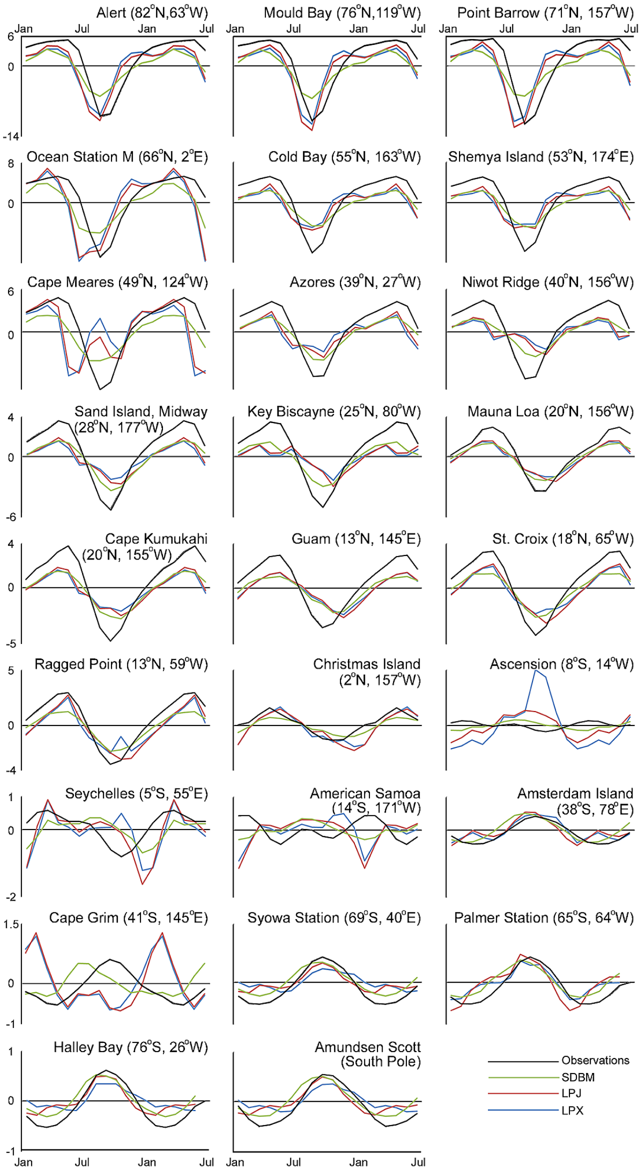

Fig. 3. Observed seasonal cycle of atmospheric CO_2 concentrations at 26 CO_2 stations over the period 1998–2005 (black line), taken from the CDIAC website (cdiac.ornl.gov) compared to the simulated seasonal cycle from the Simple Diagnostic Biosphere Model (SDBM) (green line); LPJ (red); and LPX (blue). The y-axis indicates variation in atmospheric CO_2 concentration about the mean. The x-axis is months from January through 18 months to June.

are calculated from the summed annual probability of fire, using a simple relationship.

2.4.3 LPX

LPX (Prentice et al., 2011), which is a development of LPJ-SPITFIRE (Thonicke et al., 2010), incorporates a process-based fire scheme, with ignition rates based on the seasonal distribution of lightning strikes and fuel moisture content and fire spread, intensity and residence time based on climate data and modelling the drying of different fuel types between rain days. Fire intensity influences fire mortality and carbon fluxes. The fire model runs on a daily time step.

2.5 Model protocol

All models were run on a 0.5° global grid using the CRU TS3.0 land mask as in Prentice et al. (2011). Soil texture was prescribed using the FAO soil data (FAO, 1991). The spin-up and historical drivers for the DGVM simulations were exactly as used for LPX by Prentice et al. (2011). For comparability, the same climate data were used to drive the SDBM. In addition SDBM was driven by fAPAR values from SeaWiFS observations. For cells lacking fAPAR values, values were constructed for the missing months by fitting the following equation to available data for each year:

$$\mathrm{fAPAR}(m) = \frac{1}{2}\{(U-L)\cos[2\pi\,(m-m_{\max})/12] + U + L,\} \quad (15)$$

where fAPAR(m) is the fAPAR for months m with data; U is the maximum fAPAR value in month $m_{\max}$; and L is the minimum fAPAR value. As the maximum fAPAR value typically occurs in spring or summer (Prince, 1991) when SeaWiFS data are generally available, and the minimum occurs in the winter when data may be unavailable, U is set to the highest fAPAR value, whilst L is tuned to fit the function to the data.

The SDBM was only run for 1998–2005, a limitation imposed by the availability of fAPAR data, and comparisons were confined to this period. For LPX and LPJ, outputs and therefore comparisons were possible from 1950–2006. Comparisons with NPP, GPP, annual average basin runoff, global inter-annual variability in runoff, and the seasonal cycle of CO_2 concentration were made for all three models. LPX and LPJ are compared across a wider range of benchmarks.

Comparisons of the seasonal CO_2 cycle were based on simulated monthly net ecosystem production (NEP: NPP − R_h − fire carbon flux). NEP for the SDBM was taken as the difference between monthly NPP and R_h. For LPJ, which simulates fire on an annual basis, monthly fire carbon flux was set to 1/12 the annual value. With LPX, it was possible to use monthly fire carbon flux. For each model, detrended monthly values of NEP for each grid cell were input into the atmospheric transport matrices derived from the TM2 transport model (Kaminski et al., 1996), which allowed us to derive the CO_2 seasonal cycle (Heimann, 1995; Knorr and Heimann, 1995) at the locations of the observation sites.

Average basin runoff was calculated by summing the runoff from all model grid cells within a GRDC-defined basin and dividing by the basin area. If a grid cell fell into more than one GRDC basin, the runoff was divided between basins in proportion to the fraction of the cell within each basin. Inter-annual changes in runoff were calculated by summing runoff over all cells in basins for which there were data for a given year. Seasonal cycles of runoff are dependent on the dynamics of water transport in the river, which was not modelled.

3 Results

3.1 Benchmark results

3.1.1 fAPAR

LPJ scores 0.82 and LPX scores 0.86 using NME for annual average fAPAR (Table 5). This difference in score is equivalent to a negligible (i.e. < 5 %) change in the match to the observations. Both values are considerably better than values for the mean of the data (1.00) and random resampling (1.19 ± 0.004), with the match to observations being 15 % closer and 30 % closer respectively. The models also perform well for seasonal timing (Fig. 4), with scores of 0.19 (LPJ) and 0.18 (LPX) or the equivalent of an average of 34 days different from observations. For comparison, the seasonal timing of the mean of the data and random resampling is ca. 3 months different from observations. The models also perform well for inter-annual variability: LPJ scores 0.60 and LPX scores 0.50 using NME for inter-annual variability, compared to a mean score of 1.00 and a score of 1.21 ± 0.34 from random resampling. The DGVM scores represent, respectively, a 40 % and 50 % better match to observations than the mean of the data. LPJ scores 1.07 and LPX scores 1.14 using NME for seasonal concentration, compared to 1.00 for the mean and 1.41 ± 0.006 for random resampling. This means that the seasonal concentration of fapar in the DGVMs is, respectively, 7 % and 14 % worse than the mean of the data compared to observations.

3.1.2 Vegetation cover

LPJ scores 0.78 and LPX scores 0.76 using the MM for the prediction of life forms (Table 5), again a negligible difference in performance (< 3 %) compared to observations. Both values are better than obtained for the mean of the data (0.93) or by randomly resampling (0.88 ± 0.002). Both models were also better than mean and randomly resampling for bare ground. However, both models overestimate tree cover and underestimate herb cover by around a factor of 2 (Table 5). The scores for tree cover (LPJ: 0.62, LPX: 0.56) show, respectively, a 38 % and 24 % poorer match to observations than the mean of the data (0.45), and a 15 % and 4 % poorer match to observations than randomly resampling (0.54 ± 0.002). In the same way, the two DGVMs show a 40 % poorer match to observed grass cover than the mean of the data and a 6 % poorer match than randomly resampling. Both models are worse than mean and random resampling for phenology (Table 5).

3.1.3 NPP/GPP

The models have NME scores for NPP of 0.58 (SDBM), 0.83 (LPJ) and 0.81 (LPX) (Table 5) – better than values obtained for the mean of the data (1.00) and random resampling (1.35 ± 0.09). Removing the biases in mean and variance (Table 5) improves the performance of all three models. The SDBM simulates 1.13 times higher NPP than observed, but correctly predicts the spatial variability shown by the observations, whereas the two DGVMs overestimate NPP but underestimate the spatial variance in NPP. As a result, removing the biases in the mean produces a much larger improvement in the DGVMs. In LPJ, for example, the score goes from 0.83 to 0.69, equivalent to a 29 % better match to the observations. The improvement in the SDBM is equivalent to only a 9 % better match to observations. The two DGVMs perform worse for GPP than NPP. LPX has an NME score of 0.81 for NPP but 0.98 for GPP – this is equivalent to a 17 % better match to NPP observations than to GPP observations. The SDBM performs better for GPP than the DGVMs, obtaining an NME score of 0.71, which is better than the mean of the data (1.00) and randomly resampling (1.36 ± 0.22).

3.1.4 Canopy height

LPJ scores 1.00 and LPX scores 1.04 using NME for the prediction of height (Table 5). These values lie between the mean (1.00) and random resampling (1.33 ± 0.004) scores. This poor performance is due to modelled mean heights ca. 60–65 % lower than observed and muted variance (Table 5, Fig. 6). Removing the mean bias improves the score for both DGVMs to 0.71 for LPJ and 0.73 for LPX, equivalent to a 29 % and 30 % improvement in the match to observations. Model performance is further improved by removing bias in the variance, to 0.64 for LPJ (ca. 11 %) and 0.68 for LPX (ca. 6 %).

3.1.5 Burnt fraction

There is a major difference between the two DGVMs for annual fractional burnt area (Fig. 7): LPJ scores 1.58, while LPX scores 0.85 for NME (Table 5). Thus, LPX produces a 46 % better match to the observations than LPJ. The LPJ score is worse than the mean (1.00) and random resampling (1.02 ± 0.008). The same is true for NME comparisons of inter-annual variability, with LPJ scoring 2.86, worse than the mean (1.00) and random resampling (1.35 ± 0.34), whereas the LPX score of 0.63 is better than both. LPX could also be benchmarked against the seasonality of burnt fraction. It scores 0.10 for MPD comparison of phase, much better than the mean (0.74) and random resampling (0.47 ± 0.001). However, it did not perform well for seasonal concentration, scoring 1.38 compared to the mean (1.00) and random resampling (1.33 ± 0.006).

Table 4. Scores obtained using the mean of the data (Data mean), and the mean and standard deviation of the scores obtained from bootstrapping experiments (Bootstrap mean, Bootstrap SD). NME/NMSE denotes the normalised mean error/normalised mean squared error, MPD the mean phase difference and MM/SCD the Manhattan metric/squared chord distance metrics.

Variable	Metric used	Measure	Absolute			Square		
			Data mean	Bootstrap mean	Bootstrap SD	Data mean	Bootstrap mean	Bootstrap SD
fAPAR	NME/ NMSE	Annual average	1.00	1.19	0.004	1.00	1.95	0.01
		– with mean removed	1.00	1.21	0.003	1.00	1.93	0.01
		– with mean and variance removed	1.00	1.23	0.004	1.00	2.08	0.01
		Inter-annual variability	1.00	1.21	0.34	1.00	1.92	0.79
		– with variance removed	1.00	1.30	0.36	1.00	2.15	0.84
		Seasonal concentration	1.00	1.41	0.006	1.00	2.02	0.02
		– with mean removed	1.00	1.41	0.006	1.00	2.02	0.02
		– with mean and variance removed	1.00	1.40	0.005	1.00	2.00	0.01
	MPD	Phase	0.54	0.49	0.001	N/A	N/A	N/A
Vegetation cover	MM	Life forms	0.93	0.88	0.002	0.37	0.47	0.002
		Tree vs. non-tree	0.45	0.54	0.002	0.14	0.27	0.001
		Herb vs. non-herb	0.50	0.66	0.002	0.17	0.33	0.002
		Bare ground vs. covered ground	0.48	0.56	0.002	0.18	0.35	0.002
		Evergreen vs. deciduous	0.68	0.87	0.003	0.30	0.580	0.003
		Broadleaf vs. needleleaf	0.77	0.94	0.004	0.36	0.75	0.004
Net primary production	NME/ NMSE	Annual average	1.00	1.35	0.09	1.00	2.00	0.24
		– with mean removed	1.00	1.35	0.09	1.00	2.00	0.24
		– with mean and variance removed	1.00	1.35	0.08	1.00	2.01	0.20
Gross primary production	NME/ NMSE	Annual average	1.00	1.36	0.22	1.00	2.01	0.56
		– with mean removed	1.00	1.36	0.22	1.00	2.00	0.55
		– with mean and variance removed	1.00	1.36	0.17	1.00	2.00	0.43
Canopy height	NME/ NMSE	Annual average	1.00	1.33	0.004	1.00	1.98	0.009
		– with mean removed	1.00	1.33	0.004	1.00	1.98	0.009
		– with mean and variance removed	1.00	1.33	0.004	1.00	2.00	0.009
Burnt fraction	NME/ NMSE	Annual average	1.00	1.02	0.008	1.00	1.98	0.03
		– with mean removed	1.00	1.09	0.005	1.00	1.99	0.03
		– with mean and variance removed	1.00	1.14	0.004	1.00	2.36	0.02
		Inter-annual variability	1.00	1.35	0.34	1.00	1.93	0.77
		– with variance removed	1.00	1.39	0.32	1.00	2.15	0.76
		Seasonal concentration	1.00	1.33	0.006	1.00	1.99	0.01
		– with mean removed	1.00	1.33	0.006	1.00	1.99	0.02
		– with mean and variance removed	1.00	1.33	0.005	1.00	2.00	0.01
	MPD	Phase	0.74	0.47	0.001	N/A	N/A	N/A
Runoff	NME/ NMSE	Annual average 1950–2005	1.00	1.18	0.48	1.00	1.95	0.99
		– with mean removed	1.00	1.35	0.52	1.00	1.89	0.96
		– with mean and variance removed	1.00	1.76	0.71	1.00	2.02	1.03
		Annual average 1998–2005	1.00	1.17	0.28	1.00	1.97	0.94
		– with mean removed	1.00	1.27	0.33	1.00	1.96	0.93
		– with mean and variance removed	1.00	1.18	0.05	1.00	2.00	0.16
		Inter-annual variability 1950–2005	1.00	1.40	0.14	1.00	2.00	0.32
		– with variance removed	1.00	1.45	0.172	1.00	2.01	0.60
		Inter-annual variability 1998–2005	1.00	1.33	0.34	1.00	1.83	0.83
		– with variance removed	1.00	1.34	0.34	1.00	1.87	0.82

Table 4. Continued.

Variable	Metric used	Measure	Absolute			Square		
			Data mean	Bootstrap mean	Bootstrap SD	Data mean	Bootstrap mean	Bootstrap SD
Atmospheric CO_2 concentration	NME/ NMSE	Inter-annual variability – Bousquet (Jan 1980–June 1998)	1.00	1.36	0.058	1.00	2.00	0.15
		– with variance removed	1.00	1.36	0.058	1.00	2.00	0.15
		Inter-annual variability – Rödenbeck (Jan 1982–Dec 2001)	1.00	1.38	0.081	1.00	1.99	0.22
		– with variance removed	1.00	1.38	0.082	1.00	1.99	0.22
		Inter-annual variability – Baker (Jan 1988–Dec 2004)	1.00	1.39	0.07	1.00	1.99	0.19
		– with variance removed	1.00	1.40	0.07	1.00	1.99	0.19
		Inter-annual variability – Chevalier (Jul 1988–Jun 2005)	1.00	1.38	0.07	1.00	2.00	0.17
		– with variance removed	1.00	1.39	0.07	1.00	2.00	0.17
		Inter-annual variability – Average (Jan 1980–Jun 2005)	1.00	1.37	0.05	1.00	2.00	0.14
		– with variance removed	1.00	1.37	0.05	1.00	2.00	0.14
		Amplitude	1.00	1.38	0.28	1.00	2.04	0.81
		– with mean removed	1.00	1.40	0.39	1.00	2.00	0.78
		– with mean and variance removed	1.00	1.39	0.14	1.00	2.02	0.40
	NME	Phase	0.33	0.42	0.051	N/A	N/A	N/A

3.1.6 River discharge

Comparing average runoff for 1950–2005, both DGVMs score 0.28 for NME, better than the mean (1.00) and random resampling (1.18 ± 0.48). The models perform much less well for inter-annual comparisons, with NME scores of 1.33 (LPJ) and 1.32 (LPX), worse than 1.00 for the mean and 1.45 ± 0.17 for random resampling. Agreement is slightly improved by removing variance bias (LPJ: 1.07, LPX: 1.11). Neither of the DGVMs examined here treat water-routing explicitly. Introducing a one-year lag for inter-annual comparisons (Fig. 8) produces a 21 % (LPJ) and 19 % (LPX) improvement in the match to observations, confirming that taking account of delays in water transport is important when assessing the inter-annual variation in runoff. All three models were compared for 1998–2005. For annual average comparisons, they all performed better than the mean and random resampling (Table 5). However, all models performed poorly for inter-annual variability, obtaining similar scores (1.64 to 2.38) compared to the mean (1.00) and random resampling (1.34 ± 0.34). Removing variability bias and introducing a one-year lag improved performance, with the SDBM scoring 1.37, LPJ 1.36 and LPX 1.35.

3.1.7 CO_2 concentration

The generalised form of the seasonal cycle in CO_2 concentrations at different sites can be compared for all three models. The SDBM scores 0.21 whereas LPJ scores 0.34 and LPX 0.34 in the MPD comparisons of seasonal timing, compared to the mean of the data (0.33) and random resampling (0.42 ± 0.051). Thus, the SDBM produces an estimate of peak timing that is 22 days closer to observations than the mean of the data, while the DGVMs produce estimates that are 6 days further away from the observations than the mean of the data (Fig. 3). The scores for NME comparison of seasonal concentration for the SDBM (0.68), LPJ (0.46) and LPX (0.58) are all better than the mean (1.00) and random resampling (1.38 ± 0.28). Thus, although the difference between the models is non-trivial (ca. 30 %), all three models are ca. 30–50 % closer to observations than the mean of the data. Only the DGVMs can be evaluated with respect to inter-annual variability in global CO_2 concentrations. Both models capture the inter-annual variability relatively well (Table 5). LPJ scores 0.89 and LPX scores 0.83 for the average of all inversion datasets, compared to the mean of the data (1.00) and random resampling (1.37 ± 0.05).

Table 5. Comparison metric scores for model simulations against observations. Mean and variance rows show mean and variance of simulation for annual average values, followed in brackets by the ratio of the mean/variance with observed mean or variance in Table 3. Numbers in bold indicate the model with the best performance for that variable. Italic indicates model scores better than the mean of the data score listed in Table 4. Asterisks indicate model scores that are significantly better than randomly resampling listed in Table 4. NME/NMSE denotes the normalised mean error/normalised mean squared error, MPD the mean phase difference and MM/SCD the Manhattan metric/squared chord distance metrics. fAPAR is the fraction of absorbed photosynthetically active radiation, NPP net primary productivity, and GPP gross primary productivity.

Variable	Metric used	Measure	SDBM		LPJ		LPX	
			Absolute	Squared	Absolute	Squared	Absolute	Squared
fAPAR	Mean (ratio)	Annual average	N/A	N/A	0.30 (1.63)	N/A	0.26 (1.44)	N/A
	Variance (ratio)		N/A	N/A	0.15 (0.85)	0.17 (0.86)	0.16 (0.91)	0.18 (0.90)
	NME/ NMSE	Annual average	N/A	N/A	***0.82****	**1.04***	*0.86**	1.09*
		– with mean removed			***0.75****	***0.76****	*0.76**	*0.78**
		– with mean and variance removed			***0.80****	***0.83***∗	*0.82**	*0.90**
		Inter-annual variability	N/A	N/A	*0.60**	*0.36**	***0.50****	***0.27****
		– with variance removed			*0.73**	*0.36**	***0.44****	***0.23****
		Seasonal concentration	N/A	N/A	**1.07***	**1.28***	1.14*	1.37*
		– with mean removed			**1.02***	**1.20***	1.05*	1.25*
		– with mean and variance removed			**1.03***	**1.26***	1.06*	1.31*
	MPD	Phase	N/A	N/A	*0.19**	N/A	***0.18****	N/A
Vegetation cover	Mean (ratio)	Tree vs. non-tree	N/A	N/A	0.49 (2.23)	N/A	0.42 (1.91)	N/A
		Herb vs. non-herb	N/A	N/A	0.28 (0.54)	N/A	0.31 (0.60)	N/A
		Bare ground vs. covered ground	N/A	N/A	0.23 (1.14)	N/A	0.27 (1.33)	N/A
		Evergreen vs. deciduous	N/A	N/A	0.34 (0.79)	N/A	0.28 (0.73)	N/A
		Broadleaf vs. needleleaf	N/A	N/A	0.67 (1.08)	N/A	0.65 (1.10)	N/A
	Variance (ratio)	Tree vs. non-tree	N/A	N/A	0.45 (2.03)	0.45 (1.73)	0.46 (2.06)	0.46 (1.75)
		Herb vs. non-herb	N/A	N/A	0.30 (1.18)	0.35 (1.21)	0.32 (1.27)	0.36 (1.24)
		Bare ground vs. covered ground	N/A	N/A	0.30 (1.26)	0.36 (1.20)	0.32 (1.33)	0.37 (1.23)
		Evergreen vs. deciduous	N/A	N/A	0.35 (1.06)	0.39 (1.07)	0.36 (1.18)	0.41 (1.18)
		Broadleaf vs. needleleaf	N/A	N/A	0.40 (1.02)	0.43 (1.02)	0.43 (1.07)	0.46 (1.07)
	MM	Life forms	N/A	N/A	*0.78**	0.44*	***0.76****	***0.42****
		Tree vs. non-tree	N/A	N/A	0.62	0.39	**0.56**	**0.33**
		Herb vs. non-herb	N/A	N/A	0.71	0.39	**0.67**	**0.36**
		Bare ground vs. covered ground	N/A	N/A	***0.23****	***0.10****	*0.30**	*0.156**
		Evergreen vs. deciduous	N/A	N/A	**0.93**	**0.47***	0.94	0.48*
		Broadleaf vs. needleleaf	N/A	N/A	**0.89***	**0.47***	0.92*	0.55*
NPP	Mean (ratio)	Annual average	612 (1.13)	N/A	688 (1.28)	N/A	685 (1.27)	N/A
	Variance (ratio)		297 (1.00)	351 (0.96)	242 (0.81)	325 (0.887)	283 (0.95)	355 (0.97)
	NME/ NMSE	Annual average	***0.58****	***0.35****	*0.83**	*0.68**	*0.81**	*0.67**
		– with mean removed	***0.53****	***0.32****	*0.69**	*0.52**	*0.68**	*0.51**
		– with mean and variance removed	***0.53****	***0.33****	*0.75**	*0.57**	*0.69**	*0.53**

Table 5. Continued.

Variable	Metric used	Measure	SDBM		LPJ		LPX	
			Absolute	Squared	Absolute	Squared	Absolute	Squared
GPP	Mean (ratio)	Annual average	1231 (0.80)	N/A	1354 (0.88)	N/A	1127 (0.73)	N/A
	Variance (ratio)		316 (0.49)	492 (0.60)	288 (0.45)	388 (0.47)	240 (0.37)	304 (0.37)
	NME/ NMSE	Annual average	___0.71___*	___0.57___∗	_0.80_*	_0.63_*	_0.98_*	1.19*
		– with mean removed	___0.63___*	___0.40___∗	_0.82_*	_0.58_*	1.02*	_0.93_*
		– with mean and variance removed	___0.59___∗	___0.37___∗	_0.90_*	_0.63_*	1.33*	1.45*
Canopy height	Mean (ratio)	Annual average	N/A	N/A	6.92 (0.38)	N/A	6.36 (0.35)	N/A
	Variance (ratio)		N/A	N/A	6.17 (0.52)	6.70 (0.49)	6.69 (0.57)	7.18 (0.52)
	NME/ NMSE	Annual average	N/A	N/A	__1.00*__	__1.22*__	1.04*	1.29*
		– with mean removed			___0.71___*	___0.53___*	_0.73_*	_0.55_*
		– with mean and variance removed			___0.64___*	___0.50___*	_0.68_*	_0.58_*
Burnt fraction	Mean (ratio)	Annual average	N/A	N/A	0.014 (0.50)	N/A	0.022 (0.81)	N/A
	Variance (ratio)		N/A	N/A	0.016 (0.37)	0.027 (0.29)	0.032 (0.75)	0.46 (0.49)
	NME/ NMSE	Annual average	N/A	N/A	1.58	1.18	___0.85___*	__1.01*__
		– with mean removed			1.55	1.17	___0.91___*	__1.01*__
		– with mean and variance removed			1.72	1.29	___0.99___*	__1.60*__
		Inter-annual variability	N/A	N/A	2.86	8.10	___0.63___*	___0.49___
		– with variance removed			1.90	3.08	___0.77___	___0.56___
		Seasonal concentration	N/A	N/A	N/A	N/A	__1.38__	__2.00__
		– with mean removed					__1.37__	__1.99__
		– with mean and variance removed					__1.26*__	__1.77*__
	MPD	Phase	N/A	N/A	N/A	N/A	___0.10___*	N/A
Runoff	Mean (ratio)	Annual average 1950-2005	N/A	N/A	388 (1.26)	N/A	396 (1.29)	N/A
		Annual average 1998–2005	466 (1.41)	N/A	426 (1.29)	N/A	429 (1.30)	N/A
	Variance (ratio)	Annual average 1950–2005	N/A	N/A	17.8 (1.44)	22.7 (1.50)	16.6 (1.35)	21.0 (1.38)
		Annual average 1998–2005	11.9 (1.42)	15.6 (1.48)	15.9 (1.90)	18.9 (1.79)	14.3 (1.70)	17.1 (1.62)
	NME/ NMSE	Annual average 1998–2005	N/A	N/A	___0.28___*	_0.067_*	_0.28_*	___0.054___*
		– with mean removed			___0.34___*	_0.065_*	_0.35_*	___0.052___*
		– with mean and variance removed			___0.20___*	___0.021___*	_0.24_*	_0.031_*
		Annual average 1998–2005	_0.42_*	_0.28_*	___0.23___*	_0.039_*	___0.23___*	___0.026___*
		– with mean removed	_0.55_*	_0.26_*	___0.26___*	_0.039_*	___0.26___*	___0.025___*
		– with mean and variance removed	_0.22_*	_0.033_*	___0.18___*	___0.013___*	_0.20_*	_0.018_*
		Inter-annual variability 1950–2005	N/A	N/A	1.33*	1.91*	__1.32*__	__1.88*__
		– with variance removed			__1.07*__	__1.11*__	1.11*	1.25*
		Inter-annual variability 1950–2005 with 1yr lag			__1.03*__	__1.21*__	1.06*	1.19*
		– with variance removed			___0.84___*	___0.70___*	_0.90_*	_0.79_*

Table 5. Continued.

Variable	Metric used	Measure	SDBM		LPJ		LPX	
			Absolute	Squared	Absolute	Squared	Absolute	Squared
CO_2	Variance (ratio)	Inter-annual variability	__1.64__	__2.91__	2.38	4.59	2.27	4.09
		1998–2005	__1.48__	2.65	1.59	__2.21__	1.63	2.28
		– with variance removed						
		Inter-annual variability						
		1950–2005 with 1yr lag	__1.49__	__2.00__	2.10	4.23	1.94	3.64
		– with variance removed	1.37	__1.06__	1.36	1.95	__1.35__	1.95
		Inter-annual variability – Bousquet	N/A	N/A	1.12	1.35	1.09	1.37
		(Jan 1980–June 1998)			(1.21)	(1.22)	(1.18)	(1.24)
		Inter-annual variability – Rödenbeck	N/A	N/A	1.15	1.32	1.02	1.24
		(Jan 1982–Dec 2001)			(1.30)	(1.16)	(1.15)	(1.09)
		Inter-annual variability – Baker	N/A	N/A	1.11	1.30	0.94	1.16
		(Jan 1988–Dec 2004)			(1.28)	(1.19)	(1.09)	(1.07)
		Inter-annual variability – Chevalier	N/A	N/A	1.08	1.28	0.96	1.19
		(Jul 1988–Jun 2005)			(1.26)	(1.20)	(1.11)	(1.12)
	NME/ NMSE	Inter-annual variability – Bousquet	N/A	N/A	_0.98*_	__1.1*__	___0.95*___	1.1*
		(Jan 1980–June 1998)			___0.86*___	_0.82*_	_0.87*_	___0.81*___
		– with variance removed						
		Inter-annual variability – Rödenbeck	N/A	N/A	_0.82*_	_0.59*_	___0.70*___	___0.41*___
		(Jan 1982–Dec 2001)			_0.67*_	_0.48*_	___0.64*___	___0.37*___
		– with variance removed						
		Inter-annual variability – Baker	N/A	N/A	_0.85*_	_0.78*_	___0.78*___	___0.64*___
		(Jan 1988–Dec 2004)			___0.66*___	_0.62*_	_0.72*_	___0.60*___
		– with variance removed						
		Inter-annual variability – Chevalier	N/A	N/A	_0.93*_	_0.72*_	___0.73*___	___0.51*___
		(Jul 1988–Jun 2005)			_0.79*_	_0.56*_	___0.68*___	___0.44*___
		– with variance removed						
		Inter-annual variability – Average	N/A	N/A	_0.89*_	_0.82*_	___0.83*___	___0.82*___
		(Jan 1980–Jun 2005)			___0.73*___	___0.62*___	_0.74*_	_0.64*_
		– with variance removed						
		Amplitude	_0.68*_	_0.60*_	___0.46*___	___0.27*___	_0.58*_	_0.40*_
		– with mean removed	_0.50*_	_0.26∗_	___0.40*___	___0.17*___	_0.48*_	_0.25*_
		– with mean and variance removed	___0.10∗___	___0.02*___	_0.50*_	_0.23*_	_0.59*_	_0.34*_
		Phase	___0.21*___	N/A	0.34	N/A	0.34	N/A

3.2 Sensitivity tests

3.2.1 Incorporating data uncertainties

In calculating the performance metrics, we have disregarded observational uncertainty. Few land-based datasets provide quantitative information on the uncertainties associated with site or gridded values. However, the GFED burnt fraction (Giglio et al., 2010) and the Luyssaert et al. (2007) NPP datasets do provide quantitative estimates of uncertainty. We use these datasets to evaluate the impact of taking account observational uncertainty in the evaluation of model performance by calculating NME scores for annual averages of each variable using the maximum and minimum uncertainty

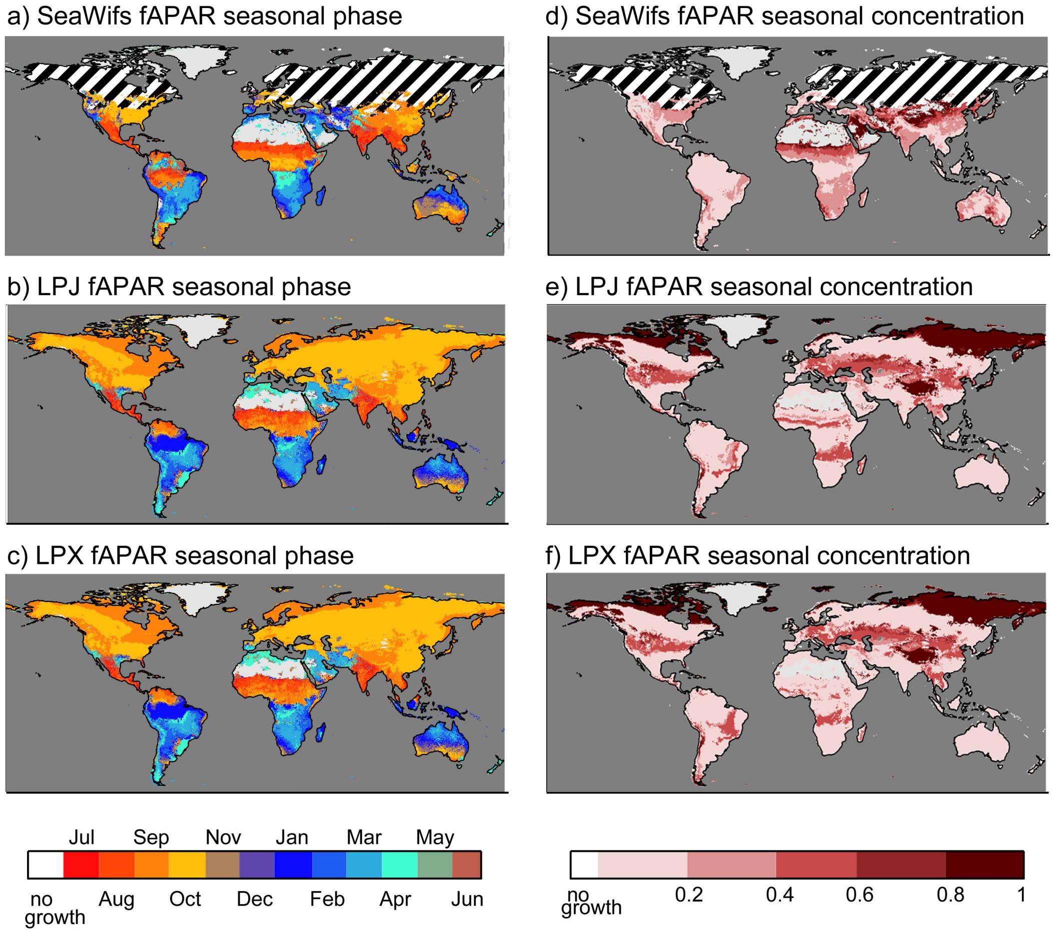
a) SeaWifs fAPAR seasonal phase
d) SeaWifs fAPAR seasonal concentration
b) LPJ fAPAR seasonal phase
e) LPJ fAPAR seasonal concentration
c) LPX fAPAR seasonal phase
f) LPX fAPAR seasonal concentration
Jul
Sep
Nov
Jan
Mar
May
no growth
Aug
Oct
Dec
Feb
Apr
Jun
no growth
0.2
0.4
0.6
0.8
1

Fig. 4. Comparison of observed and simulated seasonal phase and seasonal concentration of fraction of absorbed photosynthetically active radiation (fAPAR) averaged over the period 1998–2005 from **(a)** seasonal phase from SeaWiFS (Gobron et al., 2006) and as simulated by **(b)** LPJ and **(c)** LPX; seasonal concentration from **(d)** SeaWiFS, **(e)** LPJ and **(f)** LPX. Hashed area in **(a)** and **(d)** shows areas where no comparison is possible.

values at each site or grid cell to calculate the maximum and minimum potential distance between models and observations.

In the standard NME comparison for annual fractional burnt area, LPJ scores 1.58 while LPX scores 0.85. Taking into account the uncertainties produces minimum and maximum scores of 1.27 and 1.85 for LPJ, and 0.35 and 1.17 for LPX. Since these ranges are non-overlapping, the improvement in the match to observations shown by LPX compared to LPJ is demonstrably larger than observational uncertainty. This is not the case for the Luyssaert et al. (2007) only site-based annual average NPP comparisons, where the ranges are 0.26–1.36 (SDBM), 0.37–1.43 (LPJ) and 0.39–1.50 (LPX). Similarly, the apparent biases in mean annual NPP shown by all three models are within the observational uncertainty. Removing the slight high bias in mean annual NPP produced an improvement in the performance of the SDBM, with a change in the Luyssaert et al. (2007) only score from 0.72 to 0.59, equivalent to a 18 % better match to the observations. However, the range of scores obtained for the SDBM taking into account the observational uncertainties after removing the high bias is 0.21–1.25. As this overlaps with the scores obtained prior to removing these biases (0.26–1.36), the improvement gained from removing the influence of the mean in NPP in the SDBM is less than the observational uncertainty.

Another approach to estimating the influence of uncertainty is to use alternative realizations of the observations. This approach has been used by the climate-modelling community to evaluate performance against modern climate observations (e.g. Gleckler et al., 2008) and is used here for CO_2 inter-annual comparisons. The scores obtained in comparisons with individual inversion products range from 0.82 to 0.98 for LPJ, and from 0.70 to 0.95 for LPX. Thus, the conclusion that the two DGVMs capture the inter-annual variability equally well, based on the average scores of all inversion datasets, is supported when taking into account uncertainty expressed in the differences between the inversions.

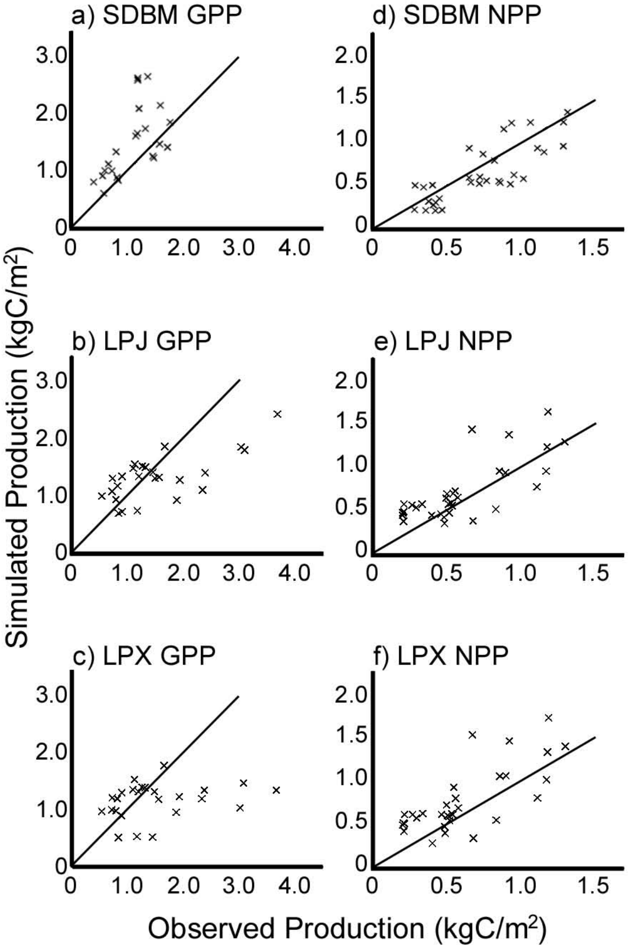

Fig. 5. Comparisons of observed and simulated NPP and GPP in kg C m^{-2}. The NPP observations (x-axis) are from the dataset made by combining sites from the Luyssaert et al. (2007) dataset and the Ecosystem/Model Data Intercomparison dataset (Olson et al., 2001). The GPP observations are derived from the Luyssaert et al. (2007) dataset. The simulated values (y-axis) are annual averages for the period 1998–2005. The observations are compared with NPP **(a)** and GPP **(b)** from the Simple Diagnostic Biosphere Model (SDBM), NPP **(c)** and GPP **(d)** from LPJ and NPP **(e)** and GPP **(f)** from LPX. The solid line shows the 1 : 1 relationship.

3.2.2 The influence of choice of dataset

The use of alternative datasets for a given variable implies that there are no grounds for distinguishing which is more reliable. It also highlights the fact that there is an element of subjectivity in the choice of datasets and that this introduces another source of uncertainty into the process of benchmarking. We have explicitly excluded from the benchmarking procedure any datasets that involve manipulations of original measurements based on statistical or physical models that are driven by the same inputs as the vegetation models being tested (e.g. MODIS NPP, remotely sensed evapotranspiration, upscaled GPP). However, such products often provide global coverage of variables that may not be as well represented in other datasets and thus could provide a useful alternative realization of the observations.

Table 6. Mean annual gross primary production (GPP) normalised mean error (NME) comparison metrics using Luyssaert et al. (2007) and Beer et al. (2010) as alternative benchmarks. In the case of Beer et al. (2010), the comparisons are made for all grid cells (global) and also from the grid cells which contain sites in the Luyssaert et al. (2007) dataset (at sites).

Variable	Measure	SDBM	LPJ	LPX
GPP from Luyssaert et al. (2007)	global at sites	N/A 0.71	N/A 0.80	N/A 0.98
GPP from Beer et al. (2010)	global at sites	0.56 0.34	0.60 0.84	0.51 0.74

Here, we test the use of the Beer et al. (2010) dataset as an alternative to the Luyssaert et al. (2007) GPP dataset. The Beer et al. (2010) GPP dataset is based on a much larger number of flux-tower measurements than are included in the Luyssaert et al. (2007) dataset, but uses both diagnostic models and statistical relationships with climate to scale up these measurements to provide global coverage. We compare the annual average GPP scores using Beer et al. (2010), calculated using all grid cells and considering only those grid cells which correspond to locations with site data in the Luyssaert et al. (2007) dataset. These comparisons (Table 6) show that LPX and SDBM perform better against the Beer et al. (2010) dataset than against the Luyssaert et al. (2007) at the site locations, while the results obtained for LPJ against the two datasets are roughly similar. There is a further improvement in performance when the models are compared against all the grid cells. The improvement in performance at the site locations presumably reflects the fact that the Beer et al. (2010) dataset smooths out idiosyncratic site characteristics; the additional improvement in performance in the global comparison reflects both the smoothing and the much larger number of flux sites included in the Beer et al. (2010) dataset. Nevertheless, the conclusion that the SDBM performs better than the DGVMs is not influenced by the choice of dataset. LPJ performs marginally better than LPX when the Luyssaert et al. (2007) dataset is used as the benchmark (0.8 versus 0.98), but worse than LPX when the Beer et al. (2010) dataset is used as a benchmark (0.6 versus 0.51). This indicates that the difference between the two DGVMs is less than the observational uncertainty.

The release of new, updated datasets poses problems for the implementation of a benchmarking system, but could be regarded as a special case of the use of alternative realizations of the observations. The GFED3 burnt fraction dataset, used here, is a comparatively recent update of an earlier burnt fraction dataset (GFED2: van der Werf et al., 2006). When LPJ and LPX are evaluated against GFED2, the NME score for

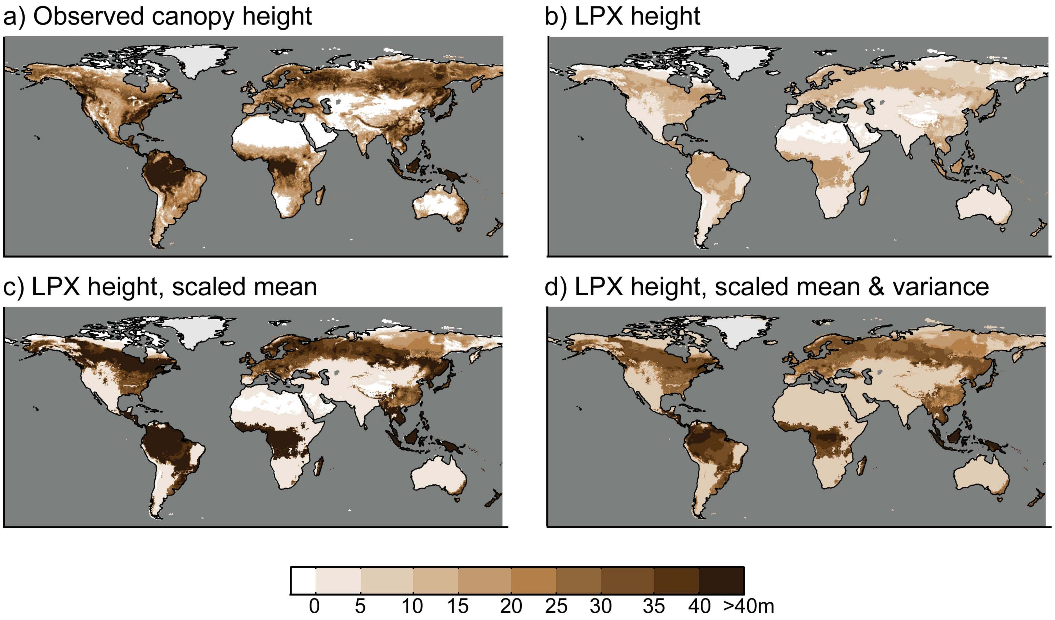
a) Observed canopy height
b) LPX height
c) LPX height, scaled mean
d) LPX height, scaled mean & variance
0
5
10
15
20
25
30
35
40
>40m

Fig. 6. Comparisons of observed and simulated height. **(a)** Observed canopy height (in 2005) from the Simard et al. (2011) dataset compared to **(b)** simulated height in the same year from LPX; **(c)** LPX-simulated height, multiplied by a factor of 2.67 so that the simulated global mean height is the same as the observations; **(d)** height from **(c)** with values reduced by a factor of 1.40 about the mean so that the simulations have the same global mean and variance as the observations.

the annual average burnt fraction changes from 1.58 (against GFED3) to 1.91 (against GFED2) for LPJ and from 0.85 (GFED3) to 0.92 (GFED2) for LPX (i.e. both models produce a better match to GFED3 than to GFED2), but the difference between the two models is preserved (i.e. LPX, with its more explicitly process-based fire model, is more realistic than LPJ).

3.2.3 Benchmarking the sensitivity to parameter tuning

Benchmarking can be used to evaluate how much tuning of individual parameters improves model performance and to ensure that the simulations capture specific processes correctly. We examine how well the current system serves in this respect by running sensitivity experiments using the SDBM. The SDBM underestimates the amplitude of CO2 seasonal cycle (Fig. 3). A better match to CO_2 observations can be achieved by tuning the light-use efficiency parameter (ε in Eq. 12). The best possible match to CO2 seasonal amplitude (0.18) is obtained when ε is equal to $1.73\,\mathrm{g\,C\,MJ^{-1}}$, but this increases both the mean and the variance of NPP compared to observations: the overall performance of the SDBM is therefore worse (Table 7). The seasonal amplitude of CO_2 depends on simulating the correct balance between NPP and R_h. Thus, given that the model produces a reasonable simulation of annual average NPP, improvement in CO_2 seasonality should come from changes in the simulation of R_h. Removing the requirement that NPP and R_h are in equilibrium, by setting total NPP to be 1.2 times R_h, improves the CO_2 seasonal amplitude score to 0.51. In the baseline simulation, the Q_{10} for R_h is 1.5 (Eq. 13). Changing this response by increasing Q_{10} to 2 degrades the simulation of the seasonal amplitude and phase of CO_2, while decreasing Q_{10} to 1.3 improves the simulation of the seasonal amplitude and phase of CO_2 (Table 7). Removing the seasonal response of R_h to moisture (i.e. removing α from Eq. 13) improves the score for seasonal amplitude (0.39) but does not change the score for the phase. However, this degrades its performance against annual average NPP from 0.58 to 0.82. We expect that R_h should be sensitive to soil moisture changes, but this analysis suggests that the treatment of this dependency in the SDBM is unrealistic.

4 Discussion and conclusion

Model benchmarking serves multiple functions, including (a) showing whether processes are represented correctly in a model, (b) discriminating between models and determining which performs better for a specific process, and (c) comparing between the model scores and those obtained by comparing mean and random resampling of observations, thus helping to identify processes that need improvement.

As found by Heimann et al. (1998), the SDBM produces a good simulation of the seasonal cycle of atmospheric CO_2 concentration. However, we show that the simulated amplitude of the seasonal cycle is too low (Table 5; Fig. 3). The SDBM's performance depends on getting the right balance

Table 7. Comparison metric scores for simulations with the Simple Diagnostic Biosphere Model (SDBM) against observations of the seasonal cycle of atmospheric CO_2 concentration and annual average NPP. Numbers in bold indicate the model with the best performance for that variable. Italic indicates model scores better than the SDBM simulation tuned using CO_2 seasonal observations. NME/NMSE denotes the normalised mean error/normalised mean squared error and MPD the mean phase difference. The details of each experiment are explained in the text.

Measure	SDBM standard run		SDBM tuned to CO_2 seasonal amplitude		SDBM NPP = 1.2 × R_h		SDBM $Q_{10} = 1.3$		SDBM $Q_{10} = 2$		SDBM constant α	
	NME	NMSE	NME	NMSE	NME	NMSE	NME	NMSE	NME	NMSE	NME	NMSE
CO_2 Amplitude	0.68	0.60	0.18	0.04	0.51	0.34	***0.15***	***0.02***	1.04	1.34	0.39	0.19
– mean removed	0.50	0.26	0.18	0.04	0.39	0.16	***0.11***	***0.01***	0.74	0.54	0.30	0.09
– mean and variance removed	**0.10**	0.02	**0.10**	0.02	**0.10**	0.02	**0.10**	***0.01***	0.18	0.07	0.12	0.02
MPD	0.21	N/A	0.21	N/A	***0.20***	N/A	***0.20***	N/A	0.26	N/A	0.21	N/A
NPP Annual Average	***0.58***	***0.36***	1.76	3.00	***0.58***	***0.36***	***0.58***	***0.36***	***0.58***	***0.36***	0.82	0.70
– mean removed	***0.53***	***0.32***	0.96	0.99	***0.53***	***0.32***	***0.53***	***0.32***	***0.53***	***0.32***	0.63	0.42
– mean and variance	***0.53***	***0.33***	**0.53**	**0.33**	***0.53***	***0.33***	***0.53***	***0.33***	***0.53***	***0.33***	0.63	0.44

of NPP and R_h. Improved simulation of CO_2 seasonal amplitude can be achieved through tuning the light-use efficiency using CO_2 station data, but this degrades the simulated NPP. The seasonal variation of R_h can be altered by changing the response of R_h to temperature (Q_{10}). Although many models (e.g. Potter et al., 1993; Cox et al., 2000) use Q_{10} values of 2, benchmarking shows that the value of 1.5 used in the SDBM provides a better match to seasonal CO_2 observations. However, reducing the Q_{10} to 1.3 improves the simulation still further. Mehecha et al. (2010), based on an analysis of FLUXNET data, have shown that Q_{10} values are 1.4 ± 0.1 independent of temperature or vegetation type. Thus, both the initial and "improved" Q_{10} values used here are consistent with observations, whereas values of 2 are not. Sensitivity analyses show that the SDBM can produce a seasonal cycle comparable to observations with respect to both amplitude and phase by removing the assumption that NPP and R_h are in equilibrium, and the dependence of R_h on seasonal changes in moisture availability. The idea that NPP and R_h are not in equilibrium is realistic; the idea that moisture availability has no impact on R_h is not. Thus, these analyses illustrate how benchmarking can be used to identify whether processes are represented correctly in a model, and pinpoint specific areas that should be targeted for investigation in further developments of the SDBM.

The benchmarking system can discriminate between models. LPJ and LPX are closely related models, differing primarily in the complexity of their treatment of fire and the feedbacks from fire disturbance to vegetation. The two DGVMs perform equally well against the benchmarks, e.g. for NPP (Fig. 9), inter-annual CO_2 concentrations (Fig. 10) and inter-annual and annual average runoff (Fig. 8). However, LPX performs better than LPJ with respect to all measures associated with fire (Fig. 7).

We were able to show several areas where both DGVMs perform poorly against the benchmarks, and use the comparisons to evaluate possible reasons. Deficiencies common to both models include a low bias in canopy height (Table 5; Fig. 6), poor simulation of the seasonal concentration of fAPAR and of the balance of tree and grass cover (Table 5), and poor simulation of the inter-annual variability in runoff (Fig. 8).

Both DGVMs score poorly against the canopy height benchmark (Fig. 6), averaging around 1/3 of observed heights (Table 5). However, they capture the spatial pattern of the differences in height reasonably well. A good match to canopy height was not expected, because LPJ and LPX do not simulate a size- or age-structured tree population but rather represent the properties of an "average individual". In contrast, the canopy height dataset represents the mean top height of forests within the grid cell. However, the models should, and do, capture broad geographic patterns of variation in height. The canopy height benchmark could provide a rigorous test for models that explicitly simulate cohorts of different ages of trees, such as the Ecosystem Demography (ED) model (Moorcroft et al., 2001). For models adopting a similar strategy to the LPJ/LPX family, the key test is whether the spatial patterns are correctly simulated. In either case, the use of remotely sensed canopy height data represents a valuable addition to the benchmarking toolkit.

Poor performance in the simulation of seasonal concentration of fAPAR (Table 5) demonstrates that both DGVMs predict the length of the growing season inaccurately: the growing season is too long at low latitudes and too short at mid-latitudes. This poor performance indicates that the phenology of both evergreen and deciduous vegetation requires improvement. Both models overestimate the amount of tree cover and underestimate grass cover (Table 5). The oversharp boundaries between forests and grasslands suggest that the models have problems in simulating the coexistence of these life forms. This probably also affects, and is exacerbated by, the simulation of fire in the models (Fig. 7).

The DGVMs simulate annual average runoff reasonably well, but inter-annual variability in runoff is poorly

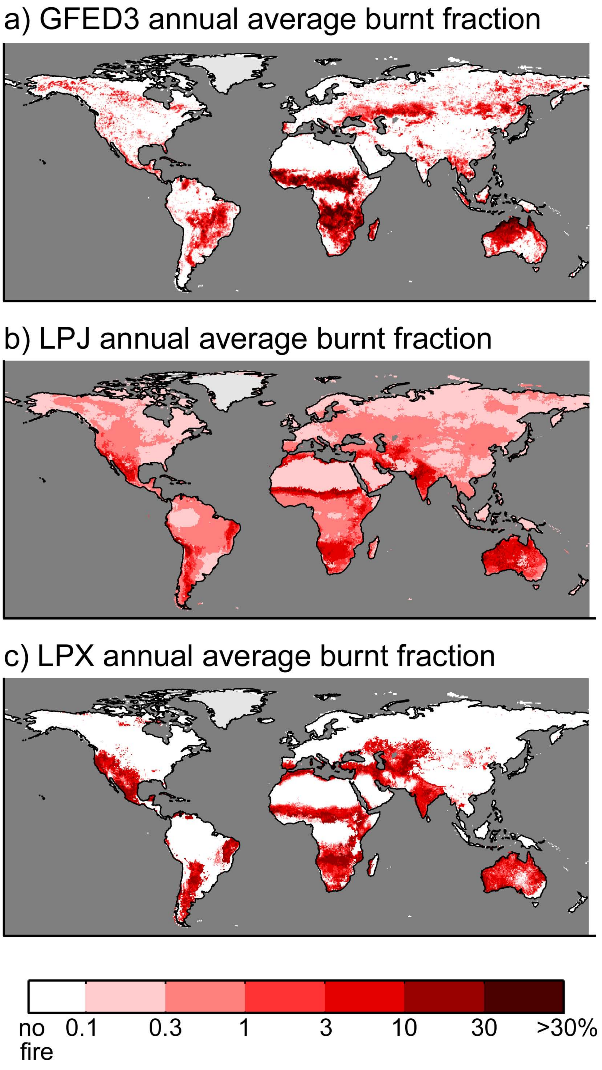
a) GFED3 annual average burnt fraction
b) LPJ annual average burnt fraction
c) LPX annual average burnt fraction
no fire
0.1
0.3
1
3
10
30
>30%

Fig. 7. Annual average burnt fraction between 1997–2005 from **(a)** GFED3 observations (Giglio et al., 2010) and as simulated by **(b)** LPJ and **(c)** LPX.

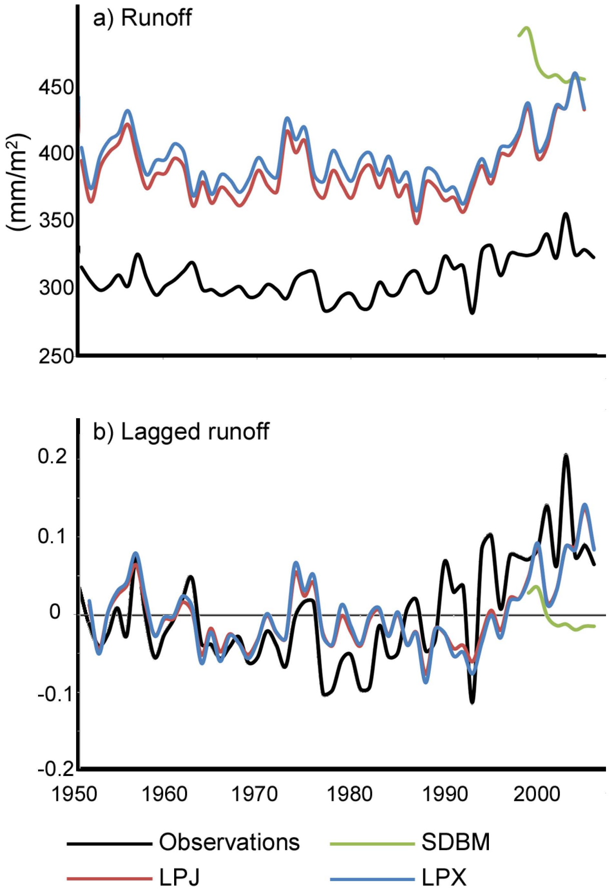

Fig. 8. Observed inter-annual runoff for 1950–2005 averaged over basins from the Dai et al. (2009) dataset (black line) compared to average simulated runoff over the same basins from LPJ (red line) and LPX (blue line). **(a)** shows inter-annual runoff, and **(b)** shows inter-annual variability in runoff where the simulated values are lagged by a year.

simulated. In large basins, water can take many months to reach the river mouth (Ducharne et al., 2003) and this delay has a major impact on the timing of peaks in river discharge. Neither LPX nor the version of LPJ evaluated here includes river routing; runoff is simulated as the instantaneous difference in the water balance. Thus, it is unsurprising that neither model produces a good match to observations of inter-annual variability. Murray et al. (2011) have pointed out that inclusion of a river routing scheme should improve the simulation of runoff in LPX, and this is supported by the fact that introducing a one-year lag improved model performance against the runoff benchmark (Fig. 8). There is already a version of LPJ (LPJmL v3.2: Rost et al., 2008) that incorporates a water storage and transport model (and also includes human extraction), and produces a more realistic simulation of inter-annual variability in runoff than the version examined here.

In this paper, we have emphasised the use of global metrics for benchmarking, although both the NME and MM metrics provide a measure of the impact of the correct simulation of geographical patterning on global performance. However, the metrics could also be used to evaluate model performance at smaller geographic scales (e.g. for specific latitudinal bands, or individual continents or biomes). For example, comparison of the mean annual burnt fraction scores for specific latitudinal bands shows that the two DGVMs simulate fire in tropical regions better than in extratropical regions or overall, with NME scores for the tropics of 1.27 (LPJ) and 0.82 (LPX) compared to the global scores of 1.58 (LPJ) and 0.85 (LPX).

Some variables, such as runoff and burnt fraction, display considerable inter-annual variability linked to climate (e.g. changes in ENSO: van der Werf et al., 2004;

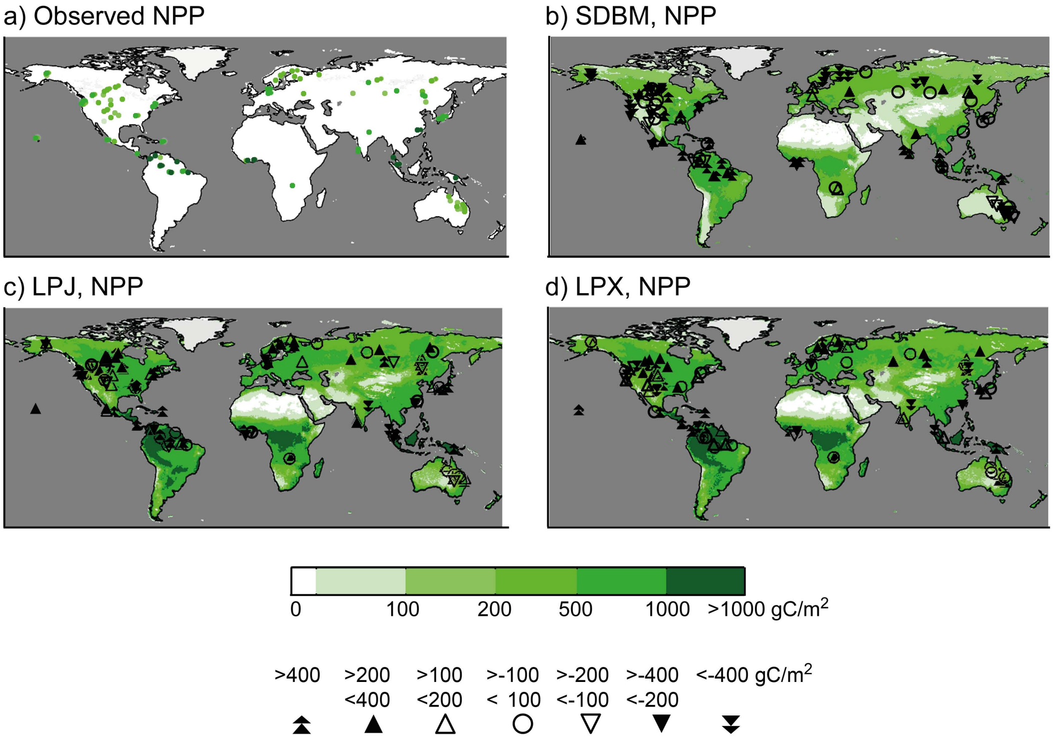

Fig. 9. Comparison of observed and simulated annual average net primary production (NPP). Observed values are from the Luyssaert et al. (2007) and Ecosystem/Model Data Intercomparison datasets (Olson et al., 2001), and the simulated values are from **(b)** Simple Diagnostic Biosphere Model (SDBM), **(c)** LPJ and **(d)** LPX. The symbols in **(b)**, **(c)** and **(d)** indicate the magnitude and direction of disagreement between simulation and observed values, where the upward and downward facing triangles represent over- and undersimulation respectively. Double triangles indicate a difference in NPP of $> 400\,\mathrm{g\,C\,m^{-2}}$, single filled triangles a difference between 200 and $400\,\mathrm{g\,C\,m^{-2}}$, single empty triangles a difference 100 and $200\,\mathrm{g\,C\,m^{-2}}$, and empty circles a difference of $< 100\,\mathrm{g\,C\,m^{-2}}$

post-volcanic cooling events: Riaño et al., 2007), and valuable information is obtained by considering this variability. The vegetation cover and canopy height datasets used for benchmarking here are single-year "snapshots": this is entirely appropriate for variables that change only slowly. Nevertheless, given that vegetation is already responding to changes in climate (Parmesan, 2006; Hickling et al., 2006; Fischlin et al., 2007), additional "snapshots" of these variables would be useful adjuncts to a benchmarking system allowing evaluation of models' ability to reproduce decadal-scale variability in vegetation properties.

In general, remote sensing data are most likely to provide the global coverage necessary for a benchmark dataset. Nevertheless, we have found considerable value in using site-based datasets for river discharge, CO_2, GPP and NPP. River discharge data are spatially integrated over basins that together cover much of the global land surface, while CO_2 station measurements intrinsically integrate land–atmosphere CO_2 fluxes over moderately large areas through atmospheric transport. The coverage of the site-based GPP and NPP datasets is more limited and currently does not represent the full range of biomes. We have shown that model performance against the Beer et al. (2010) gridded GPP dataset is better than performance against the site-specific estimates of GPP in the Luyssaert et al. (2007) dataset – a function of the much higher number of flux-tower measurements included in the newer dataset and the smoothing of individual measurements inherent in the interpolation of these measurements to produce a gridded dataset. We do not use the Beer et al. (2010) dataset as a standard benchmark, because it was derived, in part, using the same climate variables that are used for the simulation of GPP in the vegetation models. However, the apparent improvement in model performance against the Beer et al. (2010) dataset at the Luyssaert et al (2007) sites indicates the importance of making quality-controlled summaries of the primary flux-tower data available to the modelling community for benchmarking purposes.

GPP and NPP have also been derived from remotely sensed products (e.g. Running et al., 2004; Turner et al., 2006). This is not an optimal approach because the results are heavily influenced by the model used to translate the spectral vegetation indices, and the reliability of the product varies with spatial scale and for a given ecosystem type (Lu and Ji, 2006).

A more general issue with the development of benchmarking systems is the fact that target datasets are constantly

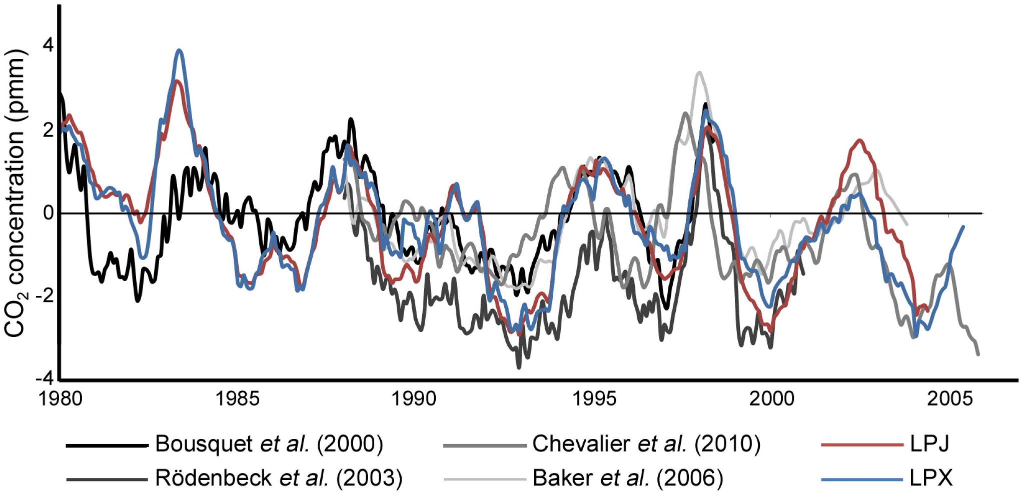

Fig. 10. Twelve-month running mean of inter-annual variability in global atmospheric CO_2 concentration between 1998–2005 from Bousquet et al. (2000), Rödenbeck et al. (2003), Baker et al. (2006) and Chevalier et al. (2010) compared to simulated inter-annual variability from LPJ and LPX.

being extended in time and upgraded in quality. This is potentially problematic if the benchmark system is to be used to evaluate improvements in model performance through time, since this requires the use of a fixed target against which to compare successive model versions, but this target may have been superseded in the interim. In the current system, for example, we use the Dai et al. (2009) dataset for runoff, which supersedes an earlier product (Dai and Trenberth, 2002) and improves upon this earlier product by including more and longer records. The use of an updated version of the same target dataset may change the numeric scores obtained for a given simulation, but our comparison of the GFED2 and GFED3 datasets suggests this is unlikely to change the interpretation of how well a model performs. Any benchmarking system will need to evolve as new data products become available. In practical terms, this may mean that data–model comparisons will have to be performed against both the old and new versions of the products in order to establish how different these products are from one another and to establish a new baseline comparison value for any given model. As with the datasets used in this study, any new datasets should be freely available to the scientific community, to allow different modelling groups to undertake comparable benchmarking exercises.

A major limitation of the benchmarking approach presented here is that it does not take into account observational uncertainties, because very few datasets provide a quantitative estimate of such uncertainties. We have shown that observational uncertainty is larger than differences in model performance with respect to site-based annual average NPP measurements, and these observational uncertainties are also greater than model biases in NPP. However, differences in the performance of LPJ and LPX with respect to annual average burnt fraction are considerably larger than observational uncertainties. Approaches such as the use of multiple datasets (e.g. our use of multiple CO_2 inversions) may be one way of assessing uncertainty where there are no grounds for selecting a particular dataset as being more accurate or realistic. However, the only comprehensive solution to the problem is for measurement uncertainties to be routinely assessed for each site/grid cell and included with all datasets.

We have not attempted to provide an overall assessment of model performance by combining the metric scores obtained from each of the benchmarks into a composite skill score, although this has been done in some previous analyses (e.g. Randerson et al., 2009), because this requires subjective decisions about how to weight the importance of each metric. Composite skill scores have been used in data-assimilation studies to obtain better estimates of model parameters (e.g. Trudinger et al., 2007). The choice of weights used in these multi-variable composite metrics alters the outcome of parameter optimization (Trudinger et al., 2007; Weng and Luo, 2011; Xu et al., 2006). Decisions about how to weight individual vegetation-model benchmarks may heavily influence model performance scores (Luo et al., 2012).

The community-wide adoption of a standard system of benchmarking, as first proposed by C-LAMP (Randerson et al., 2009) and by ILAMB (Luo et al., 2012), would help users to evaluate the uncertainties associated with specific vegetation-model simulations and help to determine which projections of the response of vegetation to future climate changes are likely to be more reliable. As such, it will help to enhance confidence in these tools. At the same time, as we have shown here, systematic benchmarking provides a good way to identify ways of improving the current models and should lead to better models in the future.

Acknowledgements. S. P. Harrison and I. C. Prentice were responsible for the design of the benchmarking system; I. C. Prentice, K. O. Willis and D. I. Kelley designed the metrics; D. I. Kelley ran the LPJ and LPX simulations, coded metrics and made the statistical analyses; H. Wang and D. I. Kelley collated and regridded datasets and coded and ran the SDBM. J. B. Fisher and M. Simard provided remote-sensed datasets. D. I. Kelley, S. P. Harrison and I. C. Prentice wrote the first draft of the paper; all authors contributed to the final version of the manuscript. D. I. Kelley is supported by a Macquarie University International Research Scholarship (iMQRES). We thank Gab Abramowitz and the ILAMB project (http://www.ilamb.org) for discussions of benchmarking strategy and Stephen Sitch for supplying the atmospheric transport matrices. Metrics were scripted using R (R Development Core Team, 2012). The benchmarking system, datasets and scripts for data–model comparison metrics are available at http://bio.mq.edu.au/bcd/benchmarks/.

Edited by: E. Falge

References

Arora, V. K. and Boer, G. J.: Fire as an interactive component of dynamic vegetation models, J. Geophys. Res., 110, G02008, doi:10.1029/2005JG000042, 2005.

Baker, D. F., Doney, S. C., and Schimel, D. S.: Variational data assimilation for atmospheric CO_2, Tellus B, 58, 359–365, 2006.

Barnston, G. A.: Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score, Boston, MA, USA, American Meteorological Society, 1992.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M. A., Baldocchi, D., Bonan, G. B., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K. W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, F. I., and Papale, D.: Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate, Science, 329, 834–838, 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, 2009.

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, doi:10.5194/gmd-4-255-2011, 2011.

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:10.1029/2010JG001593, 2011.

Bousquet, P., Peylin, P., Ciais, P., Le Quéré, C., Friedlingstein, P., and Tans, P. P.: Regional changes in carbon dioxide fluxes of land and oceans since 1980, Science, 290, 1342–1346, 2000.

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, Glob. Biogeochem. Cy., 24, GB2016, doi:10.1029/2009GB003556, 2010.

Carmona-Moreno, C., Belward, A., Malingreau, J.-P., Hartley, A., Garcia-Alegre, M., Antonovskiy, M., Buchshtaber, V., and Pivovarov, V.: Characterizing interannual variations in global fire calendar using data from Earth observing satellites, Glob. Change Biol., 11, 1537–1555, 2005.

Cha, S.: Comprehensive survey on distance / similarity measures between probability density functions, Int. J. Math. Models Methods Appl. Sci., 1, 301–307, 2007.

Chevallier, F., Ciais, P., Conway, T. J., Aalto, T., Anderson, B. E., Bousquet, P., Brunke, E. G., Ciattaglia, L., Esaki, Y., Fröhlich, M., Gomez, A., Gomez-Pelaez, A. J., Haszpra, L., Krummel, P. B., Langenfelds, R. L., Leuenberger, M., Machida, T., Maignan, F., Matsueda, H., Morguí, J. A., Mukai, H., Nakazawa, T., Peylin, P., Ramonet, M., Rivier, L., Sawa, Y., Schmidt, M., Steele, L. P., Vay, S. A., Vermeulen, A. T., Wofsy, S. and Worthy, D.: CO_2 surface fluxes at grid point scale estimated from a global 21 year reanalysis of atmospheric measurements, J. Geophys. Res., 115, D21307, doi:10.1029/2010JD013887, 2010.

Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A., and Totterdell, I. J: Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model, Nature, 408, 184–187, 2000.

Cramer, W., Kicklighter, D. W., Bondeau, A., Moore, B., Churkina, G., Nemry, B., Ruimy, A., and Schloss, A. L.: Comparing global models of terrestrial net primary productivity (NPP): overview and key results, Global Change Biol., 5, 1–15, 1999.

Dai, A. and Trenberth, K. E.: Estimates of freshwater discharge from continents: latitudinal and seasonal variations, J. Hydrometeorol., 3, 660–687, 2002.

Dai, A., Qian, T., Trenberth, K. E., and Milliman, J. D.: Changes in continental freshwater discharge from 1948 to 2004, J. Climate, 22, 2773–2792, 2009.

DeFries, R. and Hansen, M. C.: ISLSCP II Continuous Fields of Vegetation Cover, 1992–1993, in: ISLSCP Initiative II Collection, Data set, edited by: Hall, F. G., Collatz, G., Meeson, B., Los, S., Brown De Colstoun, E., and Landis, D., Oak Ridge, Tennessee, USA, available at:http://daac.ornl.gov/ from Oak Ridge National Laboratory Distributed Active Archive Center, last access: 13 January 2011, 2009.

DeFries, R. S., Townshend, J. R. G., and Hansen, M. C.: Continuous fields of vegetation characteristics at the global scale at 1-km resolution, J. Geophys. Res., 104, 16911–16923, 1999.

DeFries, R. S., Hansen, M. C., Townshend, J. R. G., Janetos, A. C., and Loveland, T. R.: A new global 1-km dataset of percentage tree cover derived from remote sensing, Glob. Change Biol., 6, 247–254, 2000.

Denman, K. L., Brasseur, G., Chidthaisong, A., Ciais, P., Cox, P. M., Dickinson, R. E., Hauglustaine, D., Heinze, C., Holland, E., Jacob, D., Lohmann, U., Ramachandran, S., da Silva Dias, P. L., Wofsy, S. C., and Zhang, X.: Couplings between changes in the climate system and biogeochemistry, in: Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., Cambridge and New York, Cambridge University Press, 499–587, 2007.

Ducharne, A., Golaz, C., Leblois, E., Laval, K., Polcher, J., Ledoux, E., and De Marsily, G.: Development of a high resolution runoff routing model, calibration and application to assess runoff from the LMD GCM, J. Hydrol., 280, 207–228, 2003.

Efron, B.: Bootstrap methods: another look at the Jackknife, Ann. Stat., 7, 1–26, 1979.

Efron, B. and Tibshirani, R. J.: An Introduction to the Bootstrap, New York, Chapman & Hall, 1993.

FAO: The Digitized Soil Map of the World (Release 1.0), edited by: Food and Agriculture Organization of the United Nations, Rome, Italy, World Soil Resources Report 67/1, 1991.

Fischlin, A., Midgley, G. F., Price, J., Leemans, R., Gopal, B., Turley, C., Rounsevell, M., Dube, P., Tarazona, J., Velichko, A., Atlhopheng, J., Beniston, M., Bond, W. J., Brander, K., Bugmann, H., Callaghan, T. V., de Chazal, J., Dikinya, O., Guisan, A., Gyalistras, D., Hughes, L., Kgope, B. S., Körner, C., Lucht, W., Lunn, N. J., Neilson, R. P., Pêcheux, M., Thuiller, W., and Warren, R.: Ecosystems, their properties, goods, and services, in: Climate Change 2007: impacts, adaptation and vulnerability, Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Parry, M. L., Canziani, O. F., Palutikof, J. P., Van Der Linden, P. J., and Hanson, C. E., Cambridge, United Kingdom, Cambridge University Press, 211–272, 2007.

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.

Fisher, J. B., Whittaker, R. J., and Malhi, Y.: ET come home: potential evapotranspiration in geographical ecology, Glob. Ecol. Biogeogr., 20, 1–18, 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.

Gallego-Sala, A. V., Clark, J. M., House, J. I., Orr, H. G., Prentice, I. C., Smith, P., Farewell, T., and Chapman, S. J.: Bioclimatic envelope model of climate change impacts on blanket peatland distribution in Great Britain, Clim. Res., 45, 151–162, 2010.

Gavin, D. G., Oswald, W. W., Wahl, E. R., and William, J. W.: A statistical approach to evaluating distance metrics and analog assignments for pollen records, Quat. Res., 60, 356–367, 2003.

Gerten, D., Schaphoff, S., Haberlandt, U., Lucht, W., and Sitch, S.: Terrestrial vegetation and water balance – hydrological evaluation of a dynamic global vegetation model, J. Hydrol., 286, 249–270, 2004.

Giglio, L., Csiszar, I., and Justice, C. O.: Global distribution and seasonality of active fires as observed with the Terra and Aqua Moderate Resolution Imaging Spectroradiometer (MODIS) sensors, J. Geophys. Res., 111, G02016, doi:10.1029/2005JG000142, 2006.

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

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res., 113, D06104, doi:10.1029/2007JD008972, 2008.

Gobron, N., Pinty, B., Taberner, M., Mélin, F., Verstraete, M. and Widlowski, J.: Monitoring the photosyntetic activity of vegetation from remote sensing data, Adv. Space Res., 38, 2196-2202, 2006.

Hall, A. and Qu, X.: Using the current seasonal cycle to constrain snow albedo feedback in future climate change, Geophys. Res. Lett., 33, L03502, doi:10.1029/2005GL025127, 2006.

Hall, F. G., Brown De Colstoun, E., Collatz, G. J., Landis, D., Dirmeyer, P., Betts, A., Huffman, G. J., Bounoua, L., and Meeson, B.: ISLSCP Initiative II global data sets: Surface boundary conditions and atmospheric forcings for land-atmosphere studies, J. Geophys. Res., 111, D22S01, doi:10.1029/2006JD007366, 2006.

Heimann, M.: The global atmospheric tracer model TM2: model description and user manual, in: The Global Atmospheric Tracer Model TM2, edited by: Deutsches Klimarechenzentrum, Max-Planck-Institut fur Meteorologie, http://mms.dkrz.de/pdf/klimadaten/servicesupport/documents/reports/ReportNo.10.pdf (last access: 7 September 2011), Hamburg, Germany, 1995.

Heimann, M., Esser, G., Haxeltine, A., Kaduk, J., Kicklighter,D. W., Knorr, W., Kohlmaier, G. H., McGuire, A. D., Melillo, J., Moore III, B., Otto, R. D., Prentice, I. C., Sauf, W., Schloss, A., Sitch, S., Wittenberg, U., and Würth, 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.

Hickling, R., Roy, D. B., Hill, J. K., Fox, R., and Thomas, C. D.: The distributions of a wide range of taxonomic groups are expanding polewards, Glob. Change Biol., 12, 450–455, 2006.

Jackson, C. S., Sen, M. K., Huerta, G., Deng, Y., and Bowman, K. P.: Error reduction and convergence in climate prediction, J. Climate, 21, 6698–6709, 2008.

Jones, P. and Harris, I.: CRU Time Series (TS) high resolution gridded datasets, edited by: Climate Research Unit, available at: http://badc.nerc.ac.uk/view/badc.nerc.ac.uk ATOM dataent 1256223773328276, BAD C, last access: 26 September 2012.

Jung, M., Reichstein, M., and Bondeau, A.: Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model, Biogeosciences, 6, 2001–2013, doi:10.5194/bg-6-2001-2009, 2009.

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., Bonan, G. B., Cescatti, A., Chen, J., de Jeu, R., Dolman, A. J., Eugster, W., Gerten, D., Gianelle, D., Gobron, N., Heinke, J., Kimball, J. S., Law, B. E., Montagnani, L., Mu, Q., Mueller, B., Oleson, K. W., Papale, D., Richardson, A. D., Roupsard, O., Running, S. W., 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.

Kaminski, T., Giering, R., and Heimann, M.: Sensitivity of the seasonal cycle of CO_2 at remote monitoring stations with respect to seasonal surface exchange fluxes determined with the adjoint of an atmospheric transport model, Phys. Chem. Earth, 21, 457–462, 1996.

Keeling, R.: Atmospheric science – Recording Earth's vital signs, Science, 319, 1771–1772, 2008.

Knorr, W. and Heimann, M.: Impact of drought stress and other factors on seasonal land biosphere CO_2 exchange studied through an atmospheric tracer transport model, Tellus B, 47, 471–489, 1995.

Le Quéré, C., Aumont, O., Bopp, L., Bousquet, P., Ciais, P., Francey, R., Heimann, M., Keeling, C. D., Keeling, R. F., Kheshgi, H., Peylin, P., Piper, S. C., Prentice, I. C., and Rayner, P. J.: Two decades of ocean CO_2 sink and variability, Tellus B, 55, 649–656, 2003.

Lenderink, G.: Exploring metrics of extreme daily precipitation in a large ensemble of regional climate model simulations, Clim. Res., 44, 151–166, 2010.

Lu, J. and Ji, J.: A simulation and mechanism analysis of long-term variations at land surface over arid/semi-arid area in north China, J. Geophys. Res., 111, doi:10.1029/2005JD006252, 2006.

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

Luyssaert, S., Inglima, I., Jung, M., Richardson, A. D., Reichstein, M., Papale, D., Piao, S. L., Schulze, E. -D., Wingate, L., Matteucci, G., Aragao, L., Aubinet, M., Beer, C., Bernhofer, C., Black, K. G., Bonal, D., Bonnefond, J. -M., Chambers, J., Ciais, P., Cook, B., Davis, K. J., Dolman, A. J., Gielen, B., Goulden, M., Grace, J., Granier, A., Grelle, A., Griffis, T., Grünwald, T., Guidolotti, G., Hanson, P. J., Harding, R., Hollinger, D. Y., Hutyra, L. R., Kolari, P., Kruijt, B., Kutsch, W., Lagergren, F., Laurila, T., Law, B. E., Le Maire, G., Lindroth, A., Loustau, D., Malhi, Y., Mateus, J., Migliavacca, M., Misson, L., Montagnani, L., Moncrieff, J., Moors, E., Munger, J. W., Nikinmaa, E., Ollinger, S. V., Pita, G., Rebmann, C., Roupsard, O., Saigusa, N., Sanz, M. J., Seufert, G., Sierra, C., Smith, M. -L., Tang, J., Valentini, R., Vesala, T. and Janssens, I. A.: CO_2 balance of boreal, temperate, and tropical forests derived from a global database, Glob. Change Biol., 13, 2509–2537, 2007.

Mahecha, M. D., Reichstein, M., Carvalhais, N., Lasslop, G., Lange, H., Seneviratne, S. I., Vargas, R., Ammann, C., Arain, A. M., Cescatti, A., Janssens, I. A., Migliavacca, M., Montagnani, L., and Richardson, A. D.: Global Convergence in the Temperature Sensitivity of Respiration at Ecosystem Level, Science, 329, 838–840, 2010.

Moise, A. F. and Delage, F. P.: New climate model metrics based on object-orientated pattern matching of rainfall, J. Geophys. Res., 116, D12108, doi:10.1029/2010JD015318, 2011.

Monteith, J. L.: Solar radiation and productivity in tropical ecosystems, J. Appl. Ecol., 9, 747–766, 1972.

Moorcroft, P. R., Hurtt, G. C., and Pacala, S. W.: A method for scaling vegetation dynamics: the Ecosystem Demography model (ED), Ecol. Monogr., 71, 557–586, 2001.

Mu, Q., Zhao, M., and Running, S. W.: Improvements to a MODIS global terrestrial evapotranspiration algorithm, Remote Sens. Environ., 115, 1781–1800, 2011.

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collin, M., and Stainforth, D. A.: Quantification of modelling uncertainties in a large ensemble of climate change simulations, Nature, 430, 768–772, 2004.

Murray, S. J., Foster, P. N., and Prentice, I. C.: Evaluation of global continental hydrology as simulated by the Land-surface Processes and eXchanges Dynamic Global Vegetation Model, Hydrol. Earth Syst. Sci., 15, 91–105, doi:10.5194/hess-15-91-2011, 2011.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, 1970.

Nevison, C. D., Mahowald, N. M., Doney, S. C., Lima, I. D., van der Werf, G. R., Randerson, J. T., Baker, D. F., Kasibhatla, P., and McKinley, G. A.: Contribution of ocean, fossil fuel, land biosphere, and biomass burning carbon fluxes to seasonal and interannual variability in atmospheric CO_2, J. Geophys. Res., 113, G01010, doi:10.1029/2007JG000408, 2008.

Olson, R. J., Scurlock, J. M. O., Prince, S. D., Zheng, D. L., and Johnson, K. R.: NPP Multi-Biome: NPP and Driver Data for Ecosystem Model-Data Intercomparison, Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA, 2001.

Parmesan, C.: Ecological and Evolutionary Responses to Recent Climate Change, Annu. Rev. Ecol. Evol. Syst., 37, 637–669, 2006.

Piani, C., Frame, D. J., Stainforth, D. A., and Allen, M.,R.: Constraints on climate change from a multi-thousand member ensemble of simulations, Geophys. Res. Lett., 32, L23825, doi:10.1029/2005GL024452, 2005.

Poorter, H., Remkes, C., and Lambers, H.: Carbon and nitrogen economy of 24 wild species differing in relative growth rate, Plant Physiol., 94, 621–627, 1990.

Potter, C. S., Randerson, J. T., Field, C. B., Matson, P. A., Vitousek, P. M., Mooney, H. A., and Klooster, S. A.: Terrestrial ecosystem production: A process model based on global satellite and surface data, Global Biogeochem. Cy., 7, 9144–9224, 1993.

Prentice, I. C., Sykes, M. T., and Cramer, W.: A simulation model for the transient effects of climate change on forest landscapes, Ecol. Model., 65, 51–70, 1993.

Prentice, I. C., Bondeau, A., Cramer, W., Harrison, S. P., Hickler, T., Lucht, W., Sitch, S., Smith, B., and Sykes, M. T.: Dynamic Global Vegetation Modelling: quantifying terrestrial ecosystem responses to large-scale environmental change Terrestrial Ecosystems in a Changing World, Springer Berlin Heidelberg, 2007.

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.: A model of regional primary production for use with coarse resolution satellite data, Int. J. Remote Sens., 12, 1313–1330, 1991.

R Development Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, available at: http://www.R-project.org/, last access: 11 July 2012.

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., Stouffer, R. J., Sumi A., and Taylor K. E.: Cilmate models and their evaluation, in: Climate change 2007: the physical science basis, Contribution of working group 1 to the fourth assessment report of the intergovernmental panel on climate change edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor M., and Miller H. L., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 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., Stockli, 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., Briggs, P. R., Haverd, V., King, E. A., Paget, M. and Trudinger, C. M.: Australian Water Availability Project (AWAP): CSIRO Marine and Atmospheric Research Component: Final Report for Phase 3, in: CAWCR Technical Report, Melbourne, Australia, The Centre for Australian Weather and Climate Research, 2009.

Riaño, D., Moreno Ruiz, J. A., Barón Martínez, J., and Ustin, S. L.: Burned area forecasting using past burned area records and Southern Oscillation Index for tropical Africa (1981–1999), Remote Sens. Environ., 107, 571–581, 2007.

Reichler, T. and Kim, J.: How well do coupled models simulate today's climate?, B. Am. Meteorol. Soc., 89, 303–311, doi:10.1175/BAMS-89-3-303, 2008.

Rödenbeck, C., Houweling, S., Gloor, M., and Heimann, M.: CO_2 flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport, Atmos. Chem. Phys., 3, 1919–1964, doi:10.5194/acp-3-1919-2003, 2003.

Rost, S., Gerten, D., Bondeau, A., Lucht, W., Rohwer, J., and Schaphoff, S.: Agricultural green and blue water consumption and its influence on the global water system, Water Resour. Res., 44, 1–17, 2008.

Running, S. W., Nemani, R. R., Heinsch, F. A., Zhao, M., Reeves, M., and Hashimoto, H.: A continuous satellite-derived measure of global terrestrial primary production, Bioscience, 54, 547–560, 2004.

Scheiter, S. and Higgins, S. I.: Impacts of climate change on the vegetation of Africa: an adaptive dynamic vegetation modelling approach, Globm Change Biol., 15, 2224–2246, 2009.

Scholze, M., Knorr, W., Arnell, N. W., and Prentice, I. C.: A climate-change risk analysis for world ecosystems, P. Natl. Acad. Sci., 103, 13116–13120, 2006.

Shukla, J., DelSole, T., Fennessy, M., Kinter, J., and Paolino, D.: Climate model fidelity and projections of climate change. Geophys. Res. Lett., 33, L07702, doi:10.1029/2005GL025579, 2006.

Simard, M., Pinto, N., Fisher, J. B., and Baccini, A.: Mapping forest canopy height globally with spaceborne lidar, J. Geophys. Res., 116, G04021, doi:10.1029/2011JG001708, 2011.

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.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res, 106, 7183–7192, doi:10.1029/2000JD900719, 2001.

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.

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., 112, G02027, doi:10.1029/2006JG000367, 2007.

Turner, D. P., Ritts, W. D., Maosheng, Z., Kurc, S. A., Dunn, A. L., Wofsy, S. C., Small, E. E., and Running, S. W.: Assessing inter-annual variation in MODIS-based estimates of gross primary production, Geosci. Remote Sens., IEEE T., 44, 1899–1907, 2006.

van der Werf, G. R., Randerson, J. T., Collatz, G. J., Giglio, L., Kasibhatla, P. S., Arellano Jr., 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.

van der Werf, G. R., Randerson, J. T., Giglio, L., Collatz, G. J., Mu, M., Kasibhatla, P. S., Morton, D. C., DeFries, R. S., Jin, Y., and van Leeuwen, T. T.: Global fire emissions and the contribution of deforestation, savanna, forest, agricultural, and peat fires (1997–2009), Atmos. Chem. Phys., 10, 11707–11735, doi:10.5194/acp-10-11707-2010, 2010.

van Oijen, M., Cameron, D. R., Butterbach-Bahl, K., Farahbakhshazad, N., Jansson, P. E., Kiese, R., Rahn, K. H., Werner, C., and Yeluripati J. B.: A Bayesian framework for model calibration, comparison and analysis: Application to four models for the biogeochemistry of a Norway spruce forest, Agr. Forest Meteorol., 151, 1609–1621, 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.

Woodward, F. I. and Lomas, M. R.: Vegetation dynamics – simulating responses to climatic change, Biol. Rev., 79, 643–670, 2004.

Xu, T., White, L., Hui, D., and Luo, Y.: Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction, Global Biogeochem. Cy., 20, GB2007, doi:10.1029/2005GB002468, 2006.

Yokoi, S., Takayabu, Y. N., Nishii, K., Nakamura, H., Endo, H., Ichikawa, H., Inoue, T., Kimoto, M., Kosaka, Y., Miyasaka, T., Oshima, K., Sato, N., Tsushima, Y., and Watanabe, M.: Application of cluster analysis to climate model performance metrics. J. Appl. Meteor. Climatol., 50, 1666–1675, 2011.

Zeng, X., Zeng, X., and Barlage, M.: Growing temperate shrubs over arid and semiarid regions in the Community Land Model; Dynamic Global Vegetation Model, Global Biogeochem. Cy., 22, GB3003, doi:10.1029/2007GB003014, 2008.