Geosci. Model Dev., 4, 255–269, 2011
www.geosci-model-dev.net/4/255/2011/
doi:10.5194/gmd-4-255-2011

Geoscientific Model Development

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

E. Blyth[1], **D. B. Clark**[1], **R. Ellis**[1], **C. Huntingford**[1], **S. Los**[2], **M. Pryor**[3], **M. Best**[3], **and S. Sitch**[4]

[1]Centre for Ecology and Hydrology, Wallingford OX10 8BB, UK
[2]Department of Geography, Swansea University, Singleton Park, Swansea, SA2 8PP, UK
[3]Hadley Centre for climate prediction and research, Met Office, Joint Centre for Hydro-Meteorological Research, Wallingford OX10 8BB, UK
[4]School of Earth and Environment, University of Leeds, Leeds, LS2 9JT, UK

Received: 15 September 2010 – Published in Geosci. Model Dev. Discuss.: 26 October 2010
Revised: 11 March 2011 – Accepted: 22 March 2011 – Published: 6 April 2011

Abstract. Evaluating the models we use in prediction is important as it allows us to identify uncertainties in prediction as well as guiding the priorities for model development. This paper describes a set of benchmark tests that is designed to quantify the performance of the land surface model that is used in the UK Hadley Centre General Circulation Model (JULES: Joint UK Land Environment Simulator). The tests are designed to assess the ability of the model to reproduce the observed fluxes of water and carbon at the global and regional spatial scale, and on a seasonal basis. Five datasets are used to test the model: water and carbon dioxide fluxes from ten FLUXNET sites covering the major global biomes, atmospheric carbon dioxide concentrations at four representative stations from the global network, river flow from seven catchments, the seasonal mean NDVI over the seven catchments and the potential land cover of the globe (after the estimated anthropogenic changes have been removed). The model is run in various configurations and results are compared with the data.

A few examples are chosen to demonstrate the importance of using combined use of observations of carbon and water fluxes in essential in order to understand the causes of model errors. The benchmarking approach is suitable for application to other global models.

Correspondence to: E. Blyth
(emb@ceh.ac.uk)

1 Introduction

Changes in atmospheric carbon dioxide and water vapour affect the global radiation budget and are important drivers of climate change. One key control is the land surface which absorbs, stores and releases carbon and water. The terrestrial cycling of carbon and water varies across the climate regions of the world and is also temporally variable: from diurnal, seasonal, inter-annual, decadal timescales and even longer.

The water vapour flux from the land to the atmosphere affects the weather patterns of the world. Spatial difference in the water held in the soils and subsequent patterns of seasonal evaporation and plant growth likely affect rainfall (Los et al., 2006). Several authors have demonstrated that improved representation of land-surface soil moisture and vegetation improves the skill of rainfall prediction (Koster et al., 2004; Beljaars et al., 1996; Van den Hurk et al., 2003). Increases in atmospheric greenhouse gas concentrations are expected to alter rainfall patterns significantly (IPCC, 2007, Fig. TS.30) and thus understanding the role of the land surface within such change is particularly important. The land surface is also expected to play a major and changing role in the global carbon cycle and changes in the land cover due to land used for food and fuel production can have impacts on the weather and climate (Cox et al., 2000).

The carbon, water and energy cycles are closely linked and climate and weather prediction models therefore need to include a robust and accurate representation of the land surface in the UK Hadley Centre Climate Prediction model (the Unified Model) to portray the regional, seasonal variability of these carbon and water stores and fluxes. The land surface model in the Unified Model is JULES (Joint UK

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

Land Environment Simulator, Blyth et al., 2006). JULES is based on the MOSES-TRIFFID model described by Cox et al. (1999) and includes mechanistic formulations of the physical, biophysical, and biochemical processes that control the radiation, heat, water and carbon fluxes in response to hourly conditions of the overlying atmosphere. JULES has integrated coupling of photosynthesis, stomatal conductance, and transpiration (Best et al., 2011; Clark et al., 2010) so that the biophysical processes in the vegetation interact with hydrological processes in the soil and energy exchange between the land and the atmosphere.

Much work has been carried out to evaluate the performance of JULES at specific sites against detailed process data (Blyth et al., 2010). Although progress in model process representation has been made in this way, it is not possible to answer the question of whether such model calibrations are effective at the larger, global scale. However, recent versions of JULES (JULES version 2) include provision for global-scale runs, meaning that it is now possible to test the models at this scale. This development allows for a new form of model calibration, testing the model against data that is appropriate to the spatial scale of the application. Thus we are in a position to present a set of benchmark datasets to quantify the performance of the global land surface model: a set of data, a set of metrics to quantify performance and a set of model runs. This paper describes the data chosen for this task, the metrics and the results of the JULES model in this Benchmarking System.

Similar initiatives, such as the CLAMP project (Randerson et al., 2009) and the study by Cadule et al. (2010) have been used to test the carbon cycle of land surface models. Indeed, Cadule et al. (2010) evaluated the performance of MOSES-TRIFFID (Met Office Surface Exchange Scheme with the Dynamic Vegetation Model TRIFFID – essentially the same as JULES) carbon cycle, however, in coupled climate-carbon cycle mode. There is a need to evaluate the land surface component offline from the climate model because the climate forcing data may be the most important factor determining the response of the land surface mode, not the land surface model itself. In addition, the Global Soil Wetness Project 2 (GSWP2) intercomparison (Dirmeyer et al., 2006) presented data and model protocols to test the ability of the models to reproduce the land-part of the water cycle. However, the data and the tests described in this study represent the first comprehensive, carbon and water sets of data and tests to define the performance of the interacting carbon and water cycle model. By evaluating the carbon and water cycles at the same time, it may be possible to provide a more stringent test of the land surface models and be possible to identify why there may be a problem, as well as when and where.

Our first criterion for defining benchmark tests is to identify and select datasets that: are independent of any model (i.e. as far as is possible, not a derived data set), are global, have several years of monthly data available, and that describe either the carbon or water surface fluxes or stores.

The datasets and metrics chosen for this exercise are described in Sect. 2. An example how the data sets can be used to test the JULES model, including a description of the JULES model simulations and the input data used is given in Sect. 3. The results of this test of the JULES model is then shown in Sect. 4. A discussion of this benchmarking exercise is then given in Sect. 5 and Conclusions of the study are summarised in Sect. 6.

2 Datasets selected for JULES benchmarking

2.1 FLUXNET data for fluxes of water and carbon dioxide

It is possible to measure hourly turbulent fluxes of water, heat and carbon dioxide using eddy correlation instruments. Sitting 2 to 10 m above the canopy or surface, they measure the hourly fluxes from an area about 100 m upstream of the instrument. Several hundred of these instruments are used to routinely measure the water, heat and carbon fluxes. The data are collated under the banner "FLUXNET" (Baldocchi et al., 2001, www.fluxnet.ornl.gov). At FLUXNET sites additional instrumentation is used to record concurrent site climatology. Since the observations are taken at a sub-diurnal frequency, it is possible to use the day- and night-time CO_2 fluxes to diagnose separately the photosynthesis and the respiration fluxes, as photosynthesis drops to zero at night, allowing for a temperature-based predictor of the respiration to be created for each site which is applied to calculate the day time fluxes (this procedure is carried out by CarboEurope). Having measurements of water and carbon fluxes at the same sites, allows us to identify the links between the water and carbon stores and emissions. Stockli et al. (2008) and Blyth et al. (2010) have both used the data to evaluate and develop the land surface models.

It was decided that a small number of FLUXNET sites would be used in this initial benchmarking system to allow researchers to see in a glance the overall performance of the model. For the present study, ten FLUXNET sites (nine for CO_2, as the data were missing) were selected to sample a range of climate zones (temperate, Mediterranean, tropical and boreal) and plant functional types and soils. Blyth et al. (2010) demonstrated that using a set of 10 was adequate for identifying key element of the performance of the model. The locations of the 10 sites are shown in Fig. 1a and the plant functional types specified by Fluxnet and climate for the selected sites are summarized in Table 1 along with the mean annual weather data. A single year of data is chosen for this comparison to highlight the response of the observations (and the model) to conditions that may not be the climatological mean for that region. For instance, we chose a very dry year for the Amazon forest site comparison which enables us to assess the ability of the model to reproduce drought conditions. The problems of energy closure (see Valentini

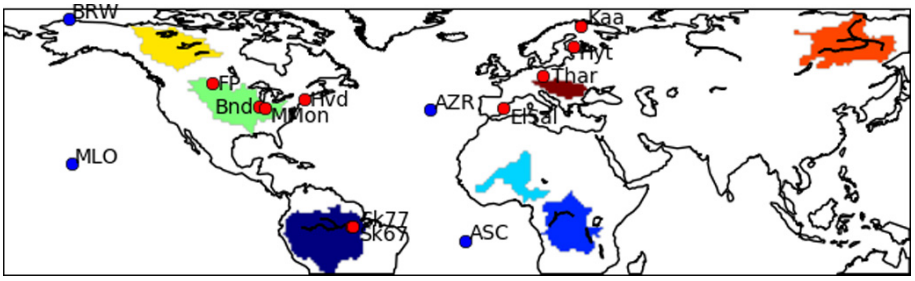

Fig. 1a. Map of FLUXNET stations, CO_2 stations and River basins.

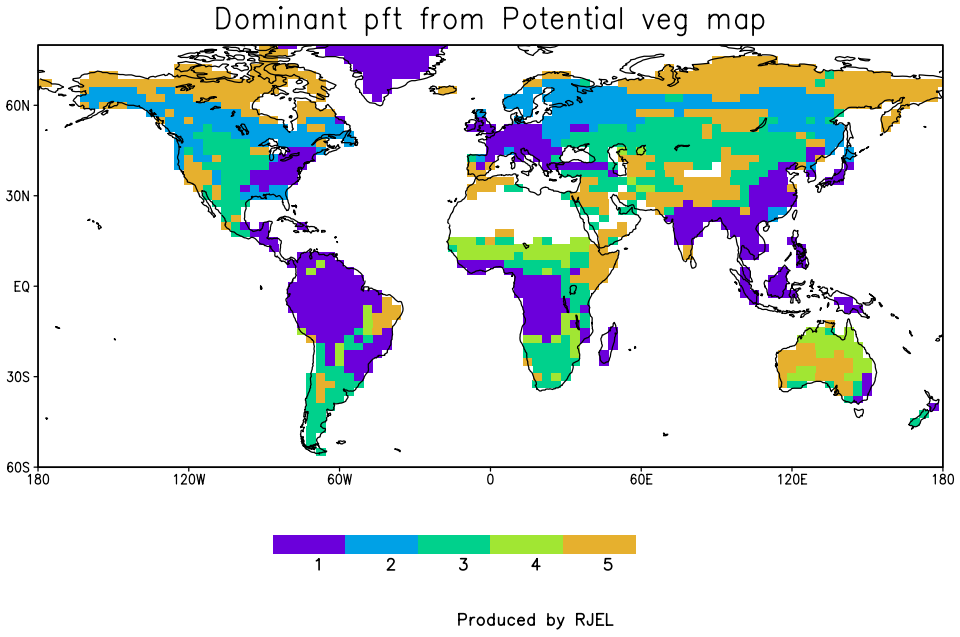

Fig. 1b. Potential vegetation map.

and Verma, 2002), are addressed by scaling the observed evaporation (or latent heat) by the ratio of the modelled to observed available energy (latent heat plus sensible heat: see Blyth et al., 2010).

The metric chosen to quantify the model performance against the benchmark data is the RMSE (Root Mean Square Errors) of the mean monthly fluxes of evaporation (scaled by the modelled available energy) and carbon dioxide. For carbon dioxide flux, both the uptake through photosynthesis and the release through respiration (plant and microbial) are quantified.

2.2 Mean monthly flask CO_2 from 4 stations around the world for 10 years

A robust independent measurement of the global carbon cycle can be found in the series of measurements of atmospheric carbon dioxide concentration made in various remote locations around the world. A glass flask is filled with air, which is then analysed to quantify its composition. Although the upward trend of carbon dioxide concentration is the most obvious and important diagnostic of these data, it also reveals the seasonal variation in the concentration of CO_2.

Heimann et al. (1998) and Fung et al. (1983) among others, show that the seasonal variation changes with location around the globe. The seasonal cycle in atmospheric CO_2 concentration results largely from the differential response of plant and soil-microbial activity to seasonal climate variation. Northern Hemisphere stations have greater seasonal cycle amplitude than those in the Southern Hemisphere, reflecting greater land-mass in the north. Also the amplitude of the seasonal cycle increases with latitude in the Northern Hemisphere, due to increasing seasonality in climate with latitude. We used the global inverse model of atmospheric tracer transport model TM2 (Kaminski et al., 1999a, b). The "station matrices" for the adjoint model (see Kaminski et al., 1999a, b) were derived from the TM2 (Heimann et al., 1998) using the mean wind fields from ERA40 the ECMWF reanalysis for the year 1987. The monthly station

Table 1. Summary climate and site descriptions of F:LUXNET sites.

Site name (year selected for comparison)	Plant functional type	Available years of data	Climate zone	Mean temperature (°C)	Annual precipitation (to nearest 10 mm)
Kaamanen (2002)	Wetlands (designated grass)	6	Boreal	−1	440
Hyytiala (2000)	Evergreen Needleleaf	10	Boreal	5	530
Morgan Monroe (2002)	Deciduous Broadleaf	7	Temperate	13	1150
Harvard (1999)	Deciduous Broadleaf	13	Temperate	9	1030
Tharandt (1999)	Evergreen Needleleaf	8	Temperate	9	640
Bondvill1e (1998)	Cropland, soybean, maize rotation (designated grass with variable LAI)	10	Temperate	13	930
El Saler (1999)	Evergreen Needleleaf	5	Mediterranean	17	440
Fort Peck (2006)	Grass	7	Mediterranean	7	320
Santarem Km 67 (2003)	Evergreen Needleleaf	3	Tropical	25	1290
Santarem Km77 2003	Grass	5	Tropical	26	1610

records were post-processed to extract the detrended, mean monthly observations between 1980–1990 and same procedure was applied to model outputs before the net monthly fluxes were supplied to the transport model. Monthly ocean-atmosphere fluxes were taken from an ocean carbon cycle (HAMOCC3, Maier-Reimer, 1993) and monthly CO_2 emission fields from fossil fuel and cement production were based on Marland et al. (1989), assuming no seasonality in emissions. The model that produced these matrices of contributions performed reasonably well in the TRANSCOM experiment. It does not represent the state-of-the-art with respect to atmospheric transport model analyses, but is a practical alternative that can be distributed to the JULES community as part of the benchmarking exercise. Given the over-view of seasonal fluxes benchmarking that is being delivered it was felt that this was adequate. Applying this transport model, the modelled fluxes of carbon dioxide from the land surface can be integrated and transported to the location of the measurement. It is then possible to compare the integrated signal of seasonal variation of carbon dioxide release and uptake by the land surface models by comparing simulated atmospheric concentrations with observations.

The metric chosen for this data is the RMSE of the mean monthly CO_2 for three stations which represent different zones: MLO (representative of the global mean), BRW (representative of the high northern latitudes), AZR (representative of the Northern Hemisphere) and ASC (representative of the Southern Hemisphere). Figure 1a shows the location of the flask stations chosen for this study.

2.3 Monthly river flow records from 7 major rivers for 10 years

Over time periods of several years, the terrestrial water cycle is in approximate balance, with precipitation over a region balancing evaporation and runoff. Both precipitation and evaporation vary strongly over the scale of a large river basin, making it difficult to estimate areal values, although recent advances in remote sensing suggest that better estimates will be available in future. Runoff also varies greatly with location, but a river network integrates the net runoff over a catchment, meaning that historical records of river flow are an important data source against which models can be compared (e.g. Miller et al., 1994). Even so, there are difficulties with using these data, including gaps in the record and uncertainty in the measurements. Further, the flow regimes in large river basins include the effects of human management such as dam operation, which alters the timing of flow, and extraction for water supply, which alters the amount of water. These human interventions are not considered in most global models.

For this study, a selection of relatively little-managed rivers are chosen that represent the same north-south gradient that we have for the FLUXNET data. A small selection was chosen so that the overall picture could be obtained with one set of graphics. Initially 8 river basins were chosen, including the Parana. However it was not possible to find a stretch of that river (in that very productive region of South America) that was not heavily managed. So we reduced the number to 7, covering the key regions of the Americas, Europe and Africa. In the Americas we have the Mackenzie (at Arctic Red river), Mississippi (at Vicksburg) and Amazon (at Obidos) rivers, while Europe and Africa are represented by the Lena (at Kyusyr), Danube (at Central Izmail), Niger (at

Malanville) and Congo (at Kinshasa). Monthly flow for these rivers was obtained from the The Global Runoff Data Centre, 56068 Koblenz, Germany, http://www.bafg.de/GRDC.

Their locations are shown in Fig. 1a. Data for the Congo were not available for the period in question, so the mean monthly data for 1903–1983 (Global Runoff Data Centre, 2008) were used instead. Although there is some modification of these rivers by human activities, we consider that they are still useful for this benchmarking exercise.

The measured flow rate ($m^3\,s^{-1}$) is converted to an equivalent per unit area of the catchment ($mm\,d^{-1}$) and the metric chosen is the RMSE of the monthly mean flow, normalised by the observed average flow.

2.4 Monthly phenology: NDVI and LAI over 7 catchments for 10 years

One of the most important aspects of the vegetation control of the carbon and water balance is through the seasonal variation of leaf growth, the phenology. The greening up draws down carbon dioxide from the atmosphere, and the litter fall and subsequent decomposition releases carbon back. The plant growth is dependent on it being warm enough for growth and sufficient soil water available to plants for transpiration. It is therefore sensitive to the weather and soil moisture status. The contrasting dynamics of temperature and moisture on seasonal plant growth are most apparent in semi-arid regions such as the Sahel and in areas of seasonal freezing such as the sub-artic tundra.

Semi-arid areas are naturally areas of the world with low population and therefore tend to have a paucity of data. However, it is possible to identify the seasonal growing patterns of vegetation from satellite data, perhaps with the exception of tropical forests where the seasonal signal is small (Sellers et al., 1996). Seasonal increase in vegetation greenness results in the annual drawdown of atmospheric CO_2 which can be seen in the atmospheric CO_2 concentration data (see Sect. 2.2). The Normalized Difference Vegetation Index (NDVI) observed from satellite is near linearly related to the fraction of photosynthetically active radiation absorbed by the green parts of vegetation and exponentially to leaf area index (LAI) (Zhangshi and Williams, 1997). Dependent on the clumping of leaves the NDVI can saturate for LAI values above 3–5 (Clevers, 1989; Carlson and Ripley, 1997).

A data set of NDVI based on AVHRR (James and Kalluri, 1994) and SeaWiFS (Vermote et al., 2001) for 25 years is used in this study (Los et al., 2007). The data is corrected for residual errors in sensor degradation (Los et al., 2005), BRDF effects (Los et al., 2001), and cloud contamination (Sellers et al., 1996). SeaWiFS and AVHRR data are merged by calculating the means seasonal cycle for their common period (1998–1999); subtracting this from the respective data sets, scaling the anomalies to the standard deviation of the AVHRR data on a per-pixel basis and adding back the mean seasonality of the AVHRR data.

To ease comparison with the results of the water-balance information, the tests are made in the areas covered by the 7 rivers in the previous test, shown in Fig. 1a. The range of observed NDVI for each catchment is linearly scaled with the range of the modelled LAI to give an "observed LAI". For the Lena catchment, the linearization was only applied up to an LAI of 3 and was assumed to saturate above that value. The resulting RMSE of the mean monthly values of LAI are used to quantify the performance of the model.

2.5 Global vegetation map – fractional coverage of the 5 PFTs

The Dynamic Vegetation Model in JULES (TRIFFID, Cox et al., 2001) aims to calculate the natural vegetation for a given climate: the locations of the world's tropical rainforests, the boreal forest, the great grass plains and the deserts. In order to test TRIFFID (or a similar model), we need a map of this "Potential Vegetation" i.e. an estimate of undisturbed global vegetation cover. One example is the SAGE (centre for Sustainability And the Global Environment) global potential vegetation data set (Ramankatty and Foley, 1999). To compare it to the model output from JULES, the Plant Functional Types (PFTs) from the SAGE data set are aggregated into the 5 PFTs in JULES using the mapping in Table 3, and then aggregated spatially, up to the spatial scale of model run ($2.5° \times 3.75°$). Figure 1b shows the dominant PFTs for the aggregated data, although the fractional coverage for each PFT is available as well.

To sample the set of climate conditions that produce a particular vegetation type, and to avoid potentially arbitrary choices of geographical regions, the areas with a dominant vegetation type of each PFT were grouped together. In this way, the climates have been self-selected. The metric chosen here is the total error in the percentage difference for each bar chart.

3 Model setup and results

3.1 Model description

The model used in this benchmarking is the JULES model. It is the land surface model used within the Hadley centre GCM. The description of the model is given in Best et al. (2011) and Clark et al. (2011). It is a community model and is distributed via its website: www.jchmr.org/jules which contains further content about the model. It is a mechanistic model of the land surface including linked processes of photosynthesis and evaporation, soil and snow physics as well as plant growth and soil microbial activity. These processes are all linked through a series of equations that quantify how the soil moisture and temperature govern the evapotranspiration, heat balance, the respiration, photosynthesis and carbon assimilation. It runs at a subdaily step, using meteorological

Table 2. Summary metrics of performance of JULES against datasets.

	Summary of model performance	Metric: RMSE of monthly values
FLUXNET - evaporation	Modelled evaporation is higher than that observed. The seasonality is captured well, except in areas of seasonally frozen soils and in the tropics.	21 W m^{-2}
Fluxnet – carbon	For the temperate forests, the photosynthesis is underestimated while the evaporation is overestimated. For the wetlands and tropical Forests, the photosynthesis is overestimated and the evaporation is underestimated. For the rest, the errors are the same so that if the evaporation is overestimated, then so is the photosynthesis. In all cases, the respiration is overestimated.	2.0 µMol m^{-2} (GPP) 1.6 µMol m^{-2} (Respiration)
Atmospheric CO_2	Seasonality is captured well, apart from low latitudes where the model gives too much seasonality.	3.7 µMol m^{-3}
River flow	Seasonality is generally captured well, except for in the dry areas. Temperate areas have too little river flow. Peak flow in cold regions is modelled poorly.	0.31 mm per month
LAI	Seasonality is captured well, apart from very cold regions, where the observed high seasonality is under-estimated in the model. The LAI or the temperate regions appears to be too low. Seasonality is low in the tropics for both observations and the model.	0.266 LAI
Land cover	There is reasonable agreement between the model and the potential vegetation map, although the model appears to have more "shrub" and "bare soil" than the observations.	8.9%

Table 3. Mapping from SAGE vegetation classes to JULES PFTs.

	BL	NL	C_3	C_4	Shrub	Bare soil
Tropical evergreen	0.9					0.1
Tropical deciduous	0.8		0.15			0.05
Temperate BL EG	0.9					0.1
Temperate NL EG		0.8	0.15			0.05
Temperate deciduous						0.1
Boreal EG		0.8	0.15			0.05
EG/Deciduous mixed	0.4	0.4	0.1			0.1
Savanna	0.2		0.75			0.05
Grassland/Steppe	0.9					0.1
Dense shrub			0.15		0.7	0.15
Open Shrub			0.6		0.3	0.1
Tundra			0.35		0.35	0.3

drivers of rainfall, incoming radiation, temperature, humidity and windspeed as inputs.

3.2 Point comparison with FLUXNET data

JULES model forcing data were extracted for each benchmark site. The model parameters were not tuned; neither to the flux data nor to the known local vegetation properties (rooting depth etc.). Instead, parameters for the model were taken as though it was embedded in the GCM. Hence, the default values of soil properties and vegetation properties were taken from the look-up tables used in the global operational version of the model, and not from the site data. The exception was the monthly Leaf Area Index, to which the results are particularly sensitive, which was specified for each site based on information available from the FLUXNET site description. The testing of the performance of the prediction of LAI by JULES is done in the next set of tests.

3.3 Global gridded runs of JULES: comparison with distributed data: atmospheric CO_2, river flow, NDVI and land cover

The distributed version of the JULES model was run with a resolution of 1 degree, using the GSWP2 forcing data. The driving data are available for 1986–1995. It is necessary to spin up the soil carbon so the model was run through this data 5 times. The final ten year run was used in the analysis. The model was run in two modes: one with the competition switched off and land cover specified from the CLIMAP, and the other with the plant competition (TRIFFID) switched on and with a spin up of 100 years. This latter run was used only in the comparison with the land cover map.

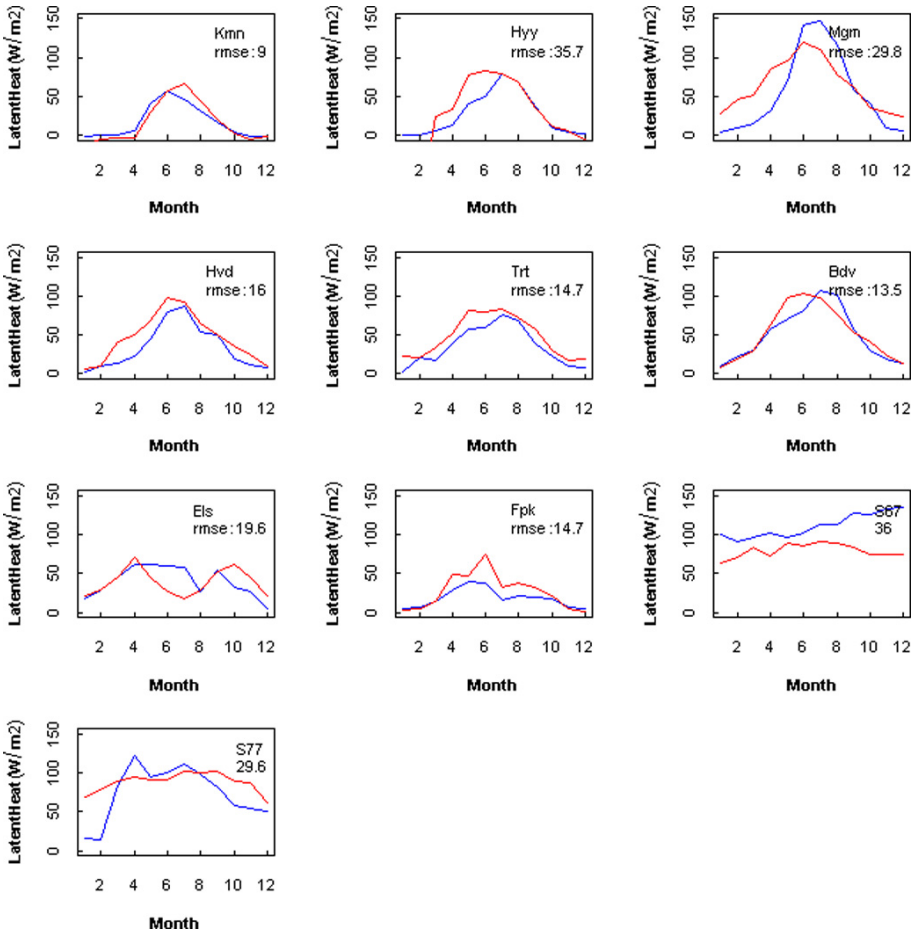

Fig. 2. Modelled (red) and observed (blue) evaporation fluxes at 10 FLUXNET sites (Kmn: Kaamanen, Hyy: Hyytiala, Mgn: Morgon Monroe, Hvd: Harvard Forest, Trt: Tharandt, Bdv: Bondville, Els: El Saler, Fpk: Fort Peck, S67: Santarem km67, S77: Santarem km77).

Outputs needed to compare to the data are monthly values of atmospheric CO_2 anomalies at the four flask stations, monthly values of leaf area index and the surface and subsurface runoff. A routing model (TRIP, Oki et al., 1999) is used to translate the runoff generated at the grid-cell into an equivalent river flow which combines the attenuation, delay and integration of the water across the catchment.

4 Results

Table 2 gives a descriptive summary of the results of each test with a quantitative estimate of the error in the process.

4.1 Metrics

In order to provide the user with a simple assessment of the error in the model, the same diagnostic is used for each of the datasets: the mean monthly value of the property that is observed. This allows us to compare very different types of observations. For some of the observations this was straightforward such as the river flows, the fluxnet data and the atmospheric CO_2. However, for distributed datasets such as the NDVI series, it is not obvious how to reduce the observations and model output to single time-series of mean monthly values. For this system, we decided to encapsulate the seasonality of the NDVI and LAI from the model by looking at the area-average value of the selected river basins. Most of the river basins are fairly uniform in climate and vegetation type, and so this represents a simple way of representing the mean response of the plants to climate in terms of phenology, which also allows us to compare the plant response directly to the response of the water balance through the river flow. The exception is the Niger River which passes through very contrasting climates.

In this analysis, a simple test of assessing the strength of the seasonality was used, as it represents a first-order test of the performance of a land-surface model. A future version should include interannual variability as this would allow us to assess the performance of the model to extremes of weather such as El Nino or La Nina years. Such an assessment is feasible with the current model configuration.

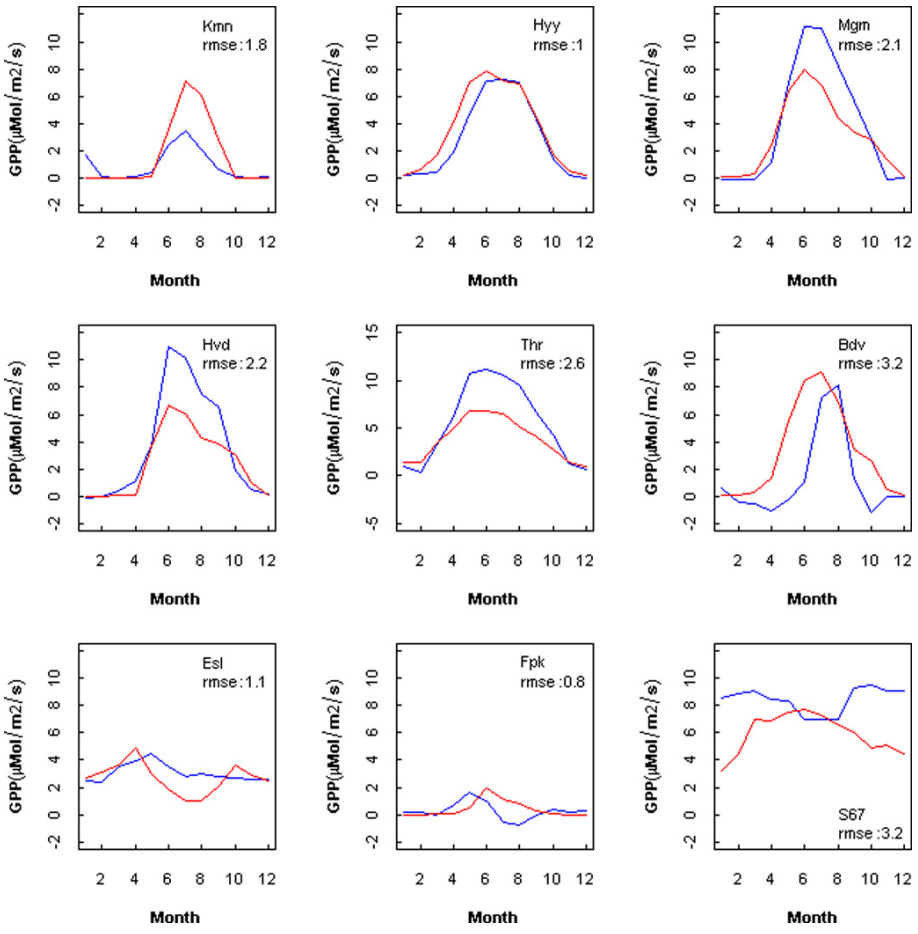

Fig. 3. Modelled (red) and observed (blue) GPP fluxes at 9 FLUXNET sites (Kmn: Kaamanen, Hyy: Hyytiala, Mgn: Morgon Monroe, Hvd: Harvard Forest, Trt: Tharandt, Bdv: Bondville, Els: El Saler, Fpk: Fort Peck, S67: Santarem km67).

No attempt was made in the benchmarking at this stage to include uncertainties in the metrics, or to combine them in any way. Combining the metric across the various measures of performance requires a deeper understanding of uncertainty and further work is required to address this. Instead the simple Root Mean Square Error of the mean monthly quantity is presented. The idea is that researchers using the JULES model can compare the model with any changes that have made to these simple diagnostics.

4.2 Model results for the 10 site-based runs

For the selected FLUXNET sites, Fig. 2 shows comparisons between the monthly average modelled evaporation and the normalised observed evaporation. Figure 3 shows the monthly Gross Primary Productivity (GPP) for each site and Fig. 4 the monthly respiration. Table 2 provides an overview of the error in the modelled seasonality of evaporation and carbon dioxide. In general, for the temperate forests (Harvard, Tharandt and Morgan Monroe), the photosynthesis is underestimated while the evaporation is overestimated. For the wetlands (Kaamanen) and tropical Forests (Santarem 67), the photosynthesis is overestimated and the evaporation is underestimated. For the rest (Fort Peck, El Saler, Bondville), the errors are the same sign: if the evaporation is overestimated, then so is the photosynthesis. It is also clear that in all cases the respiration is overestimated.

The errors in this analysis are rather large. This is mainly due to the simple approach of using a fixed, predetermined LAI for these varied sites. The errors can be reduced at a site by calibrating and adjusting the parameters for the local conditions. However, what is more interesting for this assessment is the way the evaporation error agrees or contrasts the error in the GPP. It is the relative errors that allow the model user to identify process anomalies.

4.3 Comparison with the atmospheric CO_2 observations

Figure 5 shows the results of the comparison of the observed mean-monthly atmospheric CO_2 concentrations with the model output. The two northern-latitude stations, AZR

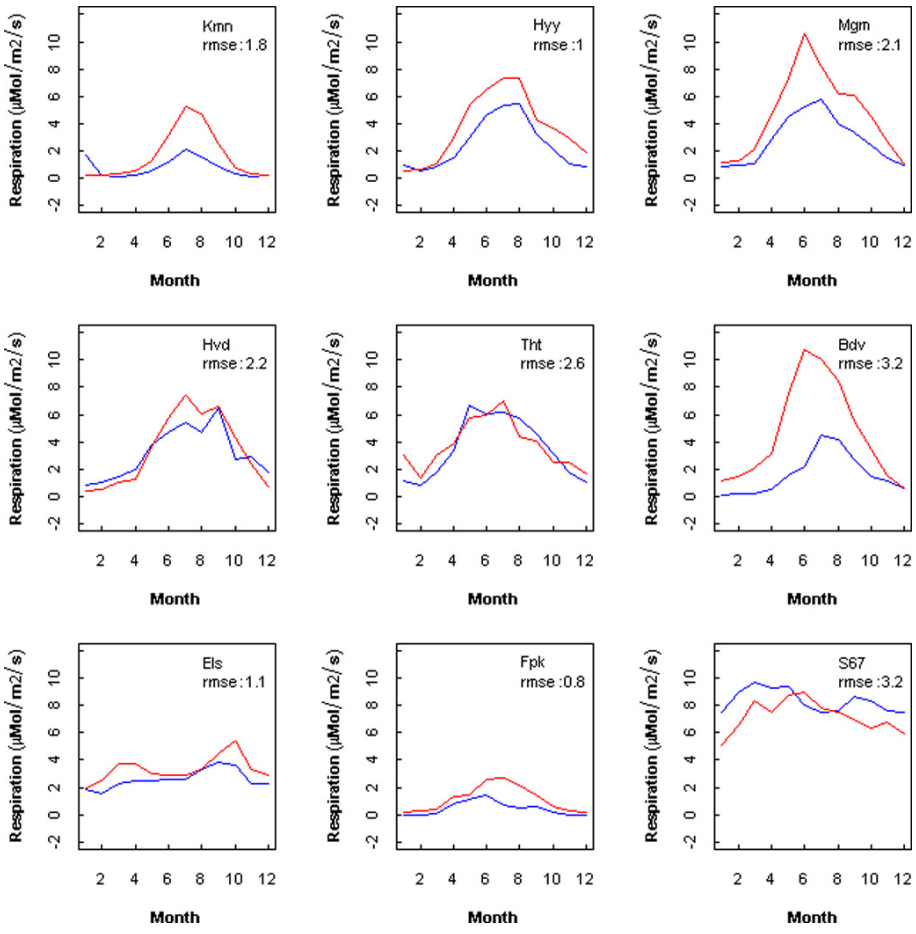

Fig. 4. Modelled (red) and observed (blue) respiration fluxes at 9 FLUXNET sites (Kmn: Kaamanen, Hyy: Hyytiala, Mgn: Morgon Monroe, Hvd: Harvard Forest, Trt: Tharandt, Bdv: Bondville, Els: El Saler, Fpk: Fort Peck, S67: Santarem km67).

and BRW (Fig. 5a and b) show strong seasonality in the observed and modelled atmospheric CO_2 concentration. This indicates that the model is correctly simulating the drawdown of CO_2 in the spring as the vegetation greens up. The autumn release of carbon due to soil-decomposition of the leaf-drop is also apparent in an increase in atmospheric concentrations of CO_2 at these stations in both the observations and the model. The greening up appears to be a bit early in the far north station (BRW, Fig. 5b) which links to the early modelled increase in LAI compared to NDVI (see the Lena catchment result, Fig. 6a) and the early increase in evaporation at the Hyytiala site (Fig. 2b).

The Southern Hemisphere station (ASC, Fig. 5c) shows very little observed seasonality of CO_2, and a large modelled seasonality, with a maximum in the months October to December. This discrepancy is probably due to the overestimate of the soil respiration, also shown in the FLUXNET data comparisons (see Sect. 4.2), which is possibly due to incorrect initial soil carbon contents, although a thorough analysis would be needed to pin-point the exact explanation.

The global average (MLO, Fig. 5d) shows that overall, the model has too high a seasonality in the net uptake of carbon which results in the strong atmospheric concentration of carbon dioxide.

4.4 Diagnostics chosen for the monthly river flow records from 7 major rivers for 10 years

Mean monthly riverflows are compared to the JULES-TRIP model and the results are shown in Fig. 6. The two cold-region rivers do not show consistent results: modelled flow in the Lena (Fig. 6a) is too low, while in the Mackenzie (Fig. 6c) it is too high. The two temperate rivers, the Mississippi (Fig. 6b) and the Danube (Fig. 6d), both show insufficient flow, implying too much loss of water through evaporation. This result agrees with the FLUXNET comparisons in these temperate regions where the evaporation is too high (sites Fort Peck, Tharandt and Harvard Forest in Fig. 2d, e and h) and are probably not a result of LAI being too high, since the LAI of these catchments is either right or too low (see Fig. 7b and d). In the humid tropics, the seasonality

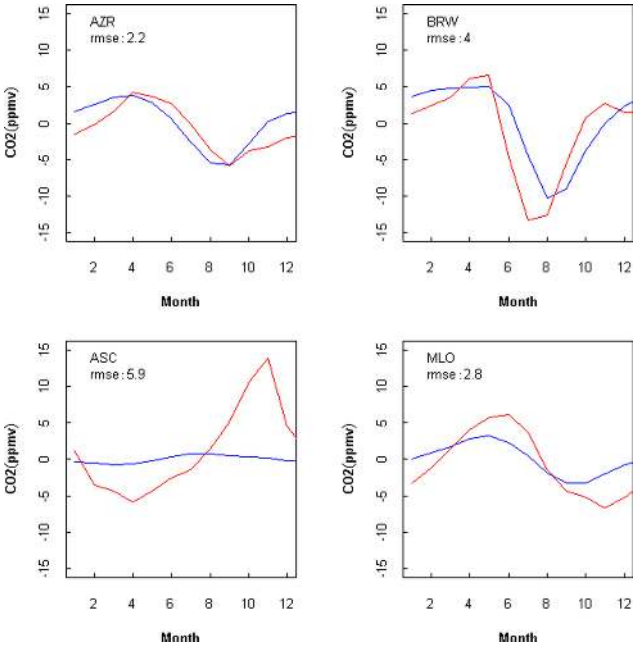

Fig. 5. Modelled (red) and observed (blue) seasonal variation of atmospheric carbon dioxide at four stations: Azores (AZR), Barrow (BRW), Ascension (ASC), and Mauna Loa (MLO).

is captured well in the Amazon, whereas the modelled flow varies too much in the Congo. In the Niger, the driest region, the model shows too much variability through the year and generally too much flow. Contributions to the excessive flow could come from losses from the river to groundwater and evaporation, and extractions, none of which are modelled.

4.5 Diagnostics chosen for the monthly NDVI for 10 years for 7 river catchments

Figure 7 shows the results of these model runs. The cold region catchments (Lena and Mackenzie, Fig. 7a and c) both show more seasonality in their observed NDVI than modelled. For the temperate catchments (Mississippi and Danube, Fig. 7b and d) the seasonality in NDVI is good, but the overall value is too low. This does not seem to affect the water balance (see Sects. 4.1 and 4.3) but is linked to the underestimation of the photosynthesis (see Sect. 4.2). The tropical catchments (Amazon and Congo, Fig. 7e and f) have a reasonable (low) seasonality to their LAI, although it is clear that the observations have at least some seasonality while the model assumes none. This does not seem to affect the water balance however. Finally, the dry catchment (Niger, Fig. 7g) has some problems matching the observed seasonality of the NDVI. The phase of the Niger NDVI is not the same as the modelled, with the LAI increasing later in the season that the observed NDVI. It is not possible to link this to any deficiencies in the water balance modelling, since the total water balance was not modelled well, with high runoff in the winter periods, compared to very low all round runoff in the observations (see Fig. 6g). It is clear that the model has some difficulties representing the true water and carbon balance of these regions.

4.6 Diagnostics chosen for the global vegetation map

Figure 8 shows the modelled fractional coverage for each PFT in the areas where the named PFT is dominant (Fig. 7a, NT, etc.) and also shows the mean observed fractional coverage of the PFTs in those grid squares. The Bare Soil fractional cover is an input to the model, so the comparison naturally does well, but does not inform us of model performance. The other known PFT covers have problems: there is too much bare soil in the grass and shrub regions. This might be partly a matter of definition (e.g. sparse vegetation could be seen as partially bare soil, or just sparse) and adjustments may need to be made. There is also much more grass in the observations than in the model. However, it should be noted that the land cover maps have a classification accuracy of about 70%, including desert and tropical forest. Confusion tends to occur between grasses and shrubs, agriculture and broadleaf seasonal forests etc, so that discrepancy of these land cover maps may not all be the models fault. The model tends to grow shrub and bare soil in its place. In addition, the model chooses shrub in many places where the observations have defined trees, either broad leaf or needle leaf.

Despite these discrepancies, the general location of the broad-leaf verses needle-leaf is well captured, even if the exact amounts are not correct.

5 Discussion

The objective of this study is to identify a range of datasets that to check the performance of the JULES model in its ability to represent the land surface components of the global water and carbon cycles. There are a very large number of datasets now available, and that relate to land surface functioning. For instance there are several global satellite products of NDVI and fPAR, and all for different time scales, spatial scales and lengths of time. There are also many river records, an ever increasing set of FLUXNET stations, many Atmospheric CO_2 stations, and several Land-Cover map products. Ideally JULES predictions (or from any other land surface model) could be considered, and with a knowledge of measurement errors, a list determined as to which ones would tell us the most about the model performance. However, pragmatism suggests that such a comprehensive prototype benchmarking system might be difficult to achieve in the first instance. Hence, to initialise this project, data chosen in this study were chosen after consultation with those most familiar with each form of data. This is sometimes referred to as proceeding based on "expert opinion". A further factor in our decision to select the measurements of

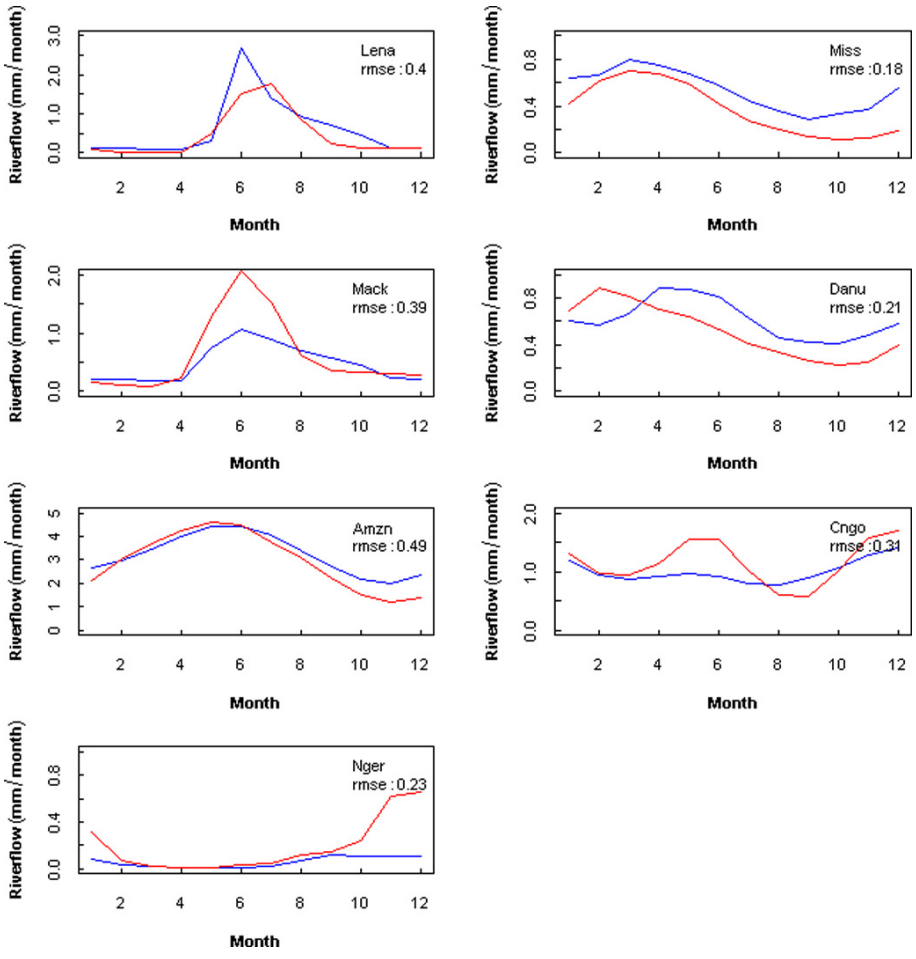

Fig. 6. Modelled (red) and observed (blue) seasonal river flow from 7 river catchments (Lena, Miss: Mississippi, Mack: Mackenzie, Danu: Danube, Amzn: Amazon, Cngo: Congo, Niger).

FLUXNET, river flow, CO_2 concentrations, satellite measurements of "greenness" and land cover is that we wished to build a benchmarking system using primary (i.e. as far as possible, not modelled, or model-enhanced) data.

We have also asked if the tests are chosen correctly to test the performance of the JULES model. Various choices were made in choosing the tests. It is possible to switch different options on and off in the model, or to over-ride particular components with measurement, whilst considering the different aspects of the time- and space-varying fluxes and stores of carbon and water. For instance, the basic processes of photosynthesis and transpiration are tested using the FLUXNET data for sites representative of the major global biomes whilst specifying the leaf area, whereas the process of growth and phenology are tested using the NDVI and atmospheric CO_2 with the model with-phenology and PFTs being specified. Plant-competition is tested with the PFT competition switched on and comparing with the land cover maps.

Ultimately we wish to use benchmarking to address is to what extent can the performance of the model in these tests be used in our assessment of whether the model is "good enough" for the purposes of modelling the global terrestrial water- and carbon-balance. The issue of what constitutes a "good enough" performance has been raised by some in the scientific community. The CLAMP system (Randerson et al., 2009) and Cadule et al. (2010) gave metrics for the model performance, depending on a qualitative assessment of the importance of the process being tested and the certainty they have in the data. However, the authors of that study made it clear that the assessments were not absolute and not final. Meanwhile, Abramowitz et al. (2008) laid out a method to determine an a priori method of assessing whether a model is "good enough" by using a statistical "model" as a benchmark for the performance of a mechanistic model. The theory in this study is that if the mechanistic model performs worse that a simple statistical one, then it has "failed" the test. This concept is a useful one, but is subject to criticism from the modelling community as statistical models almost

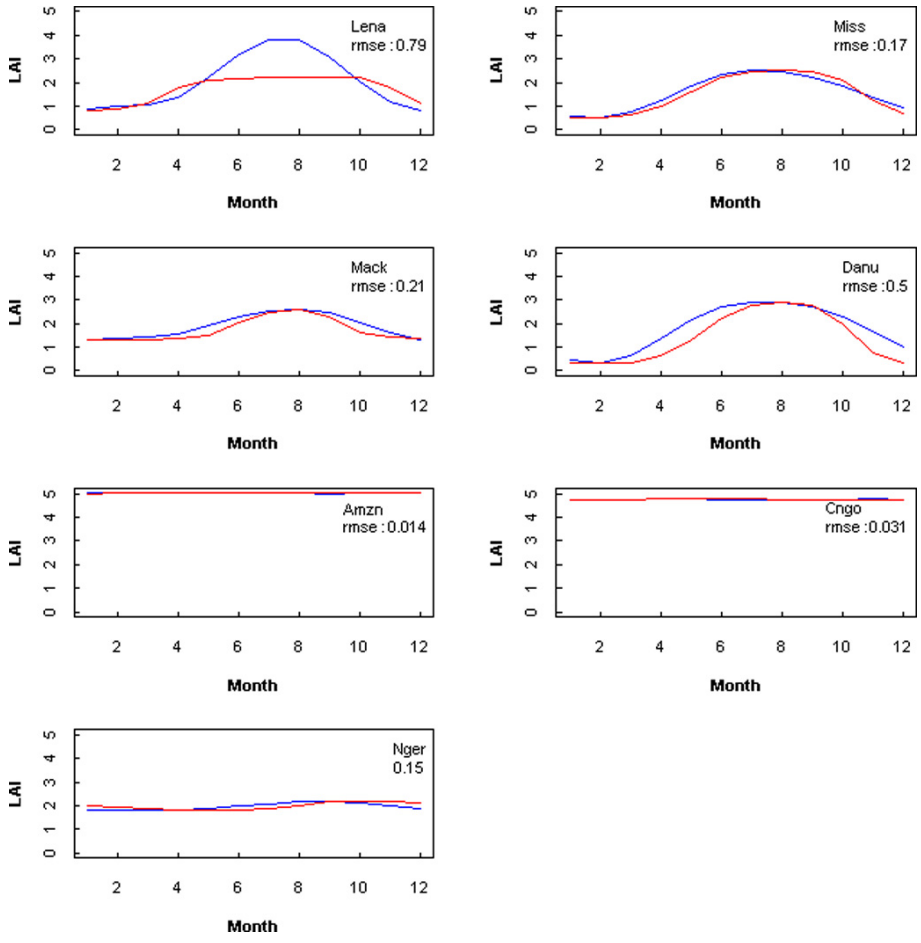

Fig. 7. Modelled (red, LAI) and observed (blue, NDVI) phenology from the 7 river catchments (Lena, Miss: Mississippi, Mack: Mackenzie, Danu: Danube, Amzn: Amazon, Cngo: Congo, Niger).

always out-perform a mechanistic model in current environmental conditions, but cannot be relied upon under changed conditions, whereas a mechanistic model that describes the correct processes can, in theory, be used for all conditions.

6 Conclusions

A benchmarking system has been built for the JULES model. It includes 5 basic datasets that cover a range of spatial (point to global) and temporal (hourly to monthly) scales: a set of hourly fluxes from 10 FLUXNET stations, monthly concentrations of atmospheric CO_2 from 4 stations representing the Northern and Southern Hemispheres and the globe, monthly NDVI observed from satellite across the world at a spatial resolution of 1 degree, the monthly river flow from 7 large rivers basins around the world and the potential (i.e. no agriculture) vegetation map of the world are all considered.

The system includes set model runs of the JULES model. At FLUXNET sites, observed meteorology and leaf area were input, and modelled fluxes of water and CO_2 were compared with eddy correlation measurements. For global runs the GSWP2 meteorological data at 1° resolution were used and model values compared with observed PFT distribution, atmospheric CO2 concentration, the NDVI and the river flows.

The tests demonstrated some weaknesses of the model, many of which are being worked on by researchers. It is important that the benchmark data remains somewhat independent of model development and is only used to indicate gross errors. However, more importantly, the tests demonstrated how to build an integrated test of the combined global terrestrial water and carbon budgets: to build a suit of tests that address the different aspects of the model. The study highlights that there are currently gaps in the data sets needed for a comprehensive system. For instance, it would be improved if global satellite data of snow cover and or some other aspect of the surface energy balance (temperature, moisture, evaporation or surface stress see Ellis et al., 2009) could be included in the system. In addition, other carbon-sensitive data could be used, such as biomass estimates or new EO products of atmospheric CO_2.

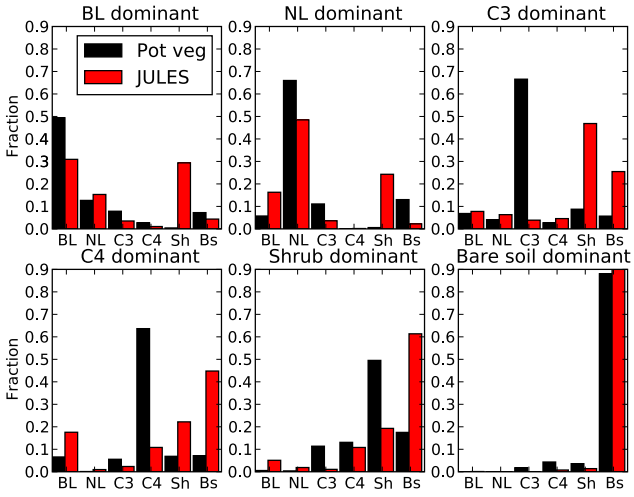

Fig. 8. For all the cells that contain the title PFT as dominant in the Potential Vegetation map of SAGE (Ramankutty and Foley, 1999), the figures show the average modelled (red) and potential (black) fraction of land cover for the PFT shown on the x-axis (BL: Broad Leaf, NL: Needle Leaf, C3: C3 grass, C4: C4 grass, Sh: Shrub, Bs: Bare Soil).

By combining the carbon and water cycle benchmarks, it has been possible to make a more comprehensive assessment of the model. For instance, it is possible to identify whether a drop in atmospheric carbon dioxide is caused by an incorrect modelling of the photosynthesis process, of the respiration process or the hydrology process. Errors in the main processes in the model responsible for the carbon and water cycle can be located for different regions by the tests chosen in this study.

It is clear from this study that such a large-scale benchmarking system is a useful tool for identifying problems in the model performance. In particular, it allows the relative importance of different problems to be assessed, for instance soil moisture control over the carbon flux, over leaf growth control. However, the benchmarking system is not a replacement for more detailed process-based field data. It is recommended that regional field data is still used for model improvement and calibration.

It is also possible that the datasets and tests described here could be used in a model-to-model comparison. The benchmark tests could be extended beyond the development of a single model.

Acknowledgements. We gratefully acknowledge the FLUXNET PIs at all the sites from which data were taken for this study and the institutions and agencies that support data collection at these sites, including T. Laurila of the Finnish Meteorological Institute, T. Vesala of the University of Helsinki, and D. Dragoni of the Indiana University, S. Wofsy of Harvard University, C. Bernhofer of the Institut für Halbleiter- und Mikrosystemtechnik (IHM) Technical University of Dresden, T. P. Meyers of Air Resources Laboratory, National Oceanic and Atmospheric Administration (NOAA/ARL), M. J. Sanz of the Fundacion Centro de Estudios Ambientales del Mediterráneo (CEAM) Parque Technogogico, and D. R. Fitzgerrald of the University of Albany. River flow data were obtained from the Global Runoff Data Centre, Koblenz.

M. Pryor was supported by the Joint DECC, Defra and MoD Integrated Climate Programme – DECC/Defra (GA01101), MoD (CBC/2B/0417_Annex C5). C. Huntingford was supported by the CEH Science Budget.

Edited by: D. Lawrence

References

Abramowitz, G., Leuning, R., Clark, M., and Pitman, A.: Evaluating the Performance of Land Surface Models, J. Climate, 21, 5468–5481. doi:10.1175/2008JCLI2378.1, 2008.

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X.,.Malhi, Y., Meyers, T., Munger, W., Oechel, W., Paw, K. T., Pilegaard, K., Schmid, H. P., Valentini, R., Verma, S., Vesala, T., Wilson, K., and Wofsy, S.: FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapour and energy flux densities, B Am. Meteorol. Soc., 82, 2415–2433, 2001.

Beljaars, A. C. M., Viterbo, P., Miller, M. J., and Betts, A. K.: The anomalous rainfall over the United States during July 1993: Sensitivity to land surface parameterization and soil moisture anomalies, Mont. Weather Rev., 124(3), 362–382, 1996.

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Ménard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B., and Harding, R. J.: The Joint UK Land Environment Simulator (JULES), Model description – Part 1: Energy and water fluxes, Geosci. Model Dev. Discuss., 4, 595–640, doi:10.5194/gmdd-4-595-2011, 2011.

Blyth, E. M., Best, M., Cox, P., Essery, R., Boucher, O., Harding, R., Prentice, I. C., Vidale, P.-L., and Woodward, I.: JULES: a new community land surface model, IGBP newsletter, 6, 9–11, 2006.

Blyth, E. M, Gash, J. H. C., Lloyd, A., Pryor, M., Weedon, G. P., and Shuttleworth, J. W.: Evaluating the JULES model energy fluxes using FLUXNET data, J. Hydrometeorol., 11, 509–519, 2010.

Cadule, P., Friedlingstein, P., Bopp, L., Sitch, S., Jones, C. D., Ciais, P., Piao, S. L., and Peylin, P.: Benchmarking coupled climate-carbon models against long-term atmospheric CO_2 measurements, Global Biogeochem. Cycles, 24, 24 pp., doi:10.1029/2009GB003556, 2010.

Carlson, T. N. and Ripley, D. A.: On the relation between NDVI. fractional vegetation cover, and leaf area index, Remote Sens. Environ., 62, 241–252, 1997.

Clark, D. B., Mercado, L. M., Sitch, S., Jones, C. D., Gedney, N., Best, M. J., Pryor, M., Rooney, G. G., Essery, R. L. H., Blyth, E., Boucher, O., Harding, R. J., and Cox, P. M.: The Joint UK Land Environment Simulator (JULES), Model description – Part 2: Carbon fluxes and vegetation, Geosci. Model Dev. Discuss., 4, 641–688, doi:10.5194/gmdd-4-641-2011, 2011.

Clevers, J. G.: The application of a weighted infrared- red vegetation index for estimating leaf area index by correcting for soil moisture, Remote Sens. Environ., 29, 25–37, 1989.

Cox, P. M.: Description of the TRIFFID dynamic global vegetation model. Technical Note 24. Hadley Centre, Met Office, 17 pp., 2001.

Cox, P. M., Betts, R. A., Bunton, C. B., Essery, R. L. H., Rowntree, P. R., and Smith, J.: The impact of new land surface physics on the GCM sensitivity of climate and climate sensitivity, Clim. Dynam., 15, 183–203, 1999.

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.

Dirmeyer, P. A., Gao, X., Zhao, M., Guo, Z., Oki, T., and Hanasaki, N.: GSWP-2 Multimodel analysis and implications for our perception of the land surface, B. Am. Meteorol. Soc., 87, 1381–1397, doi:10.1175/BAMS-87-10-1381, 2006.

Ellis, R. J., Taylor, C. M., Weedon, G. P., Gedney, N., Clark, D. B., and Los, S.: Evaluating the simulated seasonality of soil moisture with earth observation data, J. Hydrometeorol., 10, 1548–1560, 2009.

Fung, I. Y., Prentice, K., Matthews, E., Lerner, J., and Russell, J.: Three-dimensional tracer model study of atmospheric CO2: Response of seasonal exchanges with the terrestrial biosphere, J. Geophys. Res., 88, 1281–1294, 1983.

Global Runoff Data Centre: Long-Term Mean Monthly Discharges and Annual Characteristics of GRDC Stations, Global Runoff Data Centre, Koblenz, Germany, 2008.

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. Cycles, 12, 1–24, 1998.

IPCC: Summary for Policymakers, in: Climate Change: The Physical 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 University Press, Cambridge, United Kingdom and New York, NY, USA, 996 pp., 2007.

James, M. E. and Kalluri, S. N. V.: The Pathfinder AVHRR land data set: an improved coarse resolution data set for terrestrial monitoring, Int. J. Remote Sens., 15, 3347–3364, 1994.

Kaminski, T., Heimann, M., and Giering, R.: A coarse grid three-dimensional global inverse model of the atmospheric transport, part 1. A joint model and Jacobian Matrix, J. Geophys. Res., 104(D15), 18535–18553, 1999a.

Kaminski, T., Heimann, M., and Giering, R.: A coarse grid three-dimensional global inverse model of the atmospheric transport, part 2. Inversion of the transport of CO_2 in the 1980s, J. Geophys. Res., 104(D15), 18555–18581, 1999b.

Koster, R. D., Dirmeyer, P. A., Guo, Z., Bonan, G., Chan, E., Cox, P., Gordon, C. T., Kanae, S., Kowalczyk, E., Lawrence, D., Liu, P., Lu, C. H., Malyshev, S., McAvaney, B., Mitchell, K., Mocko, D., Oki, T., Oleson, K., Pitman, A., Sud, Y. C., Taylor, C. M., Verseghy, D., Vasic, R., Xue, Y., and Yamada, T.: Regions of Strong Coupling Between Soil Moisture and Precipitation, Science, 305, 1138–1140, 2004.

Los, S. O., Collatz, G. J., Bounoua, L., Sellers, P. J., and Tucker, C. J.: Global interannual variations in sea surface temperature and land surface vegetation, air temperature, and precipitation, J. Climate, 14, 1535–1549, 2001.

Los, S. O., North, P. R. J., Grey, W. M. F., and Barnsley, M. J.: A method to convert AVHRR Normalised Difference Vegetation Index time series to a standard viewing and illusmination geometry, Remote Sens. Environ., 99, 400–411, 2005.

Los, S. O. Weedon, G. P. North, P. R. J. Kaduk, J. D., Taylor, C. M., and Cox, P. M.: An observation-based estimate of the strength of rainfall-vegetation interactions in the Sahel, Geophys. Res. Lett., 33, L16402, doi:10.1029/2006GL027065, 2006.

Los, S. O., North, P. R. J., and Grey, W. M. F.: Fused AVHRR SeaWiFS Inetrannual Reanalysis (FASIR) 10-day (3 time mothly) data 1982–2006 (version 5.0) Earth system atlas, http://earthatlas.sr.unh.edu/maps, last accessed 22 October 2010, 2007.

Maier-Reimer, E.: Geochemical cycles in an ocean general circulation model – preindustrial tracer distributions, Global Biogeochem. Cycles, 7, 645–677, 1993.

Marland, J., Boden, T. A., Griffin, R. C., Huang, S. F., Kanciruk, P., and Nelson, T. R.: Estimates of CO_2 emissions from fossil fule burning and cement manufacturing, based on US Bureau of Mines manufacturing data, ORNL/CIAC-25, NDP-030, Carbon Dioxide Inf. Anal. Center, Oak Ridge National Laboratory, Oak Ridge, Tenn., 1989.

Miller, J. R., Russell, G. L., and Caliri, G.: Continental-scale river flow in climate models, J. Climate, 7, 914–928, 1994.

Oki, T., Nishimura, T., and Dirmeyer, P.: Assessment of annual runoff from land surface models using Total Runoff Integrating Pathways (TRIP), Met. Soc. Japan, 77, 235–255, 1999.

Ramankatty, N. and Foley, J. A.: Global Potential Vegetation Data, Technical Note: Climate, People, and Environment Program, University of Wisconsin, Madison, Wisconsin, USA, http://www.sage.wisc.edu/download/potveg/global_potveg.html, last access 22 October 2010, 1999.

Randerson, J. T., Hoffman, F. M.,Thornton, P. E. Mahowald, N. M. Lindsay, K., Lee, Y. H., Nevison, C., Doney, S. C., Bonan, G. Stockli, R., Covey, C., Running, S. W., and Fung, I. Z. Y.: Systematic assessment of terrestrial biogeochemistry in coupled climatecarbon models, Global Change Biology, 15, 2462–2484, doi:10.1111/j.1365-2486.2009.01912.x, 2009.

Sellers, P. J., Los, S. O., Tucker, C. J., Justice, C. O., Dazlich, D. A., Collatz, G. J., and Randall, D. A.: A revised land surface parameterization (SiB-2) for atmospheric GCMs. Part 2: The generation of global fields of terrestrial biophysical parameters from satellite data, J. Climate, 9, 706–737, 1996.

Stockli, R., Lawrence, D. M., Niu, G. Y., Oleson, K. W., Thornton, P. E., Yang, Z. L., Bonan, G. B., Denning, A. S., and Running, S. W.: Use of FLUXNET in the Community Land Model development, J. Geophys Res., 113, G01025, doi:10.1029/2007JG000562, 2008.

Valentini, R. and Verma, S.: Energy balance closure at FLUXNET sites, Agric. Forest Meteor., 113, 223–243, 2002.

Van den Hurk, B. J. J. M., Viterbo, P., and Los, S. O.: Impact of leaf area index seasonality on the annual land surface evaporation in a global circulation model, J. Geophys. Res.-Atmos., 108, 4191–4199, 2003.

Vermote, E. F., Justice, C. O., Descloitres, J., El Saleous, N., Roy, D. P., Ray, J., Margerin, B., and Gonzalez, L.: A SeaWiFS global monthly coarse-resolution reflectance dataset, Int J. Remote Sensing, 22, 1151–1158, 2001.

Zhangshi, Y. and Williams, T. H. L.: Obtaining spatial and temporal vegetation data from Landsat MSS and AVHRR/NOAA satellite images for a hydrological model, Photogr. Eng. Remote Sens., 63, 69–77, 1997.