Dear Jules,

we appreciate that you accompanied the review process of the manuscript. We have revised the "Code and data availability" section. And we also added the difference file from the last version.

We have added in the first paragraph of the conclusion some statements about the performance of the LPJmL4 model for the different evaluations and their importance for the overall performance. However, the individual sections also conclude to their respective evaluations.

More than in the companion paper we need to acknowledge data providers and thus we included some inevitable links.

For both papers we have created a doi for the data and LPJmL4 code. Data are available under the doi: http://doi.org/10.5880/pik.2017.009 and the code can be downloaded from the PIK's gitlab server: https://gitlab.pikpotsdam.de/lpjml/LPJmL and is available under the doi: http://doi.org/10.5880/pik.2018.002.

Yours faithfully,
Sibyll Schaphoff

Topical Editor Decision: Publish subject to minor revisions (review by editor) *(15 Jan 2018) by Julia Hargreaves*
Comments to the Author:
In the conclusion it is stated that the model results are adequate. My question is, adequate for what? We all know that no model is perfect, and indeed that no observations are perfect, and that in some cases there may be representation error making models and observations difficult to compare. The question is really what does the level of performance of the model say about what the problems the model may be successfully used to tackle (and of course the opposite - in what ways is it not useful). I think that a few comments on this topic may improve the usefulness of this paper!

Please make the code available, then describe how it may be obtained. Please also provide the DOI for the exact version of the code described in the paper.

Other materials may be uploaded to a public repository with a DOI, or added to the supplement. Weblinks are very ephemeral. Try to include as few weblinks as possible in the final version of the paper.

This paper and its companion paper are fairly epic. They could really do with clickable tables of contents, but unfortunately that is not supported within the journal formal. I am very grateful to the 4 reviewers for their carefully considered reviews.

For such long papers in GMD it is important that as many of the details are correct as possible. Readers may well dip into the papers, and make use of small details. If there are errors then this considerably reduces the utility of the paper. At the same time, it is unlikely that any of the peer review team are in a position to find these kind of errors. So, you can either check the paper very carefully yourselves at this stage, or you could also ask people who know the model well but are not authors on the paper to check the content for errors. It is up to you, but the latter method has been employed with considerable success (i.e. some important errors were found!) for other long papers in GMD.

Manuscript prepared for Geosci. Model Dev.
with version 2015/04/24 7.83 Copernicus papers of the LaTeX class copernicus.cls.
Date: 26 February 2018

LPJmL4 – a dynamic global vegetation model with managed land: Part II – Model evaluation

Sibyll Schaphoff[1], Matthias Forkel[2], Christoph Müller[1], Jürgen Knauer[3], Werner von Bloh[1], Dieter Gerten[1,4], Jonas Jägermeyr[1], Wolfgang Lucht[1,4], Anja Rammig[5], Kirsten Thonicke[1], and Katharina Waha[1,6]

[1]Potsdam Institute for Climate Impact Research, Telegraphenberg, PO Box 60 12 03, 14412 Potsdam, Germany
[2]TU Wien, Climate and Environmental Remote Sensing Group, Department of Geodesy and Geoinformation, Gusshausstraße 25-29, 1040 Wien, Austria
[3]Max Planck Institute for Biogeochemistry, Hans-Knöll-Str. 10, 07745 Jena, Germany
[4]Humboldt Universität zu Berlin, Department of Geography, Unter den Linden 6, 10099 Berlin, Germany
[5]Technical University of Munich, Germany
[6]CSIRO Agriculture & Food, 306 Carmody Rd, St Lucia QLD 4067, Australia

Correspondence to: Sibyll.Schaphoff@pik-potsdam.de

Abstract. The dynamic global vegetation model LPJmL4 is a process-based model that simulates climate and land-use change impacts on the terrestrial biosphere, agricultural production and the water and carbon cycle. Different versions of the model have been developed and applied to evaluate the role of natural and managed ecosystems in the Earth system and potential impacts of global environmental change. A comprehensive model description of the new model version, LPJmL4, is provided in a companion paper (Schaphoff et al., under Revision). Here, we provide a full picture of the model performance, going beyond standard benchmark procedures, give hints of the strengths and shortcomings of the model to identify the need of further model improvement. Specifically, we evaluate LPJmL4 against various datasets from in-situ measurement sites, satellite observations, and agricultural yield statistics. We apply a range of metrics to evaluate the quality of the model to simulate stocks and flows of carbon and water in natural and managed ecosystems at different temporal and spatial scales. We show that an advanced phenology scheme improves the simulation of seasonal fluctuations in the atmospheric CO_2 concentration while the permafrost scheme improves estimates of carbon stocks. The full LPJmL4 code including the new developments will be supplied Open Source through https://gitlab.pik-potsdam.de/lpjml/LPJmL. We hope that this will lead to new model developments and applications that improve model performance and possibly build up a new understanding of the terrestrial biosphere.

1 Introduction

The terrestrial biosphere is a central element in the Earth System, supporting ecosystem functioning and also providing food to human societies. Dynamic global vegetation models (DGVMs) have been developed and used to study the biosphere dynamics under climate and land-use change. LPJmL4 is a DGVM with managed land that has been developed to investigate potential impacts of climate change on the terrestrial biosphere including natural and managed ecosystems, and is now described in full detail in the companion paper (Schaphoff et al., under Revision). LPJmL and its predecessors have been originally benchmarked against ecosystem carbon and water fluxes and global maps of vegetation distribution (Sitch et al., 2003), against runoff (Gerten et al., 2004), agricultural yield statistics (Bondeau et al., 2007), satellite observations of fire activity (Thonicke et al., 2001, 2010), permafrost distribution and active layer thickness (Schaphoff et al., 2013), satellite observations of fraction of absorbed photosynthetically active radiation (FAPAR) and albedo (Forkel et al., 2014, 2015), and atmospheric CO_2 concentrations (Forkel et al., 2016). These previous evaluation studies focussed on single processes or components of the model. Here we present now a comprehensive multi-sectoral evaluation to demonstrate that LPJmL4 can consistently represent multiple aspects of biosphere dynamics.

LPJmL4 spans a wide range of processes (from biogeochemical to ecological aspects, from leaf-level photosynthesis to biome composition) and combines natural ecosystems, terrestrial water cycling, and managed ecosystems in one consistent framework. As such, it is increasingly applied for cross-sectoral studies such as the quantification of planetary boundaries (Steffen et al., 2015) and SDG interactions (Jägermeyr et al., 2017), and of multidimensional impacts of climate and land use change (e.g., Gerten et al., 2013; Ostberg et al., 2015; Warszawski et al., 2014; Zscheischler et al., 2014; Müller et al., 2016). With this complexity, its evaluation against historical observations along multiple dimensions is essential (Harrison et al., 2016). For such purpose, standardized benchmarking systems have been proposed (Luo et al., 2012; Kelley et al., 2013; Abramowitz, 2005) and iLAMB (https://www.ilamb.org/), the international land model benchmarking project, has been established. In the present evaluation of a broad range of fundamental features of the LPJmL4 model, we basically follow the benchmarking procedures, variables, performance metrics and diagnostic plots suggested by Luo et al. (2012), and Kelley et al. (2013). Thus the presented evaluation is going well beyond earlier evaluations of DGVMs and of LPJmL (and its predecessors) itself. We pay special attention to LPJmL4's capability to reproduce observed seasonal and interannual dynamics and patterns of key biogeochemical, hydrological and agricultural processes at various spatial scales. In doing so, we highlight the model's unique feature of representing the interaction of processes for both natural and agricultural ecosystems in a single, internally consistent framework.

2 Model benchmark

In the following we describe in detail the model benchmarking scheme employed here, which allows for a consistent evaluation of processes simulated by LPJmL4 at seasonal and annual resolution and at spatial scales from site level (using e.g. eddy-flux measurements for comparison) to global level (using e.g. remote sensing products). The evaluation spans the time period from 1901 to 2011. The benchmarking analysis also considers results from different model set-ups and previous model versions, in order to demonstrate advancements achieved with the current LPJmL4 version and the sensitivity of results to individual new modules.

2.1 Model setup and simulation experiments

As described in Schaphoff et al. (under Revision), we drive the model simulations with observation based monthly input data on daily mean temperatures from Climatic Research Unit (CRU TS version 3.23 University of East Anglia Climatic Research Unit; Harris (2015); Harris et al. (2014)), precipitation provided by the Global Precipitation Climatology Centre (GPCC Full Data Reanalysis Version 7.0, (Becker et al., 2013)). Shortwave downward radiation and net downward longwave radiation are reanalysis data from ERA-Interim (Dee et al., 2011). Monthly average wind speeds are based on the National Centers for Environmental Prediction (NCEP) re-analysis data and were regridded to CRU (NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado, USA, Kalnay et al. (1996b)). The number of wet days per month, which is used to allocate monthly precipitation data to individual days of the corresponding months, is derived synthetically as suggested by New et al. (2000). Dew point temperature is approximated from daily minimum temperature (Thonicke et al., 2010). Global annual values for atmospheric carbon dioxide concentration are taken from the Mauna Loa station (NOAA/ESRL, http://www.esrl.noaa.gov/gmd/ccgg/trends/).

The spatial resolution of all input data is 0.5° and the model simulations are conducted at this spatial resolution. All model simulations are based on a 5000 year spinup simulation after initializing all pools to zero. A second spinup simulation of 390 years is conducted in which human land use is introduced in 1700, using the data of Fader et al. (2010). In addition to the original data set description of Fader et al. (2010), sugar cane is now represented explicitly. Cropping intensity as calibrated following Fader et al. (2010) is kept static in the simulations, whereas sowing dates are computed dynamically as a function of climatic conditions until 1971, following Waha et al. (2012) and kept static afterwards. Soil texture is given by the Harmonized World Soil Database (HWSD) version 1 (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012; Nachtergaele et al., 2008) and parameterized based on the relationships between texture and hydraulic properties from Cosby et al. (1984). The river routing scheme is from the simulated Topological Network (STN-30) drainage direction map (Vorosmarty and Fekete, 2011). Reservoir parameters are taken from Biemans et al. (2011), locations are obtained from the GRanD database (Lehner et al., 2011).

We test the influence of specific processes that have been implemented or improved in LPJmL4 (specifically, permafrost, phenology, and fire) on overall model performance by conducting the following factorial experiments:

- LPJmL4-GSI-GlobFIRM: a simulation with all standard model features enabled as used in Schaphoff et al. (under Revision), i.e. with land use, permafrost dynamics, the growing season index (GSI) phenology scheme and the simplified fire model (GlobFIRM). This model experiment is the default LPJmL4 model experiment.
- LPJmL4-GSI-GlobFIRE-PNV: same, but for potential natural vegetation (PNV) to evaluate the role of managed land on global pattern and processes. This model experiments mimics the original LPJ model (i.e. without agriculture) but with improved phenology.
- LPJmL4-NOGSI-GlobFIRM: a simulation with land use, permafrost dynamics and the simplified fire model, but without the GSI phenology for testing the sole effect of the GSI phenology. Instead of the GSI phenology, here we use the original phenology model (Sitch et al., 2003) that is based on a growing-degree day approach. This experiment mimics the LPJmL 3.5 version (including the LPJ core, agriculture, and permafrost) as described in Schaphoff et al. (2013).
- LPJmL4-NOGSI-NOPERM-GlobFIRM: a simulation with land use and the simplified fire model but without permafrost and without the GSI phenology. This model experiment mimics the original LPJmL 3.0 model with the LPJ core (Sitch et al., 2003) and the agricultural modules (Bondeau et al., 2007).
- LPJmL4-GSI-SPITFIRE: a simulation setup as LPJmL4-GSI-GlobFIRM but with the process-based fire model (SPITFIRE, Thonicke et al. (2010)). This experiment is a LPJmL4 model run with an alternative fire module.

2.2 Evaluation data sets

Following Kelley et al. (2013) we compare LPJmL4 simulations against independent data for vegetation cover, atmospheric CO_2 concentrations, carbon stocks and fluxes, fractional burnt area, river discharge and FAPAR. Beyond these suggestions of Kelley et al. (2013), we extend the benchmarking system to data sets of eddy flux tower measurements of evapotranspiration and net ecosystem exchange rate (NEE). Ecosystem respiration (Re) is evaluated against both eddy-flux measurements and operational remote sensing data. Crop yields are evaluated against FAOSTAT data (FAO-AQUASTAT, 2014). For FAPAR, we use not just one but three different reference data sets to account for uncertainties from multiple satellite datasets (see Section 2.2.6). We also compare LPJmL4 results against data that are not fully independent of other models (mostly empiricial, data-driven modelling concepts), acknowledging the limitations of these data in a benchmark system.

However, this allows for assessing LPJmL4's performance in additional aspects, where fully data-based products are not available. These data comprise global gridded data sets of vegetation or aboveground biomass carbon (Carvalhais et al., 2014; Liu et al., 2015), cropping calendars (Portmann et al., 2010), global gross primary production (GPP) (Jung et al., 2011), Re (Jägermeyr et al., 2014), soil carbon (Carvalhais et al., 2014), and evapotranspiration (Jung et al., 2011).

We use both site-level and global gridded data because they provide complementary information but have different advantages for the comparison with simulated data like from LPJmL4. Site-level data are fully independent from model estimates and assumptions, but typically only represent a specific ecosystem with a certain vegetation and soil type, and a specific site history. Thus site-level data have only a limited representativeness for 0.5° grid cells. On the other hand, global gridded data of GPP (Beer et al., 2010; Jung et al., 2011) and Re (Jägermeyr et al., 2014) are available at the same scale and thus can be directly compared to simulation outputs of DGVMs. However, global gridded datasets usually rely on empirical modelling approaches and ancillary data to upscale and extrapolate site-level data to large regions. Nevertheless, specific site conditions like forest management affecting site age, biomass, and carbon fluxes can be hardly re-simulated for a large number of global sites within a DGVM. Although Kelley et al. (2013) reject the use of such datasets for model benchmarking because they depend on modelling approaches, we accept the additional use of such datasets because they prevent the scale mismatch between site-level data and global DGVM simulations.

2.2.1 Vegetation cover

We compare simulated vegetation cover to the ISLSCP II vegetation continuous fields of Defries and Hansen (2009) as suggested by Kelley et al. (2013). This data set is a gridded snapshot of vegetation cover for the years 1992/1993 from remote sensing data and distinguishes bare soil, herbaceous, and tree cover fractions aggregated to 0.5° resolution (Defries and Hansen, 2009; Kelley et al., 2013). Tree cover fractions are further distinguished into evergreen vs. deciduous and into broad-leaved vs. needle-leaved tree types, respectively. The herbaceous vegetation class includes woody vegetation that is less than 5m tall. Data uncertainties increase in regions where tree cover is <20% due to understorey vegetation and soil disturbing the signal, as well as above 80% due to signal saturation (Defries and Hansen, 2009; Kelley et al., 2013). To test if the simulated land cover of LPJmL4 performs better than a random-generated land cover distribution we compare the performance of LPJmL4 also to the random model as suggested by Kelley et al. (2013, Section 2.3.5), whereas the original dataset ISLSCP II vegetation continuous fields were randomly resampled.

2.2.2 Atmospheric CO_2 concentration

To evaluate the model's capacity to capture global-scale, intra- and interannual fluctuations of atmospheric CO_2 concentrations as driven by the uptake activity of the terrestrial biosphere, we compare simulated CO_2 concentrations with those recorded continuously at two remote measurements at Mauna Loa (MLO, 19.53°N, 155.58°W) and Point Barrow (BRW, 71.32°N, 156.60°W) (see Rödenbeck (2005) for further details on these measurements). We use monthly CO_2 concentrations from flask and continuous measurements from 1980 to 2010 for the comparison with LPJmL4 simulations. CO_2 observations were temporally smoothed and interpolated using a standard method (Thoning et al., 1989). The atmospheric transport model (TM3, Rödenbeck et al. (2003)) in Jacobian representation (Kaminski et al., 1999) simulates the global CO_2 transport using estimates of net biome production (NBP) (here simulated by LPJmL4, see Forkel et al. (2016)), estimated net ocean CO_2 fluxes from the Global Carbon Project (Le Quéré et al., 2015) and fossil fuel emissions from the Carbon Dioxide Information Analysis Center (CDIAC; Boden et al. (2013)). Atmospheric transport in TM3 is driven by wind fields of the NCEP reanalysis (Kalnay et al., 1996a) at a spatial resolution of 4° x 5°.

2.2.3 Terrestrial carbon stocks and fluxes

Model-independent reference data for carbon stocks and fluxes are available from Luyssaert et al. (2007) for various sites globally distributed. This data set comprises vegetation carbon, aboveground biomass, GPP and net primary production (NPP). GPP flux data from Luyssaert et al. (2007) are based on eddy-flux measurements and are subject to those uncertainties, reported in Luyssaert et al. (2007, Table 2). Contrastingly, NPP data are derived from direct measurements of continuous leaf-litter collection, allometry-based estimates of stem and branch NPP from basal measurements, root NPP estimates from soil cores, mini rhizotrons, or soil respiration, and destructive understorey harvest. Estimates here are subject to uncertainties, depending on the sampling methods (Luyssaert et al., 2007). Several individual sites of this data set can be located within one simulation unit of a 0.5° grid cell and we thus compare simulated values to the range of site measurements in that grid cell.

Alternatively to the site-based GPP data from Luyssaert et al. (2007), we also compare spatial patterns and grid cell specific GPP simulations to the GPP data set of Jung et al. (2011), as also suggested by Kelley et al. (2013). This global data set is based on a larger set of eddy flux tower measurements than the data set of Luyssaert et al. (2007), but uses additional satellite and climate data, and empirical modelling for extrapolation to full global coverage. Re is evaluated for the time period 2000 to 2009 directly against plot-scale FLUXNET (http://fluxnet.fluxdata.org/data/la-thuile-dataset/) measurements (ORNL DAAC, 2011), but also against large-scale Re estimates from an empirical model based on operational remote sensing data by the Moderate Resolution Imaging Spectroradiometer (MODIS) with a resolution of 1 km and 8 days (Jägermeyr et al., 2014).

In addition to GPP, R_e and NPP, we also compare simulated NEE fluxes with eddy flux tower measurements directly. We use 70 time series of estimated NEE from eddy flux tower sites that measure the exchanges of carbon and water fluxes continuously over a broad range of climate and biome types (ORNL DAAC, 2011). Nevertheless, eddy flux tower sites are not well distributed across the globe and sites in the temperate and boreal zone are better represented than the tropical zone.

For the global comparison of the soil and vegetation carbon stocks we use the data compiled by Carvalhais et al. (2014). The soil organic carbon (SOC) estimations are based on the Harmonized World Soil Database (HWSD) (Nachtergaele et al., 2008). Carvalhais et al. (2014) used an empirical model to calculate SOC stocks ($\mathrm{kg\,m^{-2}}$) from soil organic content ($\%$), layer thickness (m, here for the first 3 m), gravel content ($\mathrm{vol\%}$), and bulk density ($\mathrm{kg\,m^{-3}}$). They pointed out that regions as North America and northern Eurasia are less reliable as HWSD was work in progress at that time. The vegetation carbon data of Carvalhais et al. (2014) are based on a forest biomass map for temperate and boreal forests from microwave satellite observations (Thurner et al., 2014), a biomass map for tropical forests based on Lidar observations (Saatchi et al., 2011), and an additional estimate of grassland biomass. Uncertainties in biomass are in most regions between 30-40 $\%$ and are strongly related to uncertainties in belowground biomass. We also compare simulated aboveground biomass to the estimates of Liu et al. (2015), which is also based on satellite-based passive microwave data. This comparison requires additional assumptions on the separation of aboveground and belowground biomass in LPJmL4 simulations. Liu et al. (2015) estimates for 2000 a global aboveground biomass of 362 PgC with a 90 $\%$ confidence interval of 310–422 PgC.

2.2.4 Terrestrial water fluxes

River discharge measurements are taken from the ArcticNET and UNH/GRDC data sets for 287 gauges (Vörösmarty et al., 1996). From this data base, we only selected river gauges with catchment areas $\geq$ 10,000 $\mathrm{km^2}$ as the model setup and resolution are not suitable for comparison with smaller catchments. We also only selected river gauge records with a temporal coverage of more than 95 $\%$ of the observation period and an observation period longer than 2 years at a monthly resolution.

Evapotranspiration fluxes are taken from the FLUXNET data base and comprise 126 sites, of which we selected sites (n=99) with at least 3 years of data available (ORNL DAAC, 2011). Additional to site-level data, we used global gridded ET data from Jung et al. (2011), which is based on an upscaling of site-level eddy covariance observations with satellite and climate data using a machine learning approach.

Irrigation withdrawal and consumption data we compare to are from other modelling approaches. Nonetheless, human water use for irrigation is an important component in the terrestrial water cycle and we discuss modelled LPJmL4 estimates in comparison to other model-based estimates, acknowledging the limitation of this comparison and addressing different sources of uncertainty.

2.2.5 Permafrost

For the evaluation of simulated permafrost dynamics, we use the measured thaw depth data from 131 stations of the Circumpolar Active Layer Monitoring (CALM) station data set: https://www2.gwu.edu/ calm/

(Brown et al., 2000) as well as the International Permafrost Association (IPA) Circum–Arctic Map of Permafrost http://nsidc.org/data/ggd318 (Brown et al., 1998). The distribution of permafrost is based on regional elevation, physiography and surface geology. The permafrost extent represents four classes which categorize the percentage of the ground underlain by permafrost (continuous, 90-100 %; discontinuous, 50-90 %; sporadic, 10-50 %; and isolated patches of permafrost, 0-10 %).

2.2.6 Fractional area burnt

For the evaluation of simulated fire dynamics, we employ data on fractional area burnt from the Global Fire Emissions Database GFED4, Version 4 (GFED4) data set (http://www.globalfiredata.org/; Giglio et al. (2013)) for the period 1995 to 2014 and climate change initiative (CCI) Fire Version 4.1 (http://cci.esa.int/data; Chuvieco et al. (2016)) for the period 2005 to 2011. Mean annual burned area was computed for both datasets for the overlapping period (2005-2011). Both data sets are derived from satellite data. Active fire data was used in GFED4, to prolong the dataset prior to the MODIS period (i.e. for 1995-2000).

2.2.7 Fraction of absorbed photosynthetic active radiation and albedo

Data on the fraction of absorbed ~~photosnthetically~~ photosynthetically active radiation (FAPAR) are derived from three different satellite data sets to account for differences between datasets for model evaluation (see Table 4, Forkel et al. (2015)). The MODIS (~~Moderate-Resolution Imaging Spectroradiometer;~~ USGS, 2001) FAPAR (Knyazikhin et al., 1999), the Geoland2 BioPar (GEOV1) FAPAR data set (Baret et al., 2013) (hereafter called VGT2 FAPAR), and the GIMMS3g FAPAR data set (Zhu et al., 2013). The MODIS FAPAR data set is taken from the MOD15A2 product with a temporal resolution of 8 days at a spatial resolution of 1 km, covering the period 2001 to 2011. VGT2 is based on SPOT VGT with a temporal resolution of 10 days and 0.05° spatial resolution (Baret et al., 2013), covering the period 2003 to 2011. The GIMMS3g data set has a 15-day temporal resolution and 1/12° spatial resolution and covers the period from 1982 to 2011. Data on FAPAR is also subject to uncertainties from the processing of the remotely sensed data and is not available continuously for all areas. We compare the spatial patterns of the peak FAPAR, and the temporal dynamics of FAPAR in each grid cell, and seasonal variations in FAPAR averaged for Köppen-Geiger climate zones for the three different FAPAR data sets. The aggregated FAPAR represents the average monthly time series for all grid cells that belong to a certain Köppen-Geiger climate zone (see also Forkel et al. (2015)). For the Köppen-Geiger climate zones, FAPAR time series are averaged over all grid cells that belong to ~~that~~ the same Köppen-Geiger climate zone (see also Forkel et al. (2015)).

For the evaluation of the reflectance of the earth surface we used the MODIS C5 albedo time series data set from 2000-2010 (Lucht et al., 2000; Schaaf et al., 2002), that we also aggregated to Köppen-Geiger climate zones for the evaluation here.

2.2.8 Agricultural productivity

Detailed data on crop growth and productivity are available for individual sentinel sites (Rosenzweig et al., 2014). For global-scale or regional simulations, reference data are available only for crop yields and in (sub-)national aggregations (e.g., FAO-AQUASTAT, 2014) or as processed and interpolated gridded products (Iizumi et al., 2014). In all yield data statistics outside of well-controlled field experiments, yield levels and interannual variability are not only affected by variability in weather, but also by variance in management conditions, such as sowing dates, variety choices, cropping areas, fertilizer inputs, pest control and others (Schauberger et al., 2016). Consequently, it is difficult to evaluate model performance from a comparison of simulated yields with static assumptions on most management aspects with yield statistics in which the contribution of weather variability on yield variability is unknown. Müller et al. (2017) propose a combination of global gridded crop model simulations and different observation-based yield data sets to establish a benchmark for global crop model evaluation. Generally, global gridded crop models perform well in most regions for which statistical models can detect significant influence of weather on crop yield variability (Ray et al., 2015). We here evaluate LPJmL4 by comparing simulated and observed yield variability of the 10 top-producing countries of the respective crop (FAO-AQUASTAT, 2014). We refrain from comparing to individual sentinel sites, but refer to the evaluation of LPJmL crop simulations at global, national and grid cell scale in the global gridded crop model evaluation framework (Müller et al., 2017). As in Müller et al. (2017), we aggregate simulated grid-cell level yield time series to average national yield time series using the MIRCA2000 data set for spatial aggregation (Porwollik et al., 2016) and removing trends in observations and simulations with a moving window average (see Müller et al. (2017) for details).

The productivity of biomass plantations is evaluated with data from experimental sites for miscanthus, switchgrass, poplar, willow and eucalyptus production, using the data collection of Heck et al. (2016). Data on biomass productivity typically report a data range. These are site-specific management differences and reflect the diverse drivers of reported productivity, such as variation of plant species, fertiliser use and irrigation management, crop spacing or sapling size. We average the minimum and maximum values to derive the mean productivity per site.

2.2.9 Sowing dates

For evaluating the accuracy of the simulated rainfed sowing dates, we use the global data set of growing areas and growing periods, MIRCA2000 (Portmann et al., 2008, 2010) at a spatial resolution of 0.5° and a temporal resolution of one month, as proposed by Waha et al. (2012). Monthly data in MIRCA2000 were converted to daily data by assuming that the growing period starts at the first day of the month following Portmann et al. (2010). MIRCA2000 reports several growing periods in a year for some administrative units ~~and~~ for the crops wheat, rapeseed, rice, cassava and maize. For comparison we select the best corresponding growing period so that a close agreement indicates that simulated sowing dates are reasonable, but not necessarily the most frequently chosen by farmers. We do not compare simulated sowing dates for sugar cane (see SI-Fig. S94) to observed sowing dates as MIRCA2000 assumes it is grown all year round as a perennial crop.

2.3 Evaluation metrics

We employ Taylor diagrams ~~Taylor (2001)~~ (Taylor, 2001) to compare the correlation, differences in standard deviation, and the centered root mean squared error (CRMS) between simulated and observed carbon and water fluxes at FLUXNET sites (ORNL DAAC, 2011) and at gauge stations from ArcticNET and UNH/GRDC. The standard deviations of the reference data sets have been normalized to 1.0 so that multiple sites can be displayed in one figure.

Table 1. Evaluation metrics used in this study

Metric	Equation	Reference
NMSE	$\mathrm{NMSE} = \frac{\sum_{i=1}^{N}(y_i - x_i)^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$	Kelley et al. (2013)
NME	$\mathrm{NME} = \frac{\sum_{i=1}^{N}\lvert y_i - x_i\rvert}{\sum_{i=1}^{N}\lvert x_i - \overline{x}\rvert}$	Kelley et al. (2013)
ME	$\mathrm{ME} = \frac{\sum_{i=1}^{N}\lvert y_i - x_i\rvert \cdot A_i}{\sum_{i=1}^{N} A_i}$	
W	$\mathrm{W} = 1 - \frac{\sum_{i=1}^{N}(y_i - x_i)^2 \cdot A_i}{\sum_{i=1}^{N}(\lvert y_i - \overline{x}\rvert + \lvert x_i - \overline{x}\rvert)^2 \cdot A_i}$	Willmott (1982)
MM	$\mathrm{MM} = \frac{\sum_{i=1}^{N}\lvert q_{i,j} - p_{i,j}\rvert}{N}$	Kelley et al. (2013)

y_i is the simulated and x_i the observed value in grid cell i, $\overline{x}$ the mean observed value, A_i the area weight in grid cell i, and N the number of grid cells or sites, $q_{i,j}$ is the simulated and $p_{i,j}$ is the observed fraction of item j in grid cell i. Normalized mean square error – NMSE, Normalized mean error – NME, ME – Mean absolute error, W – Willmott coefficient of agreement, MM – Manhattan metric

For global gridded reference data sets, such as for carbon stocks, we show spatial patterns in maps and aggregations as latitudinal means and quantify overall differences as a spatial correlation analysis over all grid cells (see Table 4). As suggested by Kelley et al. (2013) we use the normalized mean ~~squared~~ square error (NMSE) to describe differences between model simulation and reference data sets. The NMSE is zero for perfect agreement, 1.0 if the model is as good as using the data mean as predictor and larger 1.0 if the model performs less well than that. The squared error term puts stronger emphasis on large deviations between simulations and observations and is thus stricter than the normalized mean error (see Table 1 for equations). Kelley et al. (2013) also suggests to use the Normalized mean error (NME) as a more robust metric than NMSE. NME is based on absolute residuals (NMSE on squared residuals) and thus is especially better suited for variables that can have very large values and residuals. Additionally, we use the Manhattan metric (MM) proposed by Kelley et al. (2013) for evaluation of vegetation cover. Values for MM less than 1 reflect that the model ~~perform~~ performs better than the mean value~~and additionally~~. Additionally we show the random model~~- "produced~~, which was generated by bootstrap resampling of the observations ~~" -~~ as proposed by Kelley et al. (2013, Table 4)~~for~~. The random model was used for the evaluation of vegetation distribution.

Table 2 gives an overview of variables evaluated at the local scale and ~~which measures are~~ the measures that were used for the evaluation of time series for crop yields~~, we~~. We employ a simple time series correlation analysis after removing trends with a moving-window detrending method. For comparison with point measurements, we extract the time series from corresponding 0.5° grid cells. These simulated time series may differ in terms of weather and soil conditions from the actual site as the simulations are based on gridded global data set inputs. Time period is given by the respective measurements, which differ for each observation point.

Table 2. Overview of variables ~~evaluating LPJmL4, showing measures~~ and ~~references at~~ measures used for the evaluation of LPJmL4 local scale.

	Measure				Reference	
Variable	CRMSE	Standard Deviation	Correlation	Reference to figures	Data	Citation
CO_2			x	Fig. 1 & 2	Atmospheric transport	Rödenbeck (2005)
NEE	x	x	x	Fig. 3	FLUXNET	ORNL DAAC (2011)
ET	x	x	x	Fig. 7	FLUXNET	ORNL DAAC (2011)
NPP			x	Fig. 4d		Luyssaert et al. (2007)
GPP			x	Fig. 4c		Luyssaert et al. (2007)
BIOMASS			x	Fig. 4a & 4b		Luyssaert et al. (2007)
DISCHARGE	x	x	x	Fig. 8 & SI-Fig. S19-S66	ArcticNET & UNH/GRDC	Vörösmarty et al. (1996)

Centered root mean square error (CRMSE)

To envisage the degree of agreement between simulated (LPJmL4) and observed (MIRCA2000) sowing dates, we follow Waha et al. (2012) and compute two different metrics: the Willmott coefficient of agreement (W) (Willmott, 1982) and the mean absolute error (ME), both weighted by the crop-specific cultivated area according to (Portmann et al., 2010). For an overview of all metrics used, see Table 1.

3 Results and discussion

In the following we compare the standard version LPJmL4, which refers to the experiment LPJmL4-GSI-GlobFIRM. In case of the other experiments we refer to the names defined in Section 2.1.

3.1 Vegetation cover

LPJmL4 reproduces the observed vegetation distribution better than the random model (Table 3). Such as the random model, LPJmL4 can best reproduce the distinction between bare soil and vegetated areas (MM = 0.22) and between tree-covered areas and areas without trees (MM = 0.31), but with considerably better scores than the random model (MM = 0.56 and 0.54 respectively). Moreover, LPJmL4 simulation results reach the lowest MM scores for the distinction of evergreen vs. deciduous trees (MM = 0.52) and for the distribution and composition of life forms (trees vs. herbaceous vs. bare soil; MM = 0.45), these are substantially better than the random model (MM = 0.87 and 0.88 respectively). The largest improvement of LPJmL4 simulations over the random model are found for the patterns of broadleaved vs. needleleaved trees (MM = 0.37 for LPJmL4 vs. 0.94 for the random model, see Table 3).

Table 3. Comparison metric scores for LPJmL4 simulations against observations of fractional vegetation cover data from International Satellite Land-Surface Climatology Project (ISLSCP) II vegetation continuous field (VCF) (Defries and Hansen, 2009).

Vegetation cover	Manhattan Metric (MM)	
	LPJmL4	Random model*
Life forms	0.45	0.88
Tree vs. non-tree	0.31	0.54
Herb vs. non-herb	0.42	0.66
Bare vs. covered ground	0.22	0.56
Evergreen vs. deciduous	0.52	0.87
Broadleaf vs. needleleaf	0.37	0.94

MM suggested by Kelley et al. (2013),* values taken from Kelley et al. (2013, Table. 4)

3.2 Atmospheric CO_2 concentration and NEE

3.2.1 Comparison of simulated NBP to atmospheric CO_2 concentration at MLO and BRW

LPJmL4 reproduces well observed long-term and seasonal dynamics of atmospheric CO_2 (Fig. 1 and 2). The long-term trend of atmospheric CO_2 is well reproduced in all the different model setups (Fig. 1), except for the setup with natural vegetation only (LPJmL4-GSI-GlobFIRM-PNV). The experiment with all processes included (LPJmL4-GSI-GlobFIRM) gives the best correlation and trend reproduction, which suggests that an integral representation of the LPJmL4's features is required to match observations best. Next to land-use dynamics, the inclusion of permafrost dynamics has the strongest effects on the simulated trend (LPJmL4-NOGSI-NOPERRM-GlobFIRM vs. LPJmL4-NOGSI-GlobFIRM). The use of the process-based fire model SPITFIRE leads to small overestimation of the trend in atmospheric CO_2 concentrations compared to the other model setups, especially at MLO. Seasonal variations in atmospheric CO_2 can be well reproduced by LPJmL4, especially by the standard setup (LPJmL4-GSI-GlobFIRM) (Fig. 2). The simulation of seasonal variations in atmospheric CO_2 content are especially improved by the GSI phenology scheme (LPJmL4-NOGSI-GlobFIRM vs. LPJmL4-GSI-GlobFIRM, Fig. 2 ~~(a)(b)~~top panel). All model setups (except LPJmL4-GSI-SPITFIRE) can reproduce the observed strong significant increase in the seasonal CO_2 amplitude at BRW and the weak (~~but~~ and insignificant) increase at MLO (Fig. 2 ~~(c)~~bottom panel). These results are in agreement with a previous evaluation of simulated seasonal CO_2 changes in LPJmL (Forkel et al., 2016).

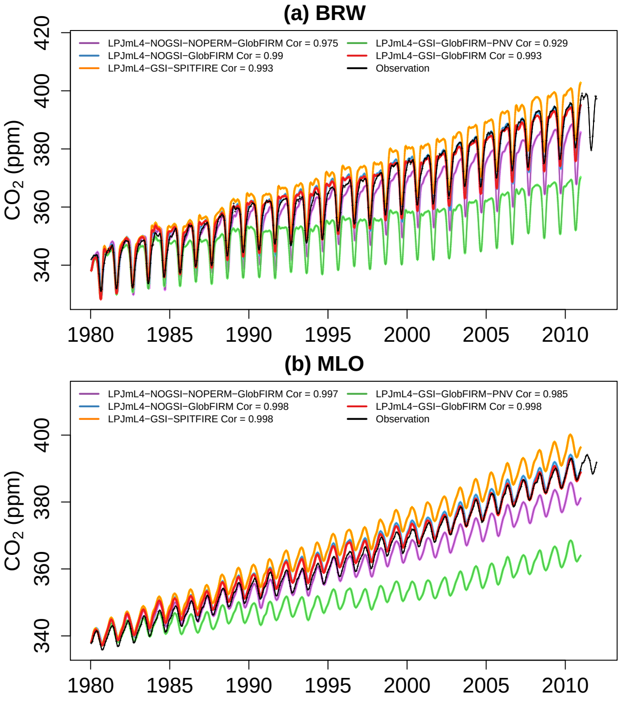


Figure 1. Comparison of the atmospheric CO_2 concentrations at ~~Mauna Loa~~ Point Barrow (~~MLO~~BRW) at the top and ~~Point Barrow~~ Mauna Loa (~~BRW~~MLO) at the bottom for the different LPJmL4 experiments.

Further analysis shows that the standard setup (LPJmL4-GSI-GlobFIRM) can best produce the mean seasonal cycle in MLO, whereas the version that omits land use (LPJmL4-GSI-GlobFIRM-PNV) performs slightly better than this in BRW (Fig. 2). The standard setup (LPJmL4-GSI-GlobFIRM) can also best reproduce the increase in the seasonal amplitude at BRW, whereas it is the only setup that produces a statistically significant but still very small increase in the seasonal amplitude at MLO, where also observations do not show a statistically significant increase.

3.2.2 Comparison of simulated NEE to eddy-flux measurements

We evaluate model performance of simulated NEE from LPJmL4 for temporal and spatial variation of NEE data from eddy flux measurements, using Taylor diagrams (Taylor, 2001). Stations are sorted

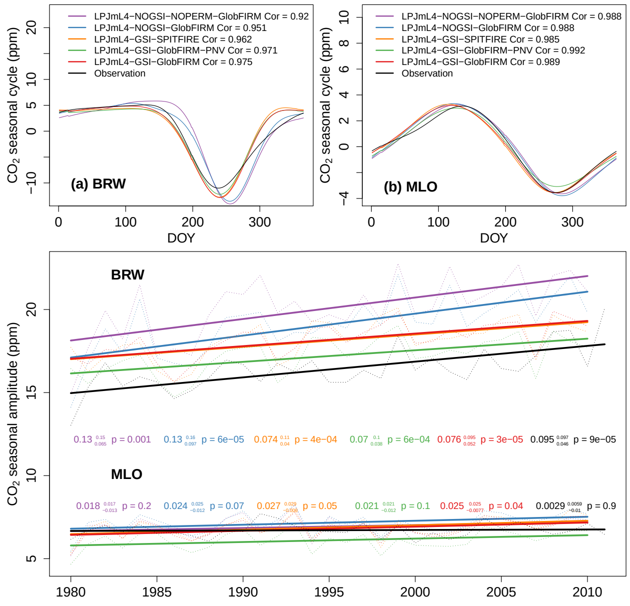


Figure 2. Comparison of the atmospheric CO_2 concentration at Mauna Loa (MLO) and Point Barrow (BRW) simulated in the different LPJmL4 experiments. Top panel, seasonal cycle; bottom panel, trend of the seasonal amplitude, ~~slope~~ slopes are given for the different LPJmL4 experiments.

from North to South (see Fig. 3) for all NEE measurements available for >3 years. The model is able to reproduce the mid-latitudes best (represented by yellow over green to light blue colors), with correlation coefficients mostly between 0.4 and 0.9 and standard deviations often within +/-30 % of the reference data. The northernmost regions are well reproduced at some flux towers, but often with higher standard deviation than in the flux tower data, which means that the simulated time series are largely in phase with but ~~are~~ more variable than the observations. In contrast, the evaluation is comparatively poor for tropical regions, especially the station at Santarém with strong negative correlations (r< -0.6) but realistic standard deviations. For this site, however, Saleska et al. (2003) have already pointed out that the eddy-flux measurements show the opposite sign compared to tree growth observations and model predictions, which also is the case for LPJmL4. We stress that this evaluation is done for a standard LPJmL4 run and standard input (the LPJmL4-GSI-GlobFIRM as described in Schaphoff et al. (under Revision)), i.e. we did not calibrate the model to site-specific conditions and also drive the model with gridded input data rather than the observed soil and weather data at individual stations. More detail for comparisons with eddy-flux tower measurements for individual locations is supplied in the supplementary material (see SI-Fig. S1-S7). Additionally we have simulated NEE by conducting simulations with station-specific meteorological observations (see SI-Fig. S17). It shows that results are similar to simulations driven by global climate data.

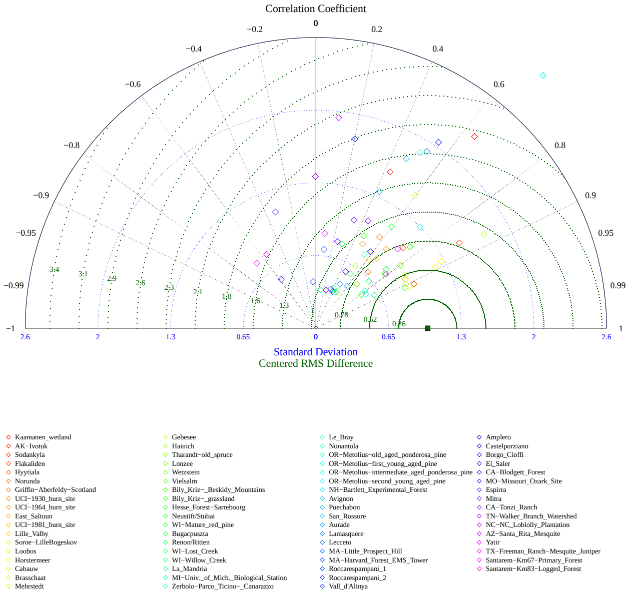


Figure 3. Net ecosystem exchange rate measured at eddy flux towers: ORNL DAAC (2011). Available online FLUXNET. Sites (colours) are ordered from north to south.

3.3 Vegetation and soil carbon stocks and vegetation productivity

3.3.1 Soil carbon and vegetation carbon stocks

The spatial correlation between simulated and observation-based estimates of SOC by Carvalhais et al. (2014) is weak (r = 0.29, Table 4) with disagreements in the sub-tropics, where LPJmL4 simulations substantially underestimate soil carbon stocks, whereas LPJmL4 report much higher soil carbon in the high northern latitudes (>50°N) and lower values for the tropical and temperate zone, compared to Carvalhais et al. (2014) (see SI-Fig. 67). Other estimates by Tarnocai et al. (2009) show much higher carbon content for the permafrost affected areas than the data set of Carvalhais et al. (2014). We thus assume that the disagreement between simulations and the Carvalhais et al. (2014) data may also result from an underestimation of carbon stocks in the Carvalhais et al. (2014) data. ~~Although that~~ However, the estimation of global soil carbon is less in LPJmL4 (1869 PgC) than estimated by Carvalhais et al. (2014) (2352±400 PgC).

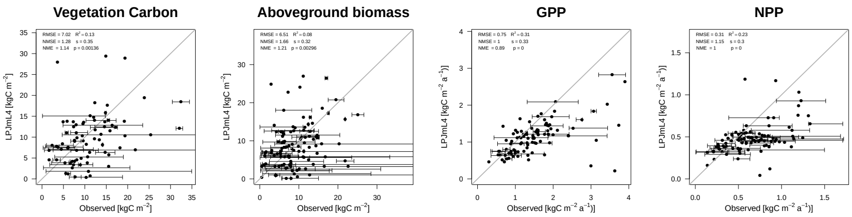
The comparison of simulated and observation-based assessments of vegetation carbon show a good spatial correlation (r = 0.84, Table 4). Globally Carvalhais et al. (2014) estimates slightly lower

Table 4. Overview of variables evaluating LPJmL4, showing measures and references at the global scale.

	Measure					Reference	
Variable	NME	NMSE	spatial Correlation	temporal Correlation	Visual Comparison	Data	Citation
GPP - Av	0.20	0.13	0.87		Fig. 5& SI-Fig. S68	GPP	Jung et al. (2011)
R_e - Av	0.67	0.55	0.67		Fig. 6 & SI-Fig. S70		Jägermeyr et al. (2014)
SoilC - Av	0.48	0.75	0.29		SI-Fig. S67	Soil carbon stocks	Carvalhais et al. (2014)
VegC - Av	0.33	0.36	0.84		SI-Fig. S69 (a)	Total Biomass	Carvalhais et al. (2014)
					SI-Fig. S69 (b)	AGB	Liu et al. (2015)
FAPAR - I-aMv	0.17	0.13	0.63	Fig. 10a		MODIS FAPAR	Knyazikhin et al. (1999)
FAPAR - I-aMv	0.18	0.15	0.59	Fig. 10b		GIMMS3g FAPAR	Zhu et al. (2013)
FAPAR - I-aMv	0.21	0.20	0.69	Fig. 10c		VGT2 FAPAR	Baret et al. (2013)
ET	1E-6	0.07	0.84		SI-Fig. S71	Latent heat flux	Jung et al. (2011)
fBA					SI-Fig. S72		GFED4 & CCI Fire (4.1)
Albedo					SI-Fig. S72	MODIS C5	Lucht et al. (2000)
Discharge						ArcticNET &	Vörösmarty et al. (1996)
Ov	0.42	0.24		$R^2 = 0.90$		UNH/GRDC	
Mav	0.36	0.19		$R^2 = 0.92$			
I-av	0.24	0.06		$R^2 = 0.97$			

Normalised mean error (NME) and Normalised mean square error (NMSE) as suggested by Kelley et al. (2013); Av – Annual average; I-aMv – Inter-annual-monthly variability; Overall variability – Ov; Monthly average variability – Mav; Inter-annual variability – I-av; Vegetation carbon – VegC; Aboveground biomass – AGB; Soil carbon – SoilC; fBA – fractional burnt area.

biomass (445±8 PgC) as simulated by LPJmL4 (507 PgC). The spatial patterns of vegetation carbon stocks are shown in SI-Fig. S69 (a) for simulations and the data product of Carvalhais et al. (2014). While the broad geographical patterns are in overall agreement with the evaluation data, the absolute values differ in some regions. Specifically, LPJmL4 simulates much higher biomass (see the latitudinal pattern of SI-Fig. S69) for the tropics, and lower biomass between 20 and 40 degrees on the northern and southern hemisphere, where Carvalhais et al. (2014) show higher values compared to LPJmL4. This is probably due to an overestimation of vegetation carbon in agricultural regions by Carvalhais et al. (2014) as Liu et al. (2015) shows similar aboveground biomass estimates there (see SI-Fig. S69 (b)). The sub-tropical region, where biomass carbon is underestimated, corresponds also to the region where LPJmL4 simulations underestimate soil carbon stocks compared to Carvalhais et al. (2014). Also the comparison of aboveground biomass estimates with the data set of Liu et al. (2015) shows a similar spatial pattern of overestimation of vegetation biomass with too high values in boreal and tropical areas. The comparison is complicated by uncertainties in the estimation of belowground biomass (Saatchi et al., 2011) and the assumed distribution between aboveground and belowground biomass in LPJmL4 simulations, where LPJmL4 assumes that belowground biomass consists of all fine root biomass and one third of all sapwood biomass. The simulation experiments without permafrost dynamics (LPJmL4-NOGSI-NOPERM-GlobFIRM) show a high overestimation of biomass in the high latitudes. Similarly, the inclusion of the GSI phenology substantially reduces the biomass overestimation in comparison to Carvalhais et al. (2014) and Liu et al. (2015), which is consistent with the finding of Forkel et al. (2014). The consideration of human land use in the simulations improves carbon stock simulations in the temperate zones (SI-Fig. S69). This clearly demonstrates the importance of permafrost, human land use and the GSI phenology for the simulation of the terrestrial carbon cycle, even though the remaining discrepancies warrant further model improvement.



(a) (b) (c) (d)

Figure 4. Evaluation of Vegetation carbon (a), aboveground biomass (b), GPP (c), and NPP (d). Observed data are provided by Luyssaert et al. (2007). Bars give the minimum and maximum of the estimation within one 0.5° cell simulated by LPJmL4.

Fig. 4a and 4b compares site data estimation with the representative LPJmL4 grid cell estimation, with an uncertainty range, which comes from the different measurements within one 0.5° grid cell. Both vegetation and aboveground carbon ~~show a slight overestimation of some simulated values, but also some strong underestimation~~ are slightly overestimated in some cases but also strongly underestimated in others. As LPJmL4 calculates a representative mean value of a 0.5° grid cell for all benchmarks, the simulated values should match to the mean values. However, it can be assumed that measurements are not evenly distributed through the age classes within one grid cell or forest and it remains unclear how representative the measurements are for a 0.5° grid cell area.

3.3.2 Gross and net primary production (GPP and NPP)

The global estimation of 123.7 $\mathrm{PgC\,a^{-1}}$ GPP from LPJmL4 (see Fig. 5) matches the estimates from Beer et al. (2010); Jung et al. (2011) of 123±8 resp. 119±6 $\mathrm{PgC\,a^{-1}}$ for the years 1982-2005, whereas the highest divergence can be observed in the tropics, where LPJmL4 estimates much lower values despite the higher biomass estimations (see Section 3.3). LPJmL4 simulated higher GPP for the temperate and boreal zones than reported by Jung et al. (2011). The different model experiments show similar pattern except for LPJmL4-GSI-GlobFIRM-PNV, which shows lower GPP in the Mediterranean (see Fig. 5). Carvalhais et al. (2014) estimates global NPP at 54±10 $\mathrm{PgC\,a^{-1}}$ and LPJmL4 at 57 $\mathrm{PgC\,a^{-1}}$ for the mean of the years 1982-2011.

The site data comparison to Luyssaert et al. (2007) shows a good agreement between site measurements and simulated GPP (see Fig. 4c) and NPP (see Fig. 4d). The overestimation of simulated biomass and the good agreement of NPP and GPP leads to the conclusion that LPJmL4 underesti-

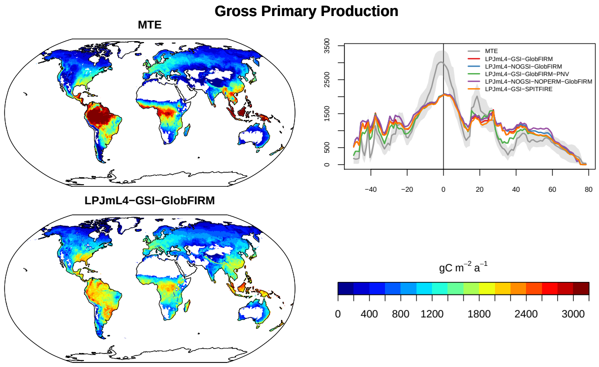


Figure 5. The maps (left side) show the spatial pattern of gross primary production (GPP, $[\mathrm{gC\,m^{-2}\,a^{-1}}]$) distribution from the standard LPJmL4 simulation against the MTE data (Jung et al., 2011). The graph on the right side shows the latitudinal pattern of ~~evapotranspiration~~ GPP distribution simulated by the different versions of LPJmL4 against data from Jung et al. (2011).

mates mortality. This warrants further investigation why LPJmL4 seems to overestimate global GPP but shows good agreement with site data. The comparison of LPJmL4 against MTE data (Jung et al., 2011) on the local scale for the same points as given by Luyssaert et al. (2007) show a good agreement, especially if outliers are excluded (SI-Fig. S68(b,c). SI-Fig. S68a compares plot data against the global data.

3.3.3 Ecosystem respiration (R_e)

Comparison of satellite-derived ecosystem respiration with those simulated by LPJmL4 reveals similar spatial patterns (Fig. 6 and SI-Fig. S70). However, LPJmL4 shows higher temperature sensitivities (Fig. 6 (a)) and consistently simulates higher R_e values in high-latitude and subtropical regions (SI-Fig. S70). Since satellite-derived ecosystem respiration is calibrated for FLUXNET data and hence exhibits marginal cross-latitude bias, the discrepancies to LPJmL4 are likely associated either with LPJmL4 parameterization or with systematic errors in the FLUXNET ~~sampling~~ processing technique. Additional details and figures are presented in Jägermeyr et al. (2014).

3.4 Water fluxes

3.4.1 Evapotranspiration

The spatial distribution of evapotranspiration of LPJmL4 shows a very similar pattern as estimated by Jung et al. (2011) (Table 4, SI-Fig. S71). It indicates a general underestimation of ET, especially in the tropics and subtropics, but in most cases within the uncertainty range. This is consistent with the underestimation of GPP in the tropics (Fig. 5), but not with the general overestimation of veg-

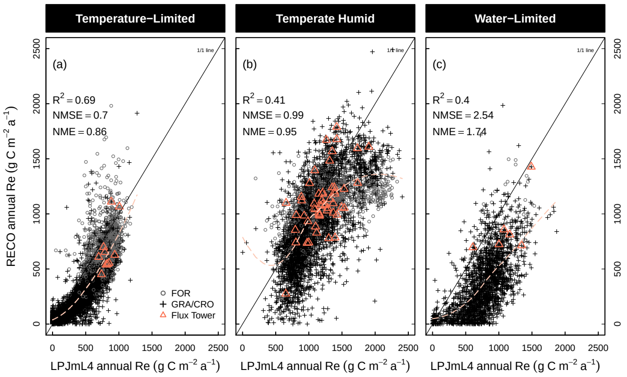


Figure 6. Ecosystem respiration (Re) evaluation of standard LPJmL4 simulations with satellite-derived estimations from (Jägermeyr et al., 2014). Compared are annual Re sums for all pixels from the displayed extent in SI-Fig. S70, separated by climate type (a)–(c). Dashed lines indicate a polynomial bias curve. Chart symbols are separated for forest (FOR) and grassland/cropland (GRA/CRO) land cover classes.

etation biomass (SI-Fig. S69). The different experiments show nearly no effects on the simulated evapotranspiration.

At site level, the evapotranspiration fluxes show a good agreement with eddy-flux tower measurements ~~in~~ (Fig. 7). LPJmL4 shows good performance in most regions, with correlation coefficients often larger than 0.6. Especially the northern and temperate stations (red to light blue symbols) show high correlation with low CRMS. Simulations of tropical and subtropical ET (dark blue to purple symbols) show weak or even negative correlations coupled with a high CRMS for some stations. We also provide more detailed time series analyses for the evapotranspiration fluxes of individual sites in the supplementary material (SI-Fig. S8-S16).

3.4.2 River discharge stations evaluation

Discharge simulated by earlier LPJmL versions was evaluated before in several studies, also in comparison with other global hydrologial and land surface models (Haddeland et al., 2011). River discharge was evaluated for major catchments globally, also accounting for effects of different precipitation datasets (Biemans et al., 2009) and regionally for the Amazon basin (Langerwisch et al., 2013) and the Ganges (Siderius et al., 2013).

Fig. 8 shows the comparison of simulated LPJmL4 and observed river discharge values for all gauges with basin area $\geq$ 10,000 km^2. Here, the most northern (blue) and also most southern (purple) gauges show good agreement, but overall the picture is mixed with respect to correlation coefficients and standard deviation. For further insights, we provide comparisons for all considered gauges in the supplementary material (SI-Fig. S19-S66). For many gauges, the simulated seasonal timing of river discharge (peaks) has improved (see SI-Fig. S19-S22) compared to the previous model evaluation of river discharge (Schaphoff et al., 2013), which is mainly a result of the newly implemented GSI-

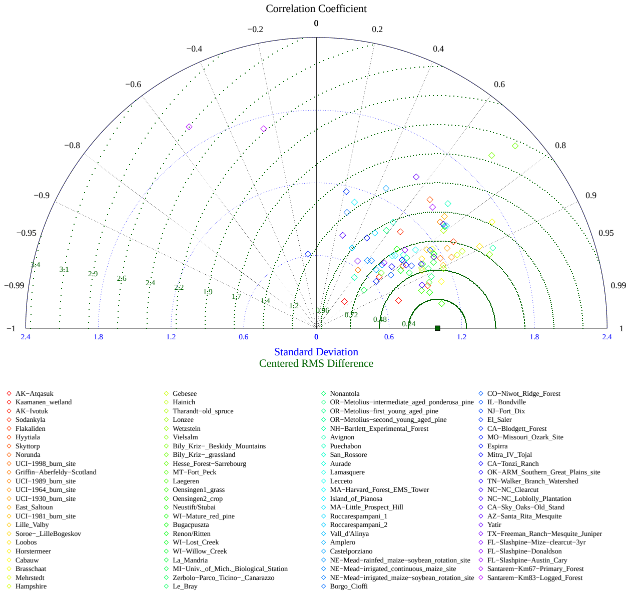


Figure 7. Evaporation rate measured at eddy flux towers: ORNL DAAC (2011). Available online FLUXNET. Site locations are ordered from north to south.

phenology scheme (Forkel et al., 2014). Especially, the discharge spring peaks in permafrost areas are affected by this improvement. At many gauges, LPJmL4 can reproduce the variability for the whole time series and specially the seasonality, with a high R^2 and a NME/NMSE, which implies 495 a better performance than the mean model. The dynamics at gauges in temperate zone (SI-Fig. S49-S50, S61) are not well reproduced in the simulations and also the NME/NMSE show high values in contrast to gauges in the subtropics and tropics (SI-Fig. S64-S66), which typically show high R^2 and low NME/NMSE.

The evaluation at the global aggregation (computed for all stations and than averaged) shows 500 very high agreement between observed and modelled discharge (see Table 4). Both the explained variance (R^2) and the NME/NMSE contribute to the good performance of the simulated discharge. The constant flow velocity in all rivers, as assumed in LPJmL4 simulations, could be varied by river for further model improvement, especially for the timing in flat areas where wetland dynamics may play an important role.

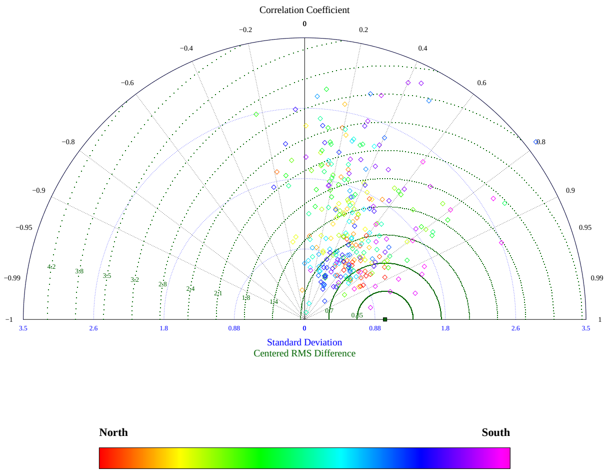


Figure 8. Comparison of simulated discharge with 287 gauges provided by ArcticNET and UNH/GRDC. Stations with basin area $\geq 10{,}000\,\mathrm{km}^2$ are taken into account. Gauges are ordered from north to south (see legend color).

3.4.3 Irrigation withdrawal and consumption

Global estimates of irrigation water withdrawal (W_d: 2545 $\mathrm{km}^3\mathrm{a}^{-1}$) and consumption ($W_c$: 1292 $\mathrm{km}^3\mathrm{a}^{-1}$) agree well with previous studies. Reported W_d values for the period 1998-2012 are 2722 $\mathrm{km}^3\mathrm{a}^{-1}$ (FAO-AQUASTAT, 2014), and modelling results range from 2217 to 3185 $\mathrm{km}^3\mathrm{a}^{-1}$ (Döll et al., 2014; Wada and Bierkens, 2014; Döll et al., 2012; Alexandratos and Bruinsma, 2012; Wada et al., 2011; Siebert and Döll, 2010). W_c estimations range between 927 and 1530 $\mathrm{km}^3\mathrm{a}^{-1}$ (Chaturvedi et al., 2015; Döll et al., 2014; Hoff et al., 2010). Döll et al. (2012) finds that 1179 $\mathrm{km}^3\mathrm{a}^{-1}$ (1098 $\mathrm{km}^3\mathrm{a}^{-1}$ in Wada and Bierkens (2014)) relate to surface water and additional 257 $\mathrm{km}^3\mathrm{a}^{-1}$ from groundwater resources. LPJmL4 does not account for fossil groundwater extraction nor desalination. However, previous studies show that 80% of groundwater withdrawals are recharged by return flows (Döll et al., 2012). It is thus plausible that studies accounting for (fossil) groundwater reach W_d estimates somewhat higher than in LPJmL4. Naturally, irrigation water estimates are associated with uncertainties in the precipitation input employed (Biemans et al., 2009). A representation of multiple cropping systems in LPJmL4 (Waha et al., 2013) and corresponding growing seasons (Waha et al., 2012) could also help to improve water withdrawal and consumption estimates and eventually river discharge, especially in tropical areas.

Simulated irrigation efficiencies are difficult to compare with observations due to inhomogeneous definitions and field measurement problems. Yet, in SI-Table S1 we relate our results to comparable literature. Our simulations meet indicative estimates of Brouwer et al. (1989) at global level. Sauer et al. (2010) provide another independent estimate of field efficiency with global average values of 42%, 78%, and 89% for the three irrigation types, respectively. Our estimates agree well with these numbers globally and regionally, even though there are some regional patterns that are not represented in our results. Sauer et al. (2010), for instance, find lower surface irrigation efficiencies in Middle East, North Africa (MENA) and sub-Saharan Africa (SSA). We simulate above-average efficiencies in MENA and particularly low ones in South Asia, which is both supported by Rosegrant et al. (2002) and Döll and Siebert (2002). Overall, the evaluation of the irrigation model in LPJmL4 demonstrates that it is well in line with reported patterns and yet it comes with much more detail depths with respect to process representation and spatio-temporal resolution than these.

3.5 Permafrost distribution and active-layer thickness

The current permafrost distribution and the active-layer thickness (Fig. 9) is well represented by the LPJmL4 model compared to independent studies (Brown et al., 1998, 2000). LPJmL4 is able to reproduce the distribution of permafrost and the measured active-layer thickness in most grid cells. The continuous permafrost zone is characterized by a thawing depth of equal or less than 1 m in LPJmL4, while the model simulates for sporadic permafrost and isolated patches a thawing depth of more than 3 m. The spatial distribution of greater thaw depth from north to south is simulated well by the model. CALM station data show a similar thawing depth as simulated by LPJmL4 (Fig. 9, bottom), but CALM station data indicate also that thawing depth can be different for the same grid cell, as other processes (e.g. exposition) not represented by LPJmL4 can play an important role.

3.6 Fire

3.6.1 Burnt area

Simulated fractional area burnt is largest in the seasonal dry tropics and temperate regions in all model versions and smallest in cold or wet environments (SI-Fig. S72). However, maximum fractional burnt area does not exceed 0.0625 in tropical and subtropical savannah and shrubland areas when the Glob-FIRM model is applied. It is comparable to GFED4 and CCI estimates only in South America, while in other tropical regions GFED4 (Giglio et al., 2013) and CCI reports fractional burnt area between 0.125 and 0.75 (SI-Fig. S72). In these regions, fractional burnt area simulated by the SPITFIRE model is overestimated with values between 0.25 and 1, specifically in the southern hemispheric Africa and northern Australia. SPITFIRE is very sensitive to vegetation, thus fuel composition where homogeneous C4 grasslands can lead to an overestimation of simulated area burnt which is specifically the case for seasonally dry South America and the Indian subcontinent. LPJmL4-GSI-SPITFIRE captures the distribution of fractional burnt area much better than LPJmL4-GSI-GlobFIRM which is too homogeneous in its response.

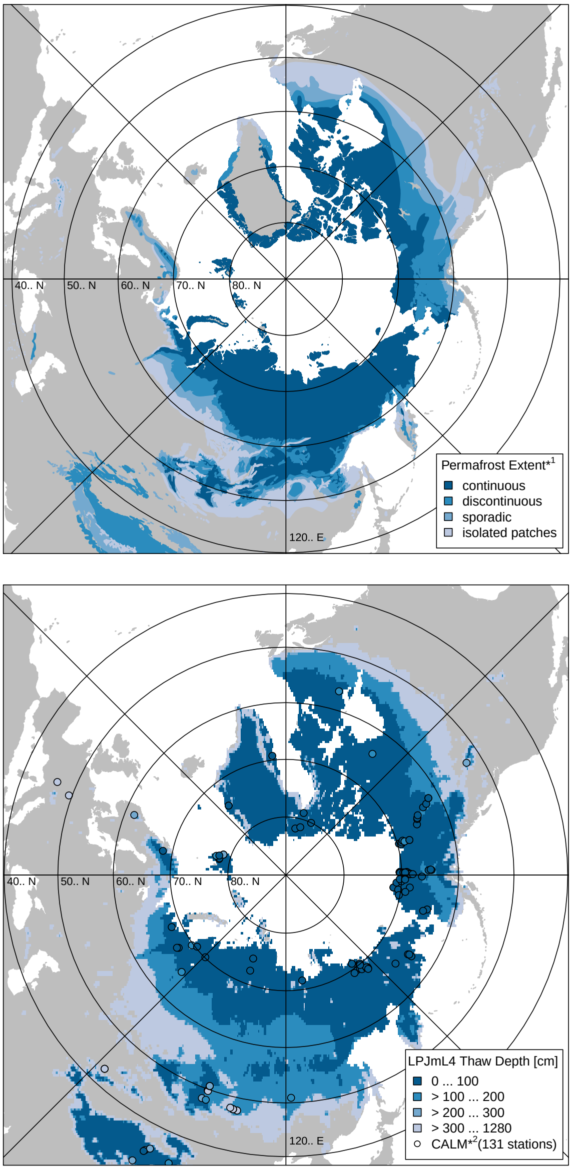


Figure 9. Observed and simulated permafrost distribution and active layer thickness. Top, contemporary permafrost extent according to the IPA Circum–Arctic Map of Permafrost (*[1] Brown et al. (1998)). Bottom, LPJmL4-simulated active-layer thickness compared to the *[2] CALM station data means both for the observation time 1991-2009 (http://www.gwu.edu/ calm/; Brown et al. (2000)). The colour scheme used at the bottom are the same for simulated thaw depth and Circumpolar Active Layer Monitoring (CALM) data.

In contrast, LPJmL4-GSI-SPITFIRE better captures the very small fractions reported for the wet tropical forests which is better comparable to GFED4. Here, the approach to simulate fire risk based on the climatic fire danger index instead of deriving a fire probability from the top-soil soil moisture is of great advantage in these regions. While LPJmL4-GSI-GlobFIRM simulates a relatively homogeneous spatial distribution of fractional burnt area in temperate and boreal forest regions, LPJmL4-GSI-SPITFIRE underestimates fractional burnt area in these biomes. LPJmL4-GSI-GlobFIRM underestimates fractional burnt area in the temperate steppe regions, whereas LPJmL4-GSI-SPITFIRE manages to spatially capture the burning conditions in these biomes, even though the total amount is overestimated. The phenology module in LPJmL4 has no effect on fractional burnt area simulated by LPJmL4-GSI-GlobFIRM, whereas including permafrost increases burnt area in the circumboreal region, specifically in Siberia, even though the spatial effect is too homogeneous.

3.6.2 Fire effects on biomass and vegetation distribution

Both fire model approaches simulate a comparable latitudinal distribution of biomass starting from the wet tropics towards dry and colder areas in the North and South. Both model versions simulate comparable values in the wet tropics around the equator and capture the gradient to seasonal dry tropics in the North (until 10°N) and South (until 20°S). The overestimation of burnt area in tropical savannahs around 20°N in LPJmL4-GSI-SPITFIRE leads to an underestimation in simulated biomass compared to the other LPJmL4 experiments. The consideration of permafrost and fire dynamics is required to reproduce observed vegetation biomass values in boreal regions.

3.6.3 Global biomass burning

The modelling errors in fractional area burnt compensate in different ways in each fire model. SPITFIRE simulates global biomass burning values of 2.7 $\mathrm{PgC\,a^{-1}}$ on average between 1996-2005 which is comparable to the 2.33 $\mathrm{PgC\,a^{-1}}$ ~~(Randerson et al., 2015)~~ suggested by Randerson et al. (2015) . Here, overestimations of burnt area in tropical savannahs and underestimations in boreal forests compensate each other. Glob-FIRM simulates more fires in boreal regions, but less spatially pronounced as in GFED4, but underestimates fractional burnt area in the subtropics and tropics. Glob-FIRM therefore estimates global biomass burning by 2.8 $\mathrm{PgC\,a^{-1}}$, similar to SPITFIRE.

3.7 Fraction of absorbed Photosynthetically Active Radiation- (FAPAR) and Albedo

Evaluations against multiple satellite datasets of FAPAR have already shown that LPJmL-GSI can well reproduce the seasonality of FAPAR and the inter-annual variability and trends in the start and end of growing season within observational uncertainties (Forkel et al., 2015). LPJmL4 shows a high spatial correlation with correlation coefficients between 0.6 and 0.71 for PEAK-FAPAR. It shows also a good agreement with the temporal variations (Fig. 10a-10c). Large parts of the wet tropics display a negative correlation between simulated and observed FAPAR, which may explain the phase-offset in the dynamics of NEE at the station Santarém. However, in these regions also the ~~difference~~ differences between datasets are large which is caused by the limitations of optical satellite observations in regions with permanent cloud cover (Forkel et al., 2015).

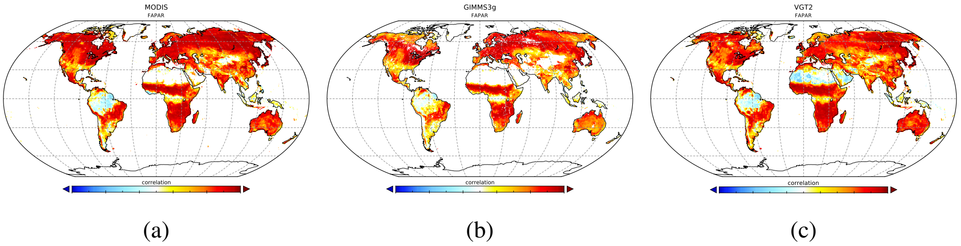


Figure 10. Evaluation of FAPAR for different data sources MODIS (a), GIMMS (b), and VGT2 (c).

LPJmL4 reproduces the global patterns of annual peak FAPAR (Fig. 11) well. Especially, in northern latitudes and in the tropics, LPJmL4 is within the range of the FAPAR datasets. However, LPJmL4 overestimates peak FAPAR especially in middle and low latitudes which originates from an overestimation of FAPAR in semi-arid regions. LPJmL4 reproduces well the temporal dynamic of FAPAR in most climate regions with very high correlations between simulated and observed FAPAR in temperate and boreal climates (climate regions Cf and D*) and with medium to high correlations in semi-arid climate regions (e.g. Am, As, Aw, Bsh, Bsk, Cs in SI-Fig. S73). LPJmL4 and the observational datasets show low correlations in wet tropics (Af) and in winter-dry temperate climates (Cw).

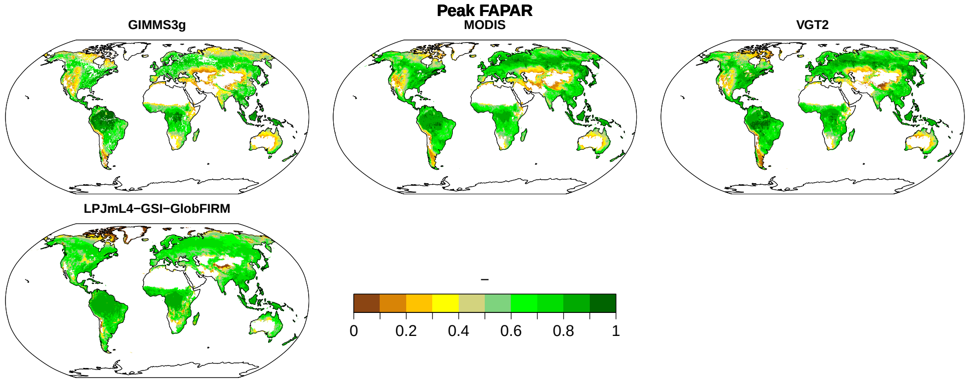


Figure 11. FAPAR mean annual peak comparison with 3 different remote sensing products.

LPJmL4 overestimates albedo in all regions (SI-Fig. S74. The temporal dynamic of snow-free albedo was well reproduced in cold steppes (climate region BSk) and in boreal regions (climate regions D*). The correlation between simulated and observed albedo is poor in tropical semi-arid and temperate climates (e.g. As, Aw, Cs, Cf). This is likely caused by soil moisture-induced changes in soil and background albedo, which has a great effect on soil reflectance (Lobell and Asner, 2002) outside the vegetation season. Such changes are not considered in LPJmL4.

3.8 Agriculture

3.8.1 Crop yields variability

The evaluation of simulated crop growth and yield can be assessed at individual sites if the model is used as a point model as in different model intercomparison simulations (Asseng et al., 2013; Bassu et al., 2014; Kollas et al., 2015; Asseng et al., 2015) where reference data are available for end-of-season properties (most importantly: crop yield) as well as within-season dynamics (e.g. development of leaf area index (LAI)). The crop yield simulations of LPJmL were evaluated in the framework of the Agricultural Model Intercomparison and Improvement Project (AgMIP) for wheat, maize, rice and soybean by ~~(Müller et al., 2017)~~ Müller et al. (2017) . They find that the performance of LPJmL is similar to that of the other gridded crop models in that model ensemble (n = 14). We here supplement the model evaluation with time series correlation analyses for the ten top-producing countries for all crops implemented in LPJmL4 (Schaphoff et al., under Revision). Results are portrayed in Fig. 12, except for field peas where no spatial data on crop-specific harvested areas exists for aggregation to national yield time series (Porwollik et al., 2016). As national yield levels are roughly calibrated in standard LPJmL simulations (Fader et al., 2010), a comparison of the mean bias is not providing insights on model performance. As management intensity is assumed to be static in the simulations (section 2.1), yield trends cannot be reproduced so that simulated and reported national yield time series have been detrended with a running mean approach (Müller et al., 2017) prior to comparison. For a more comprehensive evaluation of LPJmL's performance in yield simulations, see Müller et al. (2017).

The agreement between simulated and observed yields is not only dependent on model performance, but also on the aggregation mask used (Porwollik et al., 2016), assumptions on management and model parametrization (Folberth et al., 2016a), soil parameters (Folberth et al., 2016b) and weather data inputs (Ruane et al., 2016). LPJmL4 yield simulations are typically correlated with national yield statistics (FAO-AQUASTAT, 2014) for some of the 10 top-producing countries for each crop, but only for one ~~of these for~~ country in case of cassava (Brazil) and sugarcane (China) (Fig. 12 and supplementary material Fig. S75-S83 for the other crops).

3.8.2 Biomass yield

For the purpose of this evaluation, irrigated and rainfed biomass plants were simulated to grow globally, wherever biophysical conditions allow sustained growth. The averaged simulated yields for the 16-year period (1994–2009) were compared to reported biomass yields of switchgrass, miscanthus, poplar, willow and eucalyptus plantations on experimental test-sites located in the respective grid cell (Fig. 13). It shows that simulated yields are mostly within the range of observations for miscanthus, poplar, willow and eucalyptus, but mostly overestimates switchgrass productivity. Management options for BFTs implemented in LPJmL4 are limited to irrigation management (rainfed and fully

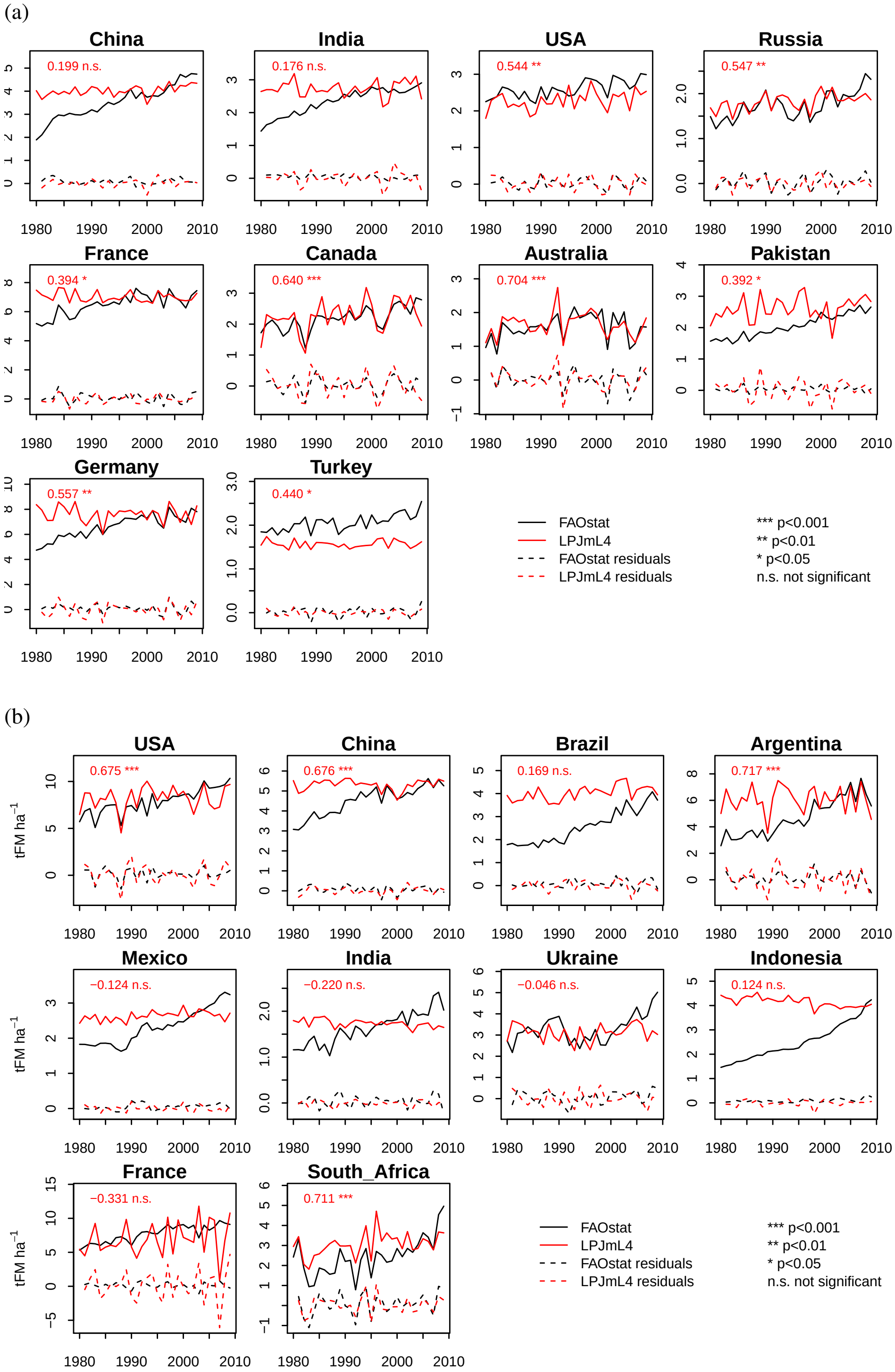
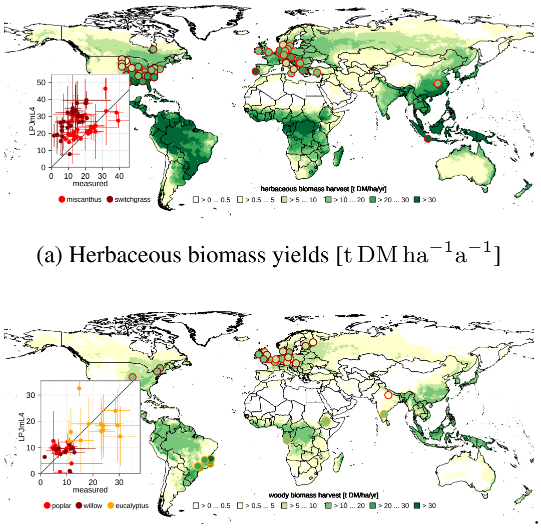


Figure 12. Evaluation of simulated yield variability for wheat (a) and maize (b) in comparison to FAO-data (FAOSTAT).



(a) Herbaceous biomass yields $[\mathrm{t\,DM\,ha^{-1}a^{-1}}]$

(b) Woody biomass yields $[\mathrm{t\,DM\,ha^{-1}a^{-1}}]$

Figure 13. Map of simulated biomass yields by LPJmL4 from rainfed herbaceous (a) and woody (b) BFTs (averages 1994–2009). Dots indicate the location of the experimental sites and measured yield, with colours scaled to map colours. Scatterplots compare observed and simulated yields in the respective grid cells. Model uncertainty is derived from simulations with and without irrigation. Observation uncertainty reflects dependencies on plantation management (adapted from Heck et al. (2016)).

irrigated), because plant species and plantation characteristics (e.g. sapling size and crop spacing) are ~~parametrised~~ parameterised as a constant scenario setting and were not varied here. The differences between rainfed and irrigated biomass yield simulations are depicted as vertical error bars in Fig. 13. The range of rainfed vs. fully irrigated biomass yields represent an approximation of management uncertainty, because simulated yields depend strongly on water availability. Nevertheless the simulated yield range is likely to represent an optimal field management for rainfed resp. irrigated plantations as nutrient limitations are not taken into account in these simulations.

3.8.3 Month of sowing

The average mean error (ME) for all crops globally is smaller than two months, with the exception of pulses (Table 5). For wheat (excl. Russia), millet, rice, sunflower and sugar beet, the agreement between simulated and observed timing of sowing is higher, with a difference of about one month. The Willmott coefficients (W) are high indicating good agreement between observations and simulations ($W > 0.85$) for all crops except pulses, sugar beet and groundnut. Both measures indicate closer agreement for pulses, groundnut, sunflower and rapeseed in temperate regions (Waha et al., 2012). Poor agreement, with differences between simulated and observed sowing dates of more than five months, is found for maize and cassava in Southeast Asia and China (for maize in East Africa),

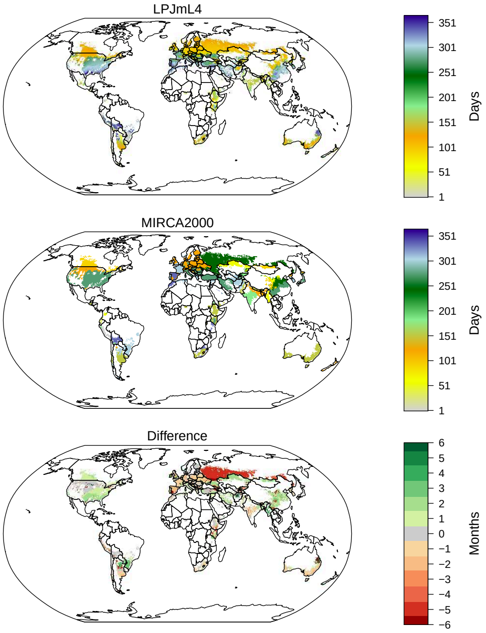


Figure 14. Evaluation of sowing dates of wheat: (from top to bottom panel) simulated (LPJmL4) sowing date, observed (MIRCA2000) sowing date and difference between simulated and observed sowing date. Green colours (red colours) in the difference map indicate that simulated sowing dates are too late (too early) compared to observations. White colours indicate crop area with less than 0.001% of the grid cell area. Regions without seasonality are not shown.

for wheat in Russia, for pulses in Southeast Asia, India, West and East Africa, the south-east region of Brazil and southern Australia, for groundnut in India and Indonesia, and for rapeseed in southern Australia and southern Europe (for wheat Fig. 14; for the other crops SI-Fig. S84-S93). Divergences are also substantial for crops growing in the southern part of the Democratic Republic of Congo, in Indo-China and in tropical climates.

There are several reasons for these disagreements between sowing dates simulated solely using climate data and the global crop calendar, please see Waha et al. (2012) for a more detailed discussion. Firstly the crop varieties in the crop calendar and simulated here differ, i.e. spring and winter varieties of wheat and rapeseed in temperate regions (e.g. in Russia). Secondly, multiple cropping in tropical regions with high cropping intensity and complex cropping systems is not considered here. Thirdly, we use ~~of~~ only one global temperature threshold for simulating sowing temperatures, which is known to vary between regions and lastly, there are other uncertainties in our method of simulating sowing dates and in the global crop calendar we use for comparison. We are also neglecting important factors such as the availability of labour and machinery, social customs, markets and prizes, the demand for certain agricultural products at certain times in the year.

Table 5. Indices of agreement between simulated (LPJmL4) and observed (MIRCA2000) sowing dates.

	All cells			Precipitation seasonality			Temperature seasonality		
Crop	W [-]	ME [days]	N	W [-]	ME [days]	N [%]	W [-]	ME [days]	N [%]
wheat	0.87	44	13962	0.86	40	15	0.87	44	85
rice	0.90	25	4995	0.90	24	82	0.87	28	18
maize	0.88	37	16333	0.89	37	48	0.85	36	52
millet	0.89	17	7851	0.92	16	63	0.89	31	37
pulses	0.63	69	14712	0.61	80	48	0.84	37	52
sugarbeet	0.37	19	2918	0.24			0.37	19	100
cassava	0.93	51	6082	0.93	51	83	0.95	57	17
sunflower	0.92	25	5876	0.87	45	22	0.93	22	78
soybean	0.94	36	8259	0.94	35	31	0.92	36	69
groundnut	0.77	34	5642	0.71	36	81	0.96	20	19
rapeseed	0.86	49	5680	0.36	135	13	0.92	37	87
wheat excl. Russia	0.94	30	11511	0.86	40	18	0.94	29	82

Mean absolute error (ME) and the Willmott coefficient of agreement (W)

The comparison to the global crop calendar, however, shows that close agreement between simulated and observed sowing dates can be achieved with purely climate-driven rules for large parts of the earth for wheat, rice, maize, millet, soybean and sunflower, as well as for pulses and groundnut in temperate regions. For about 75% of the global cropping area the difference between simulated and observed sowing dates is two months and with the exception of cassava and rapeseed 80% of the crop area displays a difference of only one month which is the minimum ~~difference possible~~ possible difference as the crop calendar reports monthly sowing dates.

4 Conclusions

This article provides a comprehensive evaluation of the now launched version 4.0 of the LPJmL DGVM that includes an operational representation of agriculture. Unique in its combination of features, the LPJmL4 model enables simulation of carbon and water fluxes linked to the dynamics of both natural and agricultural vegetation in a single, internally consistent ~~frameworks~~framework. We show that the model has great strength in reproducing carbon fluxes, especially for NBP on the global scale and NEE on the local scale. But we are also able to show that water fluxes matches well other estimates. Both, carbon and water fluxes, are the link to many ecosystem processes that the model represents and therefore are very important for the understanding of its interrelation. On the agriculture sector we synthesize that in regions with a strong weather signal the model is able to match annual yield variability. Nevertheless, in high managed countries yield variability is not well reproduced by the LPJmL4 model. This can be explained by the absent of a management module in the model. By following suggestions for objective intercomparative benchmarking systems of multiple models with dedicated software (Abramowitz, 2012; Kelley et al., 2013; Luo et al., 2012), the evaluation takes into account a number of performance metrics, diagnostic plots and a broad range of fundamental model features. This work thus goes well beyond earlier evaluations of DGVMs (see Kelley et al. (2013)) and of model evaluations published for earlier versions of LPJmL or its modules.

Pending major model improvements — anticipated as part of forthcoming LPJmL versions — are the incorporation of a scheme for calculating groundwater recharge and storage, the representation of nitrogen cycling for both natural and agricultural landscapes, consideration of ozone effects on plants (Schauberger et al., submitted) and of soil degradation, representation of wetlands with associated methane emissions, the continuous refinement of crop parameterization including multi-cropping and other management forms, and possibly a revised implementation of soil moisture (following e.g. Evaristo et al. (2015)) and stomatal conductance (following e.g. Lin et al. (2015)). As such improvements are expected to have significant effects e.g. on plant production, carbon and water fluxes – thus influencing overall model performance – any future LPJmL version will routinely be subjected to the evaluation protocol used here and, if applicable, tested against other standardized inter-model benchmarks (including participation in model intercomparisons with evaluation of single components such as in Hattermann et al. (2017)). Such continued model maintenance and benchmarking shall also keep pace with recent developments in observational and experimental data, ideally supporting identification of key uncertainties in model performance (see Medlyn et al. (2015); Smith et al. (2016)).

Besides identifying features for future model improvement, we here demonstrate adequate performance of the LPJmL4 DGVM in terms of the simulation of long-term averages and also the temporal dynamics across biogeochemical, hydrological and agricultural processes. This unique capacity renders the LPJmL4 model suitable for process-based analyses of biosphere dynamics including assessments of multi-sectoral impacts of climate change or other anthropogenic earth system ~~interference~~interferences.

5 Code and data availability

The model code of LPJmL4 is publicly available through PIK's gitlab server at https://gitlab.pik-potsdam.de/lpjml/LPJmL and an exact version of the code described here is archived under doi~~"xyz"~~: http://doi.org/10.5880/pik.2018.002 and should be referenced by Schaphoff et al. (2018b) . The output data from the model simulations described here is available at the research data repository http://dataservices.gfz-potsdam.de/portal/ under doi~~"ABC". "~~: http://doi.org/10.5880/pik.2017.009 and can be referenced by (Schaphoff et al., 2018a) .

Acknowledgements. This study was supported by the German Federal Ministry of Education and Research's (BMBF's) "PalMod 2.3 Methankreislauf, Teilprojekt 2 Modellierung der Methanemissionen von Feucht- und Permafrostgebieten mit Hilfe von LPJmL", (FKZ 01LP1507C). M.F. was funded by the TU Wien Wissenschaftspreis 2015 awarded to Wouter Dorigo. This work used eddy covariance data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE. The FLUXNET eddy covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices. We thank the coordinators of the Circumpolar Active Layer Monitoring (CALM) program for providing thaw depth, the National Snow & Ice Data Center for providing Circum-Arctic Map of Permafrost, and the R-ArcticNET for discharge data. Furthermore. we thank Jena-BGI data for providing GPP, latent heat flux, total biomass and soil carbon data. We also thank the providers of the evaluation data of FAPAR (GIMMS3g FAPAR, VGT2 FAPAR, MODIS FAPAR) and the remote sensing data of GFED4 and CCI for evaluating fractional burnt area. MODIS C5 albedo time series data product was retrieved from the online Data Pool, courtesy of the NASA Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, . We thank the Climatic Research Unit for providing global gridded temperature input, the Global Precipitation Climatology Centre for providing precipitation input and the coordinators of ERA-Interim for providing shortwave downward radiation and net downward longwave radiation. Furthermore, we thank the coordinators of MIRCA2000 for providing land use input. Finally, we thank Kirsten Elger for here great support for archiving data and the LPJmL4 code and two anonymous reviewers for their helpful comments on earlier versions of the manuscript.

References

Abramowitz, G.: Towards a benchmark for land surface models, Geophysical Research Letters, 32, n/a–n/a, doi:10.1029/2005GL024419, 2005.

Abramowitz, G.: Towards a public, standardized, diagnostic benchmarking system for land surface models, Geosci. Model Dev., pp. 819–827, doi:10.5194/gmd-5-819-2012, 2012, 2012.

Alexandratos, N. and Bruinsma, J.: World agriculture towards 2030/2050: the 2012 revision, Tech. Rep. 12, FAO, Rome, FAO, 2012.

Asseng, S., Brisson, N., Basso, B., Martre, P., Aggarwal, P. K., Angulo, C., Bertuzzi, P., Biernath, C., Challinor, A. J., Doltra, J., Gayler, S., Goldberg, R., Grant, R., Heng, L., Hooker, J., Hunt, L. A., Ingwersen, J., Izaurralde, R. C., Kersebaum, K. C., Müller, C., Kumar, S. N., Nendel, C., Leary, G. O., Olesen, J. E., Osborne, T. M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M. A., Shcherbak, I., Steduto, P., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., Wallach, D., Williams, J. R., and Wolf, J.: Uncertainty in simulating wheat yields under climate change - Supplementary Information, Nature Climate Change, doi:10.1038/NCLIMATE1916, 2013.

Asseng, S., Ewert, F., Martre, P., Rotter, R. P., Lobell, D. B., Cammarano, D., Kimball, B. A., Ottman, M. J., Wall, G. W., White, J. W., Reynolds, M. P., Alderman, P. D., Prasad, P. V. V., Aggarwal, P. K., Anothai, J., Basso, B., Biernath, C., Challinor, A. J., De Sanctis, G., Doltra, J., Fereres, E., Garcia-Vila, M., Gayler, S., Hoogenboom, G., Hunt, L. A., Izaurralde, R. C., Jabloun, M., Jones, C. D., Kersebaum, K. C., Koehler, A.-K., Muller, C., Naresh Kumar, S., Nendel, C., O/'Leary, G., Olesen, J. E., Palosuo, T., Priesack, E., Eyshi Rezaei, E., Ruane, A. C., Semenov, M. A., Shcherbak, I., Stockle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Thorburn, P. J., Waha, K., Wang, E., Wallach, D., Wolf, J., Zhao, Z., and Zhu, Y.: Rising temperatures reduce global wheat production, Nature Clim. Change, 5, 143–147, doi:10.1038/nclimate2470, 2015.

Baret, F., Weiss, M., Lacaze, R., Camacho, F., Makhmara, H., Pacholcyzk, P., and Smets, B.: GEOV1: LAI and FAPAR essential climate variables and FCOVER global time series capitalizing over existing products. Part1: Principles of development and production, Remote Sensing of Environment, 137, 299–309, doi:10.1016/j.rse.2012.12.027, 2013.

Bassu, S., Brisson, N., Durand, J.-L., Boote, K., Lizaso, J., Jones, J. W., Rosenzweig, C., Ruane, A. C., Adam, M., Baron, C., Basso, B., Biernath, C., Boogaard, H., Conijn, S., Corbeels, M., Deryng, D., De Sanctis, G., Gayler, S., Grassini, P., Hatfield, J., Hoek, S., Izaurralde, C., Jongschaap, R., Kemanian, A. R., Kersebaum, K. C., Kim, S.-H., Kumar, N. S., Makowski, D., Müller, C., Nendel, C., Priesack, E., Pravia, M. V., Sau, F., Shcherbak, I., Tao, F., Teixeira, E., Timlin, D., and Waha, K.: How do various maize crop models vary in their responses to climate change factors?, Global change biology, doi:10.1111/gcb.12520, 2014.

Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U., and Ziese, M.: A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present, Earth System Science Data, 5, 71–99, doi:10.5194/essd-5-71-2013, http://www.earth-syst-sci-data.net/5/71/2013/, 2013.

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 (New York, N.Y.), 329, 834–8, doi:10.1126/science.1184984, 2010.

Biemans, H., Hutjes, R. W. a., Kabat, P., Strengers, B. J., Gerten, D., and Rost, S.: Effects of Precipitation Uncertainty on Discharge Calculations for Main River Basins, Journal of Hydrometeorology, 10, 1011–1025, doi:10.1175/2008JHM1067.1, 2009.

Biemans, H., Haddeland, I., Kabat, P., Ludwig, F., Hutjes, R. W. a., Heinke, J., von Bloh, W., and Gerten, D.: Impact of reservoirs on river discharge and irrigation water supply during the 20th century, Water Resources Research, 47, W03 509, doi:10.1029/2009WR008929, 2011.

Boden, T., Marland, G., and Andres, R.: Global, Regional, and National Fossil-Fuel CO_2 Emissions, Carbon Dioxide Information Analysis Center (CDIAC), Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, http://cdiac.ornl.gov/trends/emis/overview.html, 2013.

Bondeau, A., Smith, P., Zaehle, S., Schaphoff, S., Lucht, W., Cramer, W., Gerten, D., Lotze-Campen, Hermann, Müller, C., Reichstein, M., and Smith, B.: Modelling the role of agriculture for the 20th century global terrestrial carbon balance, Global Change Biology, 13, 679–706, doi:10.1111/j.1365-2486.2006.01305.x, 2007.

Brouwer, C., Prins, K., and Heibloem, M.: Irrigation Water Management : Irrigation Scheduling. Training manual no. 4, Tech. Rep. 4, FAO Land and Water Development Division, Rome, Italy, http://www.fao.org/docrep/t7202e/t7202e00.htm, 1989.

Brown, J., Ferrians, O. J. J., Heginbottom, J. A., and Melnikov, E. S.: Circum-Arctic map of permafrost and ground-ice conditions, Boulder, CO: National Snow and Ice Data Center/World Data Center for Glaciology, http://nsidc.org/data/ggd318.html, 1998.

Brown, J., Hinkel, K. M., and Nelson, F. E.: The circumpolar active layer monitoring (calm) program: Research designs and initial results, Polar Geography, 24, 166–258, doi:10.1080/10889370009377698, 2000.

Carvalhais, N., Forkel, M., Khomik, M., Bellarby, J., Jung, M., Migliavacca, M., Mu, M., Saatchi, S., Santoro, M., Thurner, M., Weber, U., Ahrens, B., Beer, C., Cescatti, A., Randerson, J. T., and Reichstein, M.: Global covariation of carbon turnover times with climate in terrestrial ecosystems, Nature, 514, 213–217, 10.1038/nature13731, 2014.

Chaturvedi, V., Hejazi, M., Edmonds, J., Clarke, L., Kyle, P., Davies, E., and Wise, M.: Climate mitigation policy implications for global irrigation water demand, Mitigation and Adaptation Strategies for Global Change, 20, 389–407, doi:10.1007/s11027-013-9497-4, 2015.

Chuvieco, E., Yue, C., Heil, A., Mouillot, F., Alonso-Canas, I., Padilla, M., Pereira, J. M., Oom, D., and Tansey, K.: A new global burned area product for climate assessment of fire impacts, Global Ecology and Biogeography, 25, 619–629, doi:10.1111/geb.12440, http://dx.doi.org/10.1111/geb.12440, 2016.

Cosby, B. J., Hornberger, G. M., Clapp, R. B., and Ginn, T. R.: A Statistical Exploration of the Relationships of Soil Moisture Characteristics to the Physical Properties of Soils, Water Resour. Res., 20, 682–690, doi:10.1029/WR020i006p00682, 1984.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-

J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, doi:10.1002/qj.828, http://dx.doi.org/10.1002/qj.828, 2011.

Defries, R. and Hansen, M.: ISLSCP II Continuous Fields of Vegetation Cover, 1992-1993, ORNL Distributed Active Archive Center, https://doi.org/10.3334/ORNLDAAC/931, 2009.

Döll, P. and Siebert, S.: Global modeling of irrigation water requirements, Water Resources Research, 38, 8–1, doi:10.1029/2001WR000355, 2002.

Döll, P., Hoffmann-Dobrev, H., Portmann, F., Siebert, S., Eicker, A., Rodell, M., Strassberg, G., and Scanlon, B.: Impact of water withdrawals from groundwater and surface water on continental water storage variations, Journal of Geodynamics, 59-60, 143–156, doi:10.1016/j.jog.2011.05.001, 2012.

Döll, P., Müller Schmied, H., Schuh, C., Portmann, F. T., and Eicker, A.: Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites, Water Resources Research, 50, 5698–5720, doi:10.1002/2014WR015595, 2014.

Evaristo, J., Jasechko, S., and McDonnell, J. J.: Global separation of plant transpiration from groundwater and streamflow, Nature, 525, 91–94, doi:10.1038/nature14983, 2015.

Fader, M., Rost, S., Müller, C., Bondeau, A., and Gerten, D.: Virtual water content of temperate cereals and maize: Present and potential future patterns, Journal of Hydrology, 384, 218–231, doi:10.1016/j.jhydrol.2009.12.011, 2010.

FAO-AQUASTAT: AQUASTAT database - Food and Agriculture Organization of the United Nations (FAO), http://www.fao.org/nr/water/aquastat/data/query/index.html?lang=en, 2014.

FAO/IIASA/ISRIC/ISSCAS/JRC: Harmonized World Soil Database (version 1.2)., http://www.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/, 2012.

FAOSTAT: FAOSTAT, http://www.fao.org/faostat/en/#data/QC.

Folberth, C., Elliott, J., Müller, C., Balkovic, J., Chryssanthacopoulos, J., Izaurralde, R. C., Jones, C. D., Khabarov, N., Liu, W., Reddy, A., Schmid, E., Skalsky, R., Yang, H., Arneth, A., Ciais, P., Deryng, D., Lawrence, P. J., Olin, S., Pugh, T. A. M., Ruane, A. C., and Wang, X.: Uncertainties in global crop model frameworks: effects of cultivar distribution, crop management and soil handling on crop yield estimates, Biogeosciences Discussions, pp. 1–30, doi:10.5194/bg-2016-527, 2016a.

Folberth, C., Skalský, R., Moltchanova, E., Balkovič, J., Azevedo, L. B., Obersteiner, M., and van der Velde, M.: Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations, Nature Communications, 7, 11 872, doi:10.1038/ncomms11872, 2016b.

Forkel, M., Carvalhais, N., Schaphoff, S., v. Bloh, W., Migliavacca, M., Thurner, M., and Thonicke, K.: Identifying environmental controls on vegetation greenness phenology through model–data integration, Biogeosciences, 11, 7025–7050, doi:10.5194/bg-11-7025-2014, http://www.biogeosciences.net/11/7025/2014/, 2014.

Forkel, M., Migliavacca, M., Thonicke, K., Reichstein, M., Schaphoff, S., Weber, U., and Carvalhais, N.: Codominant water control on global interannual variability and trends in land surface phenology and greenness, Global Change Biology, 21, 3414–3435, doi:10.1111/gcb.12950, 2015.

Forkel, M., Carvalhais, N., Rödenbeck, C., Keeling, R., Heimann, M., Thonicke, K., Zaehle, S., and Reichstein, M.: Enhanced seasonal CO_2 exchange caused by amplified plant productivity in northern ecosystems, Science, 351, 696, doi:10.1126/science.aac4971, http://science.sciencemag.org/content/351/6274/696.abstract, 2016.

Gerten, D., Schaphoff, S., Haberlandt, U., Lucht, W., and Sitch, S.: Terrestrial vegetation and water balance–hydrological evaluation of a dynamic global vegetation model, Journal of Hydrology, 286, 249–270, doi:doi: DOI: 10.1016/j.jhydrol.2003.09.029, http://www.sciencedirect.com/science/article/B6V6C-4B9D86T-3/2/aa123e159770c269994f0d74c5edc335, 2004.

Gerten, D., Lucht, W., Ostberg, S., Heinke, J., Kowarsch, M., Kreft, H., Kundzewicz, Z. W., Rastgooy, J., Warren, R., and Schellnhuber, H. J.: Asynchronous exposure to global warming: freshwater resources and terrestrial ecosystems, Environmental Research Letters, 8, 034 032, doi:10.1088/1748-9326/8/3/034032, 2013.

Giglio, L., Randerson, J. T., and van der Werf, G. R.: Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4), Journal Of Geophysical Research: Biogeosciences, 118, 317–328, doi:10.1002/jgrg.20042, 2013.

Haddeland, I., Clark, D. B., Franssen, W., Ludwig, F., Voß, F., Arnell, N. W., Bertrand, N., Best, M., Folwell, S., Gerten, D., Gomes, S., Gosling, S. N., Hagemann, S., Hanasaki, N., Harding, R., Heinke, J., Kabat, P., Koirala, S., Oki, T., Polcher, J., Stacke, T., Viterbo, P., Weedon, G. P., and Yeh, P.: Multimodel Estimate of the Global Terrestrial Water Balance: Setup and First Results, Journal of Hydrometeorology, 12, 869–884, doi:10.1175/2011JHM1324.1, 2011.

Harris, I., Jones, P., Osborn, T., and Lister, D.: Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset, International Journal of Climatology, 34, 623–642, doi:10.1002/joc.3711, 2014.

Harrison, P. A., Dunford, R. W., Holman, I. P., and Rounsevell, M. D. A.: Climate change impact modelling needs to include cross-sectoral interactions, Nature Clim. Change, 6, 885–890, doi:10.1038/nclimate3039, 2016.

Hattermann, F. F., Krysanova, V., Gosling, S. N., Dankers, R., Daggupati, P., Donnelly, C., Flörke, M., Huang, S., Motovilov, Y., Buda, S., Yang, T., Müller, C., Leng, G., Tang, Q., Portmann, F. T., Hagemann, S., Gerten, D., Wada, Y., Masaki, Y., Alemayehu, T., Satoh, Y., and Samaniego, L.: Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins, Climatic Change, 141, 561–576, doi:10.1007/s10584-016-1829-4, 2017.

Heck, V., Gerten, D., Lucht, W., and Boysen, L. R.: Is extensive terrestrial carbon dioxide removal a 'green' form of geoengineering? A global modelling study, Global and Planetary Change, 137, 123–130, doi:10.1016/j.gloplacha.2015.12.008, 2016.

Hoff, H., Falkenmark, M., Gerten, D., Gordon, L., Karlberg, L., and Rockström, J.: Greening the global water system, Journal of Hydrology, 384, 177–186, doi:10.1016/j.jhydrol.2009.06.026, 2010.

Iizumi, T., Yokozawa, M., Sakurai, G., Travasso, M. I., Romanenkov, V., Oettli, P., Newby, T., Ishigooka, Y., and Furuya, J.: Historical changes in global yields: major cereal and legume crops from 1982 to 2006, Global Ecology and Biogeography, 23, 346–357, doi:10.1111/geb.12120, 2014.

Jägermeyr, J., Gerten, D., Lucht, W., Hostert, P., Migliavacca, M., and Nemani, R.: A high-resolution approach to estimating ecosystem respiration at continental scales using operational satellite data, Global change biology, 20, 1191–1210, doi:10.1111/gcb.12443, 2014.

Jägermeyr, J., Pastor, A., Biemans, h., and Gerten, D.: Reconciling irrigated food production with environmental flows for Sustainable Development Goals implementation, Nature Communications, 8, doi:10.1038/ncomms15900, 2017.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, Journal of Geophysical Research: Biogeosciences, 116, doi:10.1029/2010JG001566, 2011.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., and Woollen, J.: The NCEP/NCAR 40-year reanalysis project, Bulletin of the American meteorological Society, 77, 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, 1996a.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-Year Reanalysis Project, Bulletin of the American Meteorological Society, 77, 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, 1996b.

Kaminski, T., Heimann, M., and Giering, R.: A coarse grid three-dimensional global inverse model of the atmospheric transport: 2. Inversion of the transport of CO_2 in the 1980s, Journal of Geophysical Research: Atmospheres, 104, 18 555–18 581, doi:10.1029/1999JD900146, 1999.

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

Knyazikhin, Y., Glassy, J., Privette, J. L., Tian, Y., Lotsch, A., Zhang, Y., Wang, Y., Morisette, J. T., Votava, P., Myneni, R. B., and others: MODIS leaf area index (LAI) and fraction of photosynthetically active radiation absorbed by vegetation (FPAR) product (MOD15) algorithm theoretical basis document, Theoretical Basis Document, NASA Goddard Space Flight Center, Greenbelt, MD, 20771, 1999.

Kollas, C., Kersebaum, K. C., Nendel, C., Manevski, K., Müller, C., Palosuo, T., Armas-Herrera, C. M., Beaudoin, N., Bindi, M., Charfeddine, M., Conradt, T., Constantin, J., Eitzinger, J., Ewert, F., Ferrise, R., Gaiser, T., Cortazar-Atauri, I. G. d., Giglio, L., Hlavinka, P., Hoffmann, H., Hoffmann, M. P., Launay, M., Manderscheid, R., Mary, B., Mirschel, W., Moriondo, M., Olesen, J. E., Öztürk, I., Pacholski, A., Ripoche-Wachter, D., Roggero, P. P., Roncossek, S., Rötter, R. P., Ruget, F., Sharif, B., Trnka, M., Ventrella, D., Waha, K., Wegehenkel, M., Weigel, H.-J., and Wu, L.: Crop rotation modelling—A European model intercomparison, European Journal of Agronomy, 70, 98–111, doi:10.1016/j.eja.2015.06.007, 2015.

Langerwisch, F., Rost, S., Gerten, D., Poulter, B., Rammig, A., and Cramer, W.: Potential effects of climate change on inundation patterns in the Amazon Basin, Hydrol. Earth Syst. Sci., 17, 2247–2262, doi:10.5194/hess-17-2247-2013, 2013.

Le Quéré, C., Moriarty, R., Andrew, R. M., Canadell, J. G., Sitch, S., Korsbakken, J. I., Friedlingstein, P., Peters, G. P., Andres, R. J., Boden, T. A., Houghton, R. A., House, J. I., Keeling, R. F., Tans, P., Arneth, A., Bakker, D. C. E., Barbero, L., Bopp, L., Chang, J., Chevallier, F., Chini, L. P., Ciais, P., Fader, M., Feely, R. A., Gkritzalis, T., Harris, I., Hauck, J., Ilyina, T., Jain, A. K., Kato, E., Kitidis, V., Klein Goldewijk, K., Koven, C., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lenton, A., Lima, I. D., Metzl, N., Millero, F., Munro, D. R., Murata, A., Nabel, J. E. M. S., Nakaoka, S., Nojiri, Y., O'Brien, K., Olsen, A., Ono, T., Pérez, F. F., Pfeil, B., Pierrot, D., Poulter, B., Rehder, G., Rödenbeck, C., Saito, S., Schuster, U., Schwinger, J., Séférian, R., Steinhoff, T., Stocker, B. D., Sutton, A. J., Takahashi, T., Tilbrook, B., van der Laan-Luijkx, I. T., van der Werf, G. R., van Heuven, S., Vandemark, D., Viovy, N., Wiltshire, A., Zaehle, S., and Zeng, N.: Global Carbon Budget 2015, Earth System Science Data, 7, 349–396, doi:10.5194/essd-7-349-2015, http://www.earth-syst-sci-data.net/7/349/2015/, 2015.

Lehner, B., Liermann, C. R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., and Magome, J.: High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management, Frontiers in Ecology and the Environment, 9, 494–502, doi:10.1890/100125, 2011.

Lin, Y., Medlyn, B. E., Duursma, R. E., Prentice, I. C., Wang, H., Baig, S., Eamus, D., Resco de Dios, V., Mitchell, P., Ellsworth, D. S., Op de Beeck, M., Wallin, G., Uddling, J., Tarvainen, L., Linderson, M., Cernusak, L. A., Nippert, J. B., Ocheltree, T. W., Tissue, D. T. andMartin-StPaul, N. K., Rogers, A., Warren, J. M., De Angelis, P., Hikosaka, K., Han, Q., Onoda, Y., Gimeno, T. E., Barton, C. V. M. andBennie, J., Bonal, J. andBosc, A., Löw, M., Macinins-Ng, C., Rey, A., Rowland, L., Setterfield, S. A., Tausz-Posch, S., Zaragoza-Castells, J. andBroadmeadow, M. S. J., Drake, J. E., Freeman, M., Ghannoum, O., Hutley, L. B., Kelly, J. W., Kikuzawa, K., Kolari, P., Koyama, K., Limousin, J., Meir, P., Lola da Costa, A. C., Mikkelsen, T. N., Salinas, N., Sun, W., and Wingate, L.: Optimal stomatal behaviour around the world, Nature Clim. Change, pp. 459–464, doi:10.1038/nclimate2550, 2015.

Liu, Y. Y., van Dijk, A. I. J. M., de Jeu, R. A. M., Canadell, J. G., McCabe, M. F., Evans, J. P., and Wang, G.: Recent reversal in loss of global terrestrial biomass, Nature Clim. Change, 5, 470–474, doi:10.1038/nclimate2581, 2015.

Lobell, D. B. and Asner, G. P.: Moisture Effects on Soil Reflectance, Soil Science Society of America Journal, 66, 722–727, doi:10.2136/sssaj2002.7220, 2002.

Lucht, W., Schaaf, C., and Strahler, A.: An algorithm for the retrieval of albedo from space using semiempirical BRDF models, IEEE Transactions on Geoscience and Remote Sensing, 38, 977–998, doi:10.1109/36.841980, 2000.

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, pp. 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, Global Change Biology, 13, 2509–2537, doi:10.1111/j.1365-2486.2007.01439.x, 2007.

Medlyn, B. E., Zaehle, S., De Kauwe, M. G., Walker, A. P., Dietze, M. C., Hanson, P. J., Hickler, T., Jain, A. K., Luo, Y., Parton, W., Prentice, I. C., Thornton, P. E., Wang, S., Wang, Y.-P., Weng, E., Iversen, C. M., McCarthy, H. R., Warren, J. M., Oren, R., and Norby, R. J.: Using ecosystem experiments to improve vegetation models, Nature Climate Change, 5, 528–534, doi:10.1038/nclimate2621, 2015.

Müller, C., Stehfest, E., Minnen, J. G. v., Strengers, B., Bloh, W. v., Beusen, A. H. W., Schaphoff, S., Kram, T., and Lucht, W.: Drivers and patterns of land biosphere carbon balance reversal, Environmental Research Letters, 11, 044 002, doi:10.1088/1748-9326/11/4/044002, 2016.

Müller, C., Elliott, J., Chryssanthacopoulos, J., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Folberth, C., Glotter, M., Hoek, S., Iizumi, T., Izaurralde, R. C., Jones, C., Khabarov, N., Lawrence, P., Liu, W., Olin, S., Pugh, T. A. M., Ray, D. K., Reddy, A., Rosenzweig, C., Ruane, A. C., Sakurai, G., Schmid, E., Skalsky, R., Song, C. X., Wang, X., de Wit, A., and Yang, H.: Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications, Geoscientific Model Development, 10, 1403–1422, doi:10.5194/gmd-10-1403-2017, 2017.

Nachtergaele, F., van Velthuizen, H., Verelst, L., Batjes, N., Dijkshoorn, K., van Engelen, V., Fischer, G., Jones, A., Montanarella, L., and Petri, M.: Harmonized world soil database, Food and Agriculture Organization of the United Nations, http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/, 2008.

New, M., Hulme, M., and Jones, P.: Representing Twentieth-Century Space–Time Climate Variability. Part II: Development of 1901–96 Monthly Grids of Terrestrial Surface Climate, Journal of Climate, 13, 2217–2238, doi:10.1175/1520-0442(2000)013<2217:RTCSTC>2.0.CO;2, 2000.

ORNL DAAC, Oak Ridge, T. U.: Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC), http://fluxnet.ornl.gov/, 2011.

Ostberg, S., Schaphoff, S., Lucht, W., and Gerten, D.: Three centuries of dual pressure from land use and climate change on the biosphere, Environmental Research Letters, 10, 44 011, doi:10.1088/1748-9326/10/4/044011, 2015.

Portmann, F. T., Siebert, S., Bauer, C., and Döll, P.: Global dataset of monthly growing areas of 26 irrigated crops, Frankfurt Hydrology Paper, 2008.

Portmann, F. T., Siebert, S., and Döll, P.: MIRCA2000 - Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling, Global Biogeochemical Cycles, 24, 1–24, doi:10.1029/2008GB003435, 2010.

Porwollik, V., Müller, C., Elliott, J., Chryssanthacopoulos, J., Iizumi, T., Ray, D. K., Ruane, A. C., Arneth, A., Balkovič, J., Ciais, P., Deryng, D., Folberth, C., Izaurralde, R. C., Jones, C. D., Khabarov, N., Lawrence, P. J., Liu, W., Pugh, T. A., Reddy, A., Sakurai, G., Schmid, E., Wang, X., de Wit, A., and Wu, X.: Spatial and temporal uncertainty of crop yield aggregations, European Journal of Agronomy, doi:10.1016/j.eja.2016.08.006, 2016.

Randerson, J., van der Werf, G. R., Giglio, L., Collatz, G. J., and Kasibhatla, P. S.: Global Fire Emissions Database, Version 4, (GFEDv4), ORNL DAAC, doi:10.3334/ORNLDAAC/1293, 2015.

Ray, D. K., Gerber, J. S., MacDonald, G. K., and West, P. C.: Climate variation explains a third of global crop yield variability, Nature Communications, 6, 5989, doi:10.1038/ncomms6989, 2015.

Rödenbeck, C.: Estimating CO2 sources and sinks from atmospheric mixing ratio measurements using a global inversion of atmospheric transport, Technical Reports, Max Planck Institute for Biogeochemistry, http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:1691952, 2005.

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, Atmospheric Chemistry and Physics, 3, 1919–1964, doi:10.5194/acp-3-1919-2003, 2003.

Rosegrant, M. W., Cai, X., and Cline, S. A.: World Water and Food to 2025: Dealing with Scarcity, Tech. rep., International Food Policy Research Institute, Washigton, D.C., 2002.

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., and Khabarov, N.: Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison, Proceedings of the National Academy of Sciences, 111, 3268–3273, doi:10.1073/pnas.1222463110, 2014.

Ruane, A. C., Hudson, N. I., Asseng, S., Camarrano, D., Ewert, F., Martre, P., Boote, K. J., Thorburn, P. J., Aggarwal, P. K., and Angulo, C.: Multi-wheat-model ensemble responses to interannual climate variability, Environmental Modelling & Software, 81, 86–101, doi:10.1016/j.envsoft.2016.03.008, 2016.

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. A., Salas, W., Zutta, B. R., Buermann, W., Lewis, S. L., Hagen, S., Petrova, S., White, L., Silman, M., and Morel, A.: Benchmark map of forest carbon stocks in tropical regions across three continents, Proceedings of the National Academy of Sciences, 108, 9899–9904, doi:10.1073/pnas.1019576108, 2011.

Saleska, S. R., Miller, S. D., Matross, D. M., Goulden, M. L., Wofsy, S. C., Da Rocha, H. R., De Camargo, P. B., Crill, P., Daube, B. C., De Freitas, H. C., and others: Carbon in Amazon Forests: Unexpected Seasonal Fluxes and Disturbance-Induced Losses, Science, 302, 1554–1557, doi:10.1126/science.1091165, 2003.

Sauer, T., Havlík, P., Schneider, U. a., Schmid, E., Kindermann, G., and Obersteiner, M.: Agriculture and resource availability in a changing world: The role of irrigation, Water Resources Research, 46, doi:10.1029/2009WR007729, 2010.

Schaaf, C. B., Gao, F., Strahler, A. H., Lucht, W., Li, X., Tsang, T., Strugnell, N. C., Zhang, X., Jin, Y., Muller, J.-P., Lewis, P., Barnsley, M., Hobson, P., Disney, M., Roberts, G., Dunderdale, M., Doll, C., d'Entremont, R. P., Hu, B., Liang, S., Privette, J. L., and Roy, D.: First operational BRDF, albedo nadir reflectance products from MODIS, The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring, 83, 135–148, doi:10.1016/S0034-4257(02)00091-3, 2002.

Schaphoff, S., Heyder, U., Ostberg, S., Gerten, D., Heinke, J., and Lucht, W.: Contribution of permafrost soils to the global carbon budget, Environmental Research Letters, 8, 014 026, doi:10.1088/1748-9326/8/1/014026, 2013.

Schaphoff, S., von Bloh, W., Rammig, A., Thonicke, K., Forkel, M., Biemans, H., Gerten, D., Heinke, J., Jägermyer, J., Knauer, J., Lucht, W., Müller, C., Rolinski, S., and Waha, K.: LPJmL4 model output for the publications in GMD: LPJmL4 - a dynamic global vegetation model with managed land: Part I – Model description and Part II – Model evaluation, doi:10.5880/pik.2017.009, 2018a.

Schaphoff, S., von Bloh, W., Thonicke, K., Biemans, H., Forkel, M., Heinke, J., Jägermyer, J., Müller, C., Rolinski, S., Waha, K., Stehfest, E., de Waal, L., Heyder, U., Gumpenberger, M., and Beringer, T.: LPJmL4 model code., doi:10.5880/pik.2018.002, 2018b.

Schaphoff, S., von Bloh, W., Rammig, A., Thonicke, K., Forkel, M., Biemans, H., Gerten, D., Heinke, J., Jägermyer, J., Knauer, J., Lucht, W., Müller, C., Rolinski, S., and Waha, K.: The LPJmL4 Dynamic Global Vegetation Model with managed Land: Part I - Description of a consistently calculated vegetation, hydrology and agricultural global model, Geoscientific Model Development, under Revision.

Schauberger, B., Rolinski, S., and Müller, C.: A network-based approach for semi-quantitative knowledge mining and its application to yield variability, Environmental Research Letters, 11, 123 001, doi:10.1088/1748-9326/11/12/123001, 2016.

Siderius, C., Biemans, H., Wiltshire, A., Rao, S., Franssen, W. H. P., Kumar, P., Gosain, A. K., van Vliet, M. T. H., and Collins, D. N.: Snowmelt contributions to discharge of the Ganges, Science of the Total Environment, 468, S93–S101, doi:10.1016/j.scitotenv.2013.05.084, 2013.

Siebert, S. and Döll, P.: Quantifying blue and green virtual water contents in global crop production as well as potential production losses without irrigation, Journal of Hydrology, 384, 198–217, doi:10.1016/j.jhydrol.2009.07.031, 2010.

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, Global Change Biology, 9, 161–185, doi:10.1046/j.1365-2486.2003.00569.x, 2003.

Smith, W. K., Reed, S. C., Cleveland, C. C., Ballantyne, A. P., Anderegg, W. R. L., Wieder, W. R., Liu, Y. Y., and Running, S. W.: Large divergence of satellite and Earth system model estimates of global terrestrial CO_2 fertilization, Nature Climate Change, 6, 306–310, doi:10.1038/nclimate2879, 2016.

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R., Carpenter, S. R., de Vries, W., de Wit, C. A., Folke, C., Gerten, D., Heinke, J., Mace, G. M., Persson, L. M., Ramanathan, V., Reyers, B., and Sörlin, S.: Planetary boundaries: Guiding human development on a changing planet, Science, doi:10.1126/science.1259855, 2015.

Tarnocai, C., Canadell, J. G., Schuur, E. A. G., Kuhry, P., Mazhitova, G., and Zimov, S.: Soil organic carbon pools in the northern circumpolar permafrost region, Global Biogeochemical Cycles, 23, doi:10.1029/2008GB003327, 2009.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, Journal of Geophysical Research: Atmospheres, 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 Ecology and Biogeography, 10, 661–677, doi:10.1046/j.1466-822X.2001.00175.x, 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, http://www.biogeosciences.net/7/1991/2010/, 2010.

Thoning, K., Tans, P., and Komhyr, W.: Atmospheric carbon dioxide at Mauna Loa Observatory. II-Analysis of the NOAA GMCC data, 1974-1985, Journal of Geophysical Research, 94, 8549–8565, doi:10.1029/JD094iD06p08549, 1989.

Thurner, M., Beer, C., Santoro, M., Carvalhais, N., Wutzler, T., Schepaschenko, D., Shvidenko, A., Kompter, E., Ahrens, B., Levick, S. R., and Schmullius, C.: Carbon stock and density of northern boreal and temperate forests, Global Ecology and Biogeography, 23, 297–310, doi:10.1111/geb.12125, 2014.

University of East Anglia Climatic Research Unit; Harris, I.C.; Jones, P. .: CRU TS3.23: Climatic Research Unit (CRU) Time-Series (TS) Version 3.23 of High Resolution Gridded Data of Month-by-month Variation in Climate (Jan. 1901- Dec. 2014)., Centre for Environmental Data Analysis, http://dx.doi.org/10.5285/4c7fdfa6-f176-4c58-acee-683d5e9d2ed5, 2015.

Vorosmarty, C. and Fekete, B.: ISLSCP II River Routing Data (STN-30p), 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., ORNL Distributed Active Archive Center, https://doi.org/10.3334/ORNLDAAC/1005, 2011.

Vörösmarty, C. J., Fekete, B., and Tucker, B.: River Discharge Database, Version 1.0 (RivDIS v1.0), Volumes 0 through 6. A contribution to IHP-V Theme 1. Technical Documents in Hydrology Series., UNESCO, Paris, 1996.

Wada, Y. and Bierkens, M. F. P.: Sustainability of global water use: past reconstruction and future projections, Environmental Research Letters, 9, 104 003, doi:10.1088/1748-9326/9/10/104003, 2014.

Wada, Y., van Beek, L. P. H., Viviroli, D., Dürr, H. H., Weingartner, R., and Bierkens, M. F. P.: Global monthly water stress: 2. Water demand and severity of water stress, Water Resources Research, 47, doi:10.1029/2010WR009792, 2011.

Waha, K., van Bussel, L. G. J., Müller, C., and Bondeau, A.: Climate-driven simulation of global crop sowing dates, Global Ecology and Biogeography, 21, 247–259, doi:10.1111/j.1466-8238.2011.00678.x, 2012.

Waha, K., Müller, C., Bondeau, a., Dietrich, J., Kurukulasuriya, P., Heinke, J., and Lotze-Campen, H.: Adaptation to climate change through the choice of cropping system and sowing date in sub-Saharan Africa, Global Environmental Change, 23, 130–143, doi:10.1016/j.gloenvcha.2012.11.001, 2013.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., and Schewe, J.: The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project framework, Proceedings of the National Academy of Sciences, 111, 3228–3232, doi:10.1073/pnas.1312330110, 2014.

Willmott, C. J.: Some comments on the evaluation of model performance, Bulletin American Meteorological Society, pp. 1309–1313, doi:10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2, 1982.

Zhu, Z., Bi, J., Pan, Y., Ganguly, S., Anav, A., Xu, L., Samanta, A., Piao, S., Nemani, R. R., and Myneni, R. B.: Global Data Sets of Vegetation Leaf Area Index (LAI)3g and Fraction of Photosynthetically Active Radiation (FPAR)3g Derived from Global Inventory Modeling and Mapping Studies (GIMMS) Normalized Difference Vegetation Index (NDVI3g) for the Period 1981 to 2011, Remote Sensing, 5, 927–948, doi:10.3390/rs5020927, 2013.

Zscheischler, J., Mahecha, M., Von Buttlar, J., Harmeling, S., Jung, M., Rammig, A., Randerson, J. T., Schöllkopf, B., Seneviratne, S. I., Tomelleri, E., Zaehle, S., and Reichstein, M.: Few extreme events dominate global interannual variability in gross primary production, Environmental Research Letters, 9, 035 001, doi:10.1088/1748-9326/9/3/035001, 2014.