Are terrestrial biosphere models fit for simulating the global land carbon sink?

Christian Seiler[1], Joe R. Melton[1], Vivek K. Arora[2], Stephen Sitch[3], Pierre Friedlingstein[4], Almut Arneth[5], Daniel Goll[6], Atul K. Jain[7], Emilie Joetzjer[8], Sebastian Lienert[9], Danica Lombardozzi[10], Sebastiaan Luyssaert[11], Julia E. M. S. Nabel[12], Hanqin Tian[13], Nicolas Vuichard[14], Anthony P. Walker[15], Wenping Yuan[16], Sönke Zaehle[17]

[1]Climate Processes Section, Environment and Climate Change Canada, Victoria, BC V8P 5C2, Canada
[2]Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, Victoria, BC V8P 5C2, Canada
[3]College of Life and Environmental Sciences, University of Exeter, Exeter EX4 4RJ, UK
[4]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, UK
[5]Karlsruhe Institute of Technology, Institute of Meteorology and Climate Research/Atmospheric Environmental Research, 82467 Garmisch-Partenkirchen, Germany
[6]Laboratoire des Sciences du Climat et de l'Environnement LSCE/IPSL, F-91191 Gif sur Yvette Cedex, France
[7]Department of Atmospheric Sciences, University of Illinois, Urbana, IL 61821, USA
[8]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France; now at INRAE, UMR1434 SILVA, Champenoux, France
[9]Climate and Environmental Physics, Physics Institute and Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland
[10]National Center for Atmospheric Research, Climate and Global Dynamics, Terrestrial Sciences Section, Boulder, CO 80305, USA
[11]Faculty of Science, Department of Ecological Science, Vrije Universiteit Amsterdam, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands
[12]Max Planck Institute for Meteorology, 20146 Hamburg, Germany
[13]School of Forestry and Wildlife Sciences, Auburn University, 602 Ducan Drive, Auburn, AL 36849, USA
[14]Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, 91198 Gif-sur-Yvette, France
[15]Climate Change Science Institute and Environmental Sciences Division, Oak Ridge National Lab, Oak Ridge, TN 37831, USA
[16]School of Atmospheric Sciences, Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies, Zhuhai Key Laboratory of Dynamics Urban Climate and Ecology, Sun Yat-sen University, Zhuhai, Guangdong 510245, China
[17]Max Planck Institute for Biogeochemistry, P.O. Box 600164, Hans-Knöll-Str. 10, 07745 Jena, Germany

Key Points:

- Poor model skill can result not only from model deficiencies but also from observational uncertainties.
- Although model performance is mostly reasonable, given how uncertain reference data are, ample potential for model improvements remains.
- The effectiveness of future model development depends on our ability to account for and reduce observational uncertainties.

Corresponding author: Christian Seiler, `christian.seiler@ec.gc.ca`

Abstract

The Global Carbon Project estimates that the terrestrial biosphere has absorbed about one-third of anthropogenic CO_2 emissions during the 1959-2019 period. This sink-estimate is produced by an ensemble of terrestrial biosphere models collectively referred to as the TRENDY ensemble and is consistent with the land uptake inferred from the residual of emissions and ocean uptake. The purpose of our study is to understand how well TRENDY models reproduce the processes that drive the terrestrial carbon sink. One challenge is to decide what level of agreement between model output and observation-based reference data is adequate considering that reference data are prone to uncertainties. To define such a level of agreement, we compute benchmark scores that quantify the similarity between independently derived reference datasets using multiple statistical metrics. Models are considered to perform well if their model scores reach benchmark scores. Our results show that reference data can differ considerably, causing benchmark scores to be low. Model scores are often of similar magnitude as benchmark scores, implying that model performance is reasonable given how different reference data are. While model performance is encouraging, ample potential for improvements remains, including a reduction in a positive leaf area index bias, improved representations of processes that govern soil organic carbon in high latitudes, and an assessment of causes that drive the inter-model spread of gross primary productivity in boreal regions and humid tropics. The success of future model development will increasingly depend on our capacity to reduce and account for observational uncertainties.

Plain Language Summary

Earth's natural vegetation absorbs about one-third of CO_2 emissions caused by human activities. This value is produced by a group of models rather than through direct observations. Our study assesses how well models reproduce the processes that drive the CO_2 exchange between land and atmosphere using a wide range of datasets that are mainly derived from field measurements and satellite images. These reference datasets are prone to errors that are not quantified in a consistent manner. To account for such errors, we first compare different reference datasets against each other. We then compare model output against reference data and assess whether the differences are comparable to the differences among the reference datasets. We conclude that the performance of models is encouraging given how uncertain reference data are, but that ample potential for improvements remains.

1 Introduction

Effective climate policies demand reliable estimates of global carbon fluxes and trends. The Global Carbon Project coordinates an annual publication on the Global Carbon Budget, which assesses and reports (i) CO_2 emissions from fossil fuel combustion and oxidation from all energy and industrial processes (E_{FOS}) and land use change (E_{LUC}), (ii) atmospheric CO_2 concentration growth rate (G_{ATM}), and (iii) the uptake of CO_2 by the ocean (S_{OCEAN}) and natural vegetation (S_{LAND}), all expressed in GtC yr^{-1} (Friedlingstein et al., 2020):

$$E_{FOS} + E_{LUC} = G_{ATM} + S_{OCEAN} + S_{LAND} + B_{IM}. \tag{1}$$

The components of the carbon budget are computed independently and the budget imbalance (B_{IM}) reflects the remaining uncertainty associated with imperfect spatial and/or temporal data coverage, observational errors, and omission of smaller terms. The land sink term S_{LAND} arises from the combined effects of CO_2 fertilization, nitrogen deposition, and climate change. Estimates for the 1959-2019 period show that anthropogenic CO_2 emissions associated with fossil fuel combustion (365 GtC) and land use change (85 GtC) are approximately balanced by the increase of atmospheric CO_2 (205 GtC) and the uptake of CO_2 by oceans (105 GtC) and land (145 GtC). The natural terrestrial ecosystems would have therefore absorbed about one-third of anthropogenic CO_2 emissions, which emphasizes the pivotal role of the terrestrial biosphere in the global climate system. Note that the values above are rounded to the nearest 5 GtC and B_{IM} is estimated to equal 0 GtC for this period.

The value for S_{LAND} is not based on direct observations, but on the mean value from an ensemble of terrestrial biosphere models (TBMs) collectively referred to as the trends in the land carbon cycle project (TRENDY) ensemble. Results from TRENDY simulations have been used extensively to explore different aspects of the global carbon cycle (e.g. Forzieri et al. (2018); Fernández-Martínez et al. (2019); Bastos et al. (2020); Kondo et al. (2020); Piao et al. (2020)). Friedlingstein et al. (2020) presented a brief assessment of model performance for key processes that are relevant for S_{LAND} (their Figure B2). Using a skill score system developed by the International Land Model Benchmarking Project (ILAMB; Collier et al. (2018)), the authors concluded that (i) TRENDY models show high skill scores for runoff, and to a lesser extent for vegetation biomass, gross primary productivity (GPP), and ecosystem respiration, and that (ii) skill scores are lowest for leaf area index (LAI) and net ecosystem exchange (NEE), with the widest disparity among models for soil organic carbon. The ILAMB skill scores summarize how well model output resembles reference data across multiple statistical metrics, including the bias, centralized root-mean square error, the timing of seasonal peaks, inter-annual variability, spatial correlation, and spatial variability (see section 2.4 for details).

One challenge of model evaluation is accounting for observational uncertainty. Observational uncertainty can be understood as an estimate characterizing the range of values within which the true value of a measurand, i.e. the quantity to be measured, lies (JCGM, 2008). Any measurement consists of a series of transformations from the event observed to the final value, and each transformation may introduce and propagate errors (Merchant et al., 2017). For instance, sources of uncertainty in satellite LAI products include uncertainties in the input data (e.g. surface reflectance, radiance, albedo, land cover type), the radiative transfer model, the inversion technique, and the prior information (Fang et al., 2012). Unfortunately, observational uncertainty is not reported consistently among reference datasets (Merchant et al., 2017). To account for observational uncertainty nevertheless, a pragmatic and common approach is to evaluate model output against multiple reference datasets per variable, which may underestimate uncertainty if reference data are not sufficiently independent and overestimate uncertainty if one reference dataset is strongly inferior compared to others (Covey et al., 2002). The ILAMB framework addresses observational uncertainty by using multiple reference datasets that are weighed depending on their estimated quality and spatiotemporal coverage (Collier et al., 2018). However, the ILAMB approach does not indicate what score a model should actually yield given how uncertain reference data are. This makes the interpretation of the ILAMB scores challenging, as it remains unclear to what extent low scores are related to observational uncertainty. The purpose of our study is to evaluate how well TBMs reproduce processes that drive the terrestrial carbon sink term S_{LAND}. As a novel contribution, we will demonstrate how well models should score given that reference data are imperfect.

2 Methods

2.1 Simulation protocol

The TRENDY model ensemble consists of a variety of terrestrial ecosystem models intended for climate simulations. Some TRENDY models are characterized as land surface models (LSMs), which were initially developed to simulate land-atmosphere fluxes of mass, energy, and momentum required as inputs for the atmospheric component of global climate models. Other TRENDY models are dynamic global vegetation models (DGVMs), which were designed to simulate terrestrial carbon pools and fluxes, as well as biogeography and plant demography. To represent carbon cycle dynamics in global climate models, model developers have begun to incorporate DGVMs into LSMs in the early 2000s (Fisher & Koven, 2020). In this paper we use the more general term Terrestrial Biosphere Models (TBMs; G. Bonan (2019)) to describe all TRENDY models regardless of their original purpose. Model results evaluated in this study form part of TRENDY version 9, which was used for quantifying the global carbon budget of 2020 (Friedlingstein et al., 2020). We selected 15 TBMs for which most variables were available at the time of writing (Table 1).

TRENDY models are run for three simulations that are designed to disentangle the role of changes in CO_2, climate, as well as land-use and land-cover change (LULCC). The first simulation (S1) is driven by time-varying atmospheric CO_2 concentration but land cover state is fixed for the year 1700 and repeating climate is used from the period 1901-1920. The S1 simulation is designed to infer the effect of increasing atmospheric CO_2. The second simulation (S2) is driven with increasing CO_2 concentrations and climate varying in time, but keeps the land cover state fixed to its pre-industrial state of 1700. Finally, in the third simulation (S3) all forcings (CO_2, climate, and LULCC) are time varying. Models with a coupled carbon-nitrogen cycle are also forced with historical nitrogen deposition (S1, S2, S3), pre-industrial nitrogen fertilization (S1, S2) and historical nitrogen fertilization (S3). Our study only assess results for S3, as S1 and S2 are counterfactual.

The term S_{LAND} in equation 1 corresponds to the net biome productivity (NBP) in the S2 simulation, where NBP equals gross primary productivity minus ecosystem respiration minus CO_2 fluxes associated with disturbance. The S_{LAND} term is a counterfactual value that represents the strength of the terrestrial carbon sink under pre-industrial land cover had land use change not taken place. Given the hypothetical nature of global S_{LAND}, we cannot evaluate it against observations. However, we can evaluate NBP, and the processes that drive it, in the S3 experiment where CO_2, climate, and LULCC forcings all vary in time. The variable NBP under S3 approximates S_{LAND} (3.4 GtC yr^{-1} with a standard deviation of $\pm$0.9 GtC yr^{-1}) minus E_{LUC} (1.6$\pm$0.7 GtC yr^{-1}). Note that E_{LUC} values can be obtained from TBMs or, as for the Global Carbon Budget, from bookkeeping models (BLUE, HandN2017, and OSCAR) (Friedlingstein et al., 2020).

The S3 TRENDY simulation protocol (version 9) consists of a preindustrial spin up for the year 1700 and two transient runs for the periods 1701-1900 and 1901-2019, respectively (Friedlingstein et al., 2020). The preindustrial spin up uses a constant atmospheric CO_2 concentration of 276.59 ppm, repeating climate data from the early decades of the 20th century (i.e. 1901-1920), and land cover that uses crops and pasture distribution corresponding to the year 1700. Since TBMs use different sets of plant functional types (PFTs) their land covers are different although they are all expected to represent the crop and pasture distribution using the specified common LULCC forcing. The first transient run for the 1701-1900 period uses the same climate as for the spin up, but time-varying CO_2 concentrations and land cover. The second transient run uses time-varying CO_2, climate, and land use for the 1901-2017 period. Note that the two transient runs are typically combined in a single run, where meteorological data from the 1901-1920 period are repeatedly used during the 1701-1900 period. Meteorological inputs required by TRENDY models may include surface downwelling shortwave and longwave radiation, near-surface air temperature, precipitation, near-surface specific humidity, surface pressure, and near-surface horizontal wind speed. Models were forced by either the merged monthly Climate Research Unit (CRU) and 6-hourly Japanese 55-year Reanalysis (JRA-55) data or by the monthly CRU data (Harris et al., 2014; Kobayashi et al., 2015). The LULCC forcing was given by the Land-Use Harmonization 2 (LUH2) dataset (Hurtt et al., 2020). For the purpose of our study, all S3 model outputs were spatially interpolated to a common resolution of $1^\circ \times 1^\circ$ using bilinear interpolation. In the case of the Canadian Land Surface Scheme Including Biogeochemical Cycles (CLASSIC; Table 1), we reran the model at the $1°\times1°$ resolution rather than spatially interpolating the original $2.8125°\times 2.8125°$ grid.

2.2 In situ reference data

In situ reference data include the variables gross primary productivity (GPP), ecosystem respiration (RECO), net ecosystem exchange (NEE), vegetation carbon (CVEG), leaf area index (LAI), latent heat flux (HFLS), and streamflow (Table 2). The variable NEE is defined as RECO minus GPP, such that negative NEE values imply a net land carbon sink. In situ observations that fell into the same model grid cell were averaged prior to the comparison against model output. In situ reference data are compared against model output at the grid cell level. An evaluation that accounts for the presence of particular plant functional types at a site would have been desirable, but most model data were reported on a grid cell level only. All comparisons are conducted for locations and time steps that models and reference data have in common. Time-invariant reference data (vegetation carbon) were compared against model output averaged from 1980 to 2019. Details on each in situ reference dataset are provided next.

The FLUXNET2015 database includes 204 eddy covariance sites with measurements made sometime during the 1997-2014 period (Pastorello et al., 2020) (Table 2; Figure Appendix B1a). The corresponding variables are GPP, ecosystem respiration, NEE, and latent heat flux. Only sites with at least 3 years of data were considered. We assessed NEE using two versions of the FLUXNET2015 database. The first version uses all available sites with at least 3 years of data. This dataset was then filtered for sites that were located in forests where no disturbance occurred over the last 50 years as documented in Besnard et al. (2018) and for months that have $\geq 95\%$ of high quality data. The first and second version of this reference dataset is here referred to as NEE-FLUXNET and NEE-FLUXNETB, respectively.

Aboveground biomass measurements were obtained from two datasets. The first database consists of 1974 measurements that were compiled from literature by Xue et al. (2017). The second database consists of 1645 measurements from 274 sites provided by the Forest Observation System (Schepaschenko et al., 2019). We merged both datasets and replaced Xue et al. (2017) with the more recent Schepaschenko et al. (2019) data when a site was present in both datasets. We then converted aboveground biomass to total vegetation carbon using an empirical relation between root biomass y and shoot biomass x ($y = 0.489 \times x^{0.890}$) (Mokany et al., 2006), as well as a carbon-to-biomass ratio of 0.5. It must be noted that empirical data on root-shoot ratios are likely to be subject to a sampling bias towards smaller rather than larger trees, as the former are easier to excavate (Huang et al., 2021). Since root-shoot ratios tend to be larger for smaller trees, this sampling bias may result in an overestimation of root-shoot ratios. The conversion from aboveground biomass to total vegetation carbon was necessary as the TRENDY dataset provides only total biomass without separation into below and aboveground components. Measurements located within the same model grid cells were averaged, leading to a total of 592 grid cells with at least one in situ measurement (Figure Appendix B1c).

LAI observations were taken from the Committee on Earth Observation Satellites (CEOS) which consists of 141 sites with monthly measurements during the 1999-2017 period (Figure Appendix B1b) (Garrigues et al., 2008). The values are based on a transfer function that upscales ground LAI measurements to a moderate resolution grid cell using high spatial resolution surface reflectances.

Annual stream flow gauge records were obtained from the Global Runoff Data Centre (GRDC) for the world's 50 largest basins (Figure Appendix B1d) (Dai & Trenberth, 2002). Measurements were made some time between 1980 and 2010, depending on the basin.

2.3 Globally gridded reference data

Globally gridded reference datasets include the variables GPP, NBP, vegetation carbon, soil organic carbon, LAI, latent heat flux, and runoff. The variable NBP is defined as GPP minus RECO minus CO_2 emissions associated with disturbance and LULCC, such that positive NBP values imply a net land carbon sink. All gridded reference data were spatially interpolated to a common resolution of $1^\circ \times 1^\circ$ using bilinear interpolation. All comparisons are conducted for grid cells and time steps that models and reference data have in common. Time-invariant reference data (vegetation carbon and soil organic carbon) were compared against model output averaged from 1980 to 2019. Details on each globally gridded reference dataset are provided next.

2.3.1 Gross primary productivity

We used three different globally gridded GPP reference datasets. The first dataset is based on satellite imagery from the Moderate Resolution Imaging Spectroradiometer (MODIS) for the period 2000-2016 (Zhang et al., 2017). The dataset estimates GPP as the product of light absorption by chlorophyll and the efficiency that converts the absorbed energy to carbon fixed by plants through photosynthesis. The required inputs to the Zhang et al. (2017) algorithm include a range of MODIS products (surface temperature, land surface water index, enhanced vegetation index, and land cover classification), as well as air temperature and radiation fluxes from NCEP Reanalysis II (Kanamitsu et al., 2002).

The second reference GPP data, referred to as GOSIF, consists of solar-induced chlorophyll fluorescence (SIF) soundings from the global Orbiting Carbon Observatory-2 (OCO-2). The dataset is based on a linear correlation between SIF soundings and GPP measurements from 91 eddy covariance measurements sites from FLUXNET for the period 2000-2017 (Li & Xiao, 2019).

The third GPP reference data, referred to as FluxCom, is based on a variety of machine-learning algorithms that upscale eddy covariance data using remote sensing data and meteorological data as global predictors (Tramontana et al., 2016; Jung et al., 2020). Remote sensing data employed by FluxCom include land surface temperature (LST; MOD11A226), land cover (MCD12Q127), fraction of absorbed photosynthetically active radiation by a canopy (fPAR; MOD15A228), and bidirectional reflectance distribution function (BRDF)-corrected reflectances (MCD43B429) from MODIS. Meteorological inputs for FluxCom were taken from the Climate Research Unit National Centers for Environmental Prediction version 8. The FluxCom values used in our study are the median values computed over a FluxCom ensemble for the 1980-2013 period. The GPP FluxCom ensemble consists of six ensemble members that vary with respect to the employed machine learning algorithm (Artificial Neural Network, Multivariate Adaptive Regression Splines, and Random forest) and partitioning method (Lasslop et al., 2010; Reichstein et al., 2005). It should be noted that neither the satellite based GPP estimates nor the FluxCom product explicitly account for the CO_2 fertilization effect, which compromises the respective carbon flux trends (De Kauwe et al., 2016; Jung et al., 2020).

2.3.2 Net biome productivity

Globally gridded reference NBP was obtained from the three inversion models Copernicus Atmosphere Monitoring Service (CAMS) (Chevallier, 2013), the Jena CarboScope (Rödenbeck et al., 2018), and CarbonTracker 2019 (CT2019) (Jacobson et al., 2020). Inversion models attempt to reproduce observed atmospheric CO_2 concentrations by adjusting CO_2 fluxes at the surface. This process requires an atmospheric transport model and *apriori* estimates of surface CO_2 fluxes. The prior fluxes are usually derived from TBMs. For CAMS, atmospheric CO_2 concentrations are taken from 81 sites provided by the National Oceanic and Atmospheric Administration (NOAA) Earth System Research Laboratory archive. The inversion is based on the global atmospheric transport model Laboratoire de Météorologie Dynamique (LMDZ) and covers the period 1979-2019 (Hourdin et al., 2006). Land-atmosphere fluxes are based on priors from the Organizing Carbon and Hydrology in Dynamic Ecosystems (ORCHIDEE) (Krinner et al., 2005) and GFED wild fire emissions. CO_2 emissions from wild fires are compensated by the same annual flux of opposite sign representing the regrowth of burnt vegetation.

The second inversion-based NBP estimate from Jena CarboScope (Run ID s99oc v2020) uses 48 CO_2 measurement sites mostly from NOAA (Rödenbeck et al., 2018). The atmospheric transport is simulated by the Transport Model 3 (TM3) for the period 1999-2019. As for CAMS, the land CO_2 flux of Jena CarboScope represents the net flux resulting from GPP, ecosystem respiration, and disturbances, such as wild fires and LULCC. While Rödenbeck et al. (2018) refer to the Jena CarboScope land CO_2 flux as NEE, we refer to it as NBP, as it includes the effects of disturbances and LULUC.

The third inversion-based NBP estimate from CT2019 uses 460 CO_2 measurement sites provided by the GLOBALVIEW+ data product version 5.0 (Masarie et al., 2014). The transport model employed by CT2019 is the Transport Model 5 (TM5), which is run for the period 1999-2019 (Huijnen et al., 2010). The *apriori* land-atmosphere fluxes are taken from the Carnegie-Ames Stanford Approach (CASA) biogeochemical model (Potter et al., 1993). Carbon emissions from fires are prescribed from the Global Fire Emissions Database (GFED) (van der Werf et al., 2017), and are not modified by the optimization process.

2.3.3 Vegetation carbon

We used three globally gridded and time-invariant vegetation carbon reference datasets. Two of the three datasets originally consisted of aboveground biomass. As for our in situ measurements, we converted aboveground biomass to vegetation carbon using the empirical relation between root biomass and shoot biomass provided by Mokany et al. (2006). Again, this was necessary as most TRENDY models only reported total rather than aboveground biomass.

The first reference dataset, here referred to as GEOCARBON-Mokany, integrates local high-quality biomass data with a boreal forest biomass map by Santoro et al. (2015) and a pan-tropical biomass map by Avitabile et al. (2016), which is based on data from Saatchi et al. (2011) and Baccini et al. (2012). The dataset covers only areas that are dominated by trees in the Global Land Cover 2000 map (Bartholome & Belward, 2005). The boreal biomass estimates are based on radar imagery provided by the Envisat Advanced Synthetic Aperture Radar (ASAR). The pan-tropical biomass maps are based on Light Detection and Ranging (LiDAR) observations that were calibrated with in situ measurements of tree allometry. Baccini et al. (2012) upscaled data using a random forest machine learning algorithm and satellite imagery, including the MODIS Nadir BRDF-Adjusted Reflectance (NBAR), MODIS land surface temperature, and shuttle radar topography mission (SRTM) digital elevation data.

The second vegetation carbon reference dataset, here referred to as Zhang-Mokany, was obtained from Zhang and Liang (2020), who integrated ten existing local and global aboveground biomass maps using a data fusion technique. It must be noted that one of the ten maps is the pan-tropical biomass map by Avitabile et al. (2016), which also forms part of the above-mentioned dataset by Santoro et al. (2015). Zhang and Liang (2020) evaluated each of the ten datasets against in situ observations and high-resolution airborne lidar data.

The third vegetation carbon dataset was obtained by Huang et al. (2021), who upscaled in situ measurements of root biomass using a machine learning algorithm (Random Forest) and globally gridded predictors of shoot biomass, tree height, soil properties, and climatological data. The shoot biomass presented by Huang et al. (2021) was derived from the above ground biomass by Santoro et al. (2021). Adding root and shoot mass, and converting biomass to carbon mass using a carbon-to-biomass ratio of 0.5, we obtained a globally gridded dataset for vegetation carbon associated with trees.

2.3.4 Soil organic carbon

Reference data for soil organic carbon in the top 100 cm were obtained from the Harmonized World Soil Database (HWSD) (Wieder, 2014) and from SoilGrids250m (SG250m) (Hengl et al., 2017). The HWSD data provided by the Food Agriculture Organization (FAO) combines existing regional and national updates of soil information worldwide with the information contained by the FAO Soil Map of the World (Wieder, 2014). The values correspond to the top 100 cm soil depth. The SoilGrids250m (SG250m) dataset provides a globally gridded dataset of soil organic carbon at various depths between the surface and 200 cm belowground. The estimates are produced by an ensemble of machine learning methods that used 150,000 soil profiles and 158 remote sensing-based soil covariates. Our study considers only the top 100 cm to ensure that the values are comparable to estimates from the HWSD dataset. It must be noted that both reference datasets differ considerably, with lower values in HWSD compared to SG250m, in part due to a poor representaion of wetlands and permafrost soils in HWSD Tifafi et al. (2018).

2.3.5 Leaf area index

We used three globally gridded reference LAI that are derived from satellite imagery. MODIS LAI (MOD15A2H, collection 6) (R. Myneni et al., 2015) is based on the inversion of a three dimensional canopy radiative transfer model that simulates surface reflectance from canopy structural characteristics (Knyazikhin et al., 1998).

A second LAI reference dataset was provided by Claverie et al. (2016) for the period 1982-2010. This dataset is based on an artificial neural network that relates LAI to surface reflectance from the Advanced Very High Resolution Radiometer (AVHRR). The artificial neural network was calibrated with LAI from MODIS (MCD15A2) and in situ data from BELMANIP2 (445 sites) (Baret et al., 2006). The performance of the algorithm was assessed against in situ observations from the DIRECT database (113) (Garrigues et al., 2008).

A third LAI dataset was provided by the Copernicus Global Land Service for the period 1999-2019 (Verger et al., 2014). This product uses an artificial neural network that gives instantaneous estimates from reflectances by SPOT/VEGETATION satellite imagery. The data are filtered to reduce the impacts of atmospheric effects and snow cover, temporally smoothed, and gap-filled. For the purpose of this study only non-gap filled grid cell values were used.

2.3.6 Latent heat flux and runoff

We used two globally gridded reference latent heat flux datasets. The first dataset provided by FluxCom covers the period 2001-2013 (Jung et al., 2019). As for GPP, FluxCom upscales FLUXNET observations, where remote sensing data and meteorological data serve as global predictors. Our study uses median values from 36 FluxCom ensemble members that vary with respect to the employed meteorological forcing (Climate Research Unit National Centers for Environmental Prediction version 8, WATCH Forcing Data ERA Interim, the Global Soil Wetness Project 3, and Clouds and the Earth's Radiant Energy System in combination with the Global Precipitation Climatology Project), the machine learning algorithm (Artificial Neural Network, Multivariate Adaptive Regression Splines, and Random forest), and the energy balance closure correction (none, Bowen ratio correction and residual approach).

Our second reference dataset was taken from the Conserving Land-Atmosphere Synthesis Suite (CLASSr), which covers the period 2003-2009 (Hobeichi et al., 2019). The CLASSr provides estimates of simultaneously balanced surface water and energy budget components. Each variable presents a weighted mean computed from multiple data products that are, to some extent, observation-based. The data are observationally constrained with in situ measurements, and each term is adjusted to allow for energy and water balance closure. Latent heat flux provided by CLASSr is based on blending data from remote sensing, reanalysis, and TBMs.

The CLASSr dataset described above also provides monthly runoff. The values are based on 11 runoff estimates from eight hydrological models that are constrained by observational streamflow records from around 600 downstream stations. To obtain benchmark scores for streamflow we converted monthly CLASSr runoff to annual streamflow for the earth's 50 largest river basins and compared annual values against gauge measurements from GRDC.

2.4 Automated Model Benchmarking R package (AMBER)

The Automated Model Benchmarking R package developed by Seiler (2020) quantifies model performance using a skill score system that is based on the ILAMB framework (Collier et al., 2018). The method employs five scores that assess the model's annual mean bias (S_{bias}), monthly centralized root-mean-square-error (S_{rmse}), the timing of the seasonal peak (S_{phase}), inter-annual variability (S_{iav}), and spatial distribution (S_{dist}). The exact definition of each skill score is provided in Appendix A. The main steps for computing a score usually include (i) computing a dimensionless statistical metric, (ii) scaling this metric onto a unit interval, and (iii) computing a spatial mean. All scores are dimensionless and range from zero to one, where increasing values imply better performance. These properties allow us to average skill scores across different statistical metrics in order to obtain an overall score for each variable ($S_{overall}$) (Collier et al., 2018):

$$S_{overall} = \frac{S_{bias} + 2S_{rmse} + S_{phase} + S_{iav} + S_{dist}}{1 + 2 + 1 + 1 + 1}. \qquad (2)$$

To reward models that reproduce a realistic response to changes in the meteorological forcing, we increase the weight of S_{rmse} by a factor of two. In the case of GPP FluxCom we assign S_{iav} a weight of zero, since the reference data are known to underestimate interannual variability (Jung et al., 2020).

Model scores are calculated by comparing model output against observation-based reference data (Figure 1). Benchmark scores are computed by comparing multiple reference datasets of a variable among each other. The purpose of benchmark scores is to quantify the similarity between equally plausible reference datasets, which indicates what level of agreement between model output and reference data can be expected, given how uncertain reference data are. For instance, consider the three inversion-based NBP reference datasets CAMS, CT2019, and CarboScope. Comparing CT2019 using CAMS as a reference yields an overall score ($S_{overall}$) of 0.57. Comparing CarboScope using CAMS as a reference yields an $S_{overall}$ value of 0.56. The benchmark score is then chosen to equal the minimum of both scores (0.56), which accounts for the full uncertainty range. This benchmark score only applies when using CAMS as reference data. Using CT2019 or CarboScope as reference data may yield different benchmark scores for the following reason. Recall that evaluating CT2019 using CAMS as a reference data yields an overall score $S_{overall}$ of 0.57. Evaluating CAMS using CT2019 as a reference data, on the other hand, yields an $S_{overall}$ value of 0.58. The difference arises due to the normalization of a statistical metric. In the case of S_{bias}, the bias is divided by the standard deviation of the reference data σ_{ref} (Equation A2). If we evaluate CT2019 using CAMS as a reference, the value of σ_{ref} is given by CAMS, and if we evaluate CAMS using CT2019 as a reference, the value of σ_{ref} is given by CT2019. We can therefore have different benchmark scores for different reference datasets for the same variable in question.

The final benchmarking step in Figure 1 consists of comparing model scores against benchmark scores. If model scores reach benchmark scores, then the degree of similarity between model output and reference data is the same as between two independent reference datasets. Using this criteria, we then judge models to perform sufficiently well, given how uncertain reference data are. Note that model scores may also exceed benchmark scores when, for instance, model values are enclosed by the uncertainty range span by two or more reference data. All AMBER outputs for TRENDY are available at `https://cseiler.shinyapps.io/TRENDY2020/` (last visited on November 22, 2021).

3 Results

3.1 Gross primary productivity and ecosystem respiration

Reference data estimate global annual GPP fluxes to range from 108.9 (FluxCom) to 123.8 PgC yr^{-1} (GOSIF; Table 3). The corresponding TRENDY multi-model mean values lie within this uncertainty range, with values ranging between 115.0 and 119.3 PgC yr^{-1}, depending on the choice of reference data. The multi-model mean values vary with the choice of reference data, because all comparisons are conducted for grid cells and time steps that models and reference data have in common. If the spatiotemporal coverage varies among reference data, so do the multi-model mean values. In relative terms, the mean bias across models ranges from -6% when evaluating models against GOSIF and +6% when choosing FluxCom as reference data. The biases of the individual models range between -27% and +25%, with 7/15 models lying within the uncertainty range of the reference data. Note that differences between reference values, listed in Table 3, may be caused by differences in the observational period and grid. Although all reference data are regridded to a common horizontal resolution of $1^{\circ}\times1^{\circ}$, datasets may still differ with respect to the distribution of grid cells with missing data. Reducing reference data to a common period and identical grid leads to similar results, with 5/15 models within the uncertainty range of global mean values, which is depicted in Figure 2a).

Zonal mean values are well reproduced, but the inter-model spread is large, with values ranging from 5 to 10 gC m^{-2} day^{-1} at the equator (Figure 2a). The models reproduce the seasonal GPP cycle well across regions, with a tendency to overestimate the GPP amplitude in the boreal region of North America and Eurasia (Figure 3). Two models with particularly large positive biases in the boreal regions are LPX-Bern and CLM5.0. This bias is confined to the boreal regions and does not extend across the globe. Evaluations against FLUXNET data confirm that both models simulate larger-than-observed GPP values in boreal regions (Figure B2 e and l). GPP benchmark scores for globally gridded data equals 0.72, and multi-model mean scores range between 0.61 and 0.64 (Figure 4). None of the models reach GPP benchmark scores, but some come close with model scores of 0.70 (ISAM, ORCHIDEE, and SDGVM).

Concerning ecosystem respiration, our evaluation relies on in situ measurements only. This is because the currently available gridded reference datasets, which rely on spatially upscaled eddy covariance measurements, yield results that are inconsistent with inversion-based estimates in the tropics (Jung et al., 2020). Evaluating modeled ecosystem respiration against FLUXNET data shows that annual mean values are reasonably well reproduced with correlation coefficients ranging between 0.44 (ORCHIDEE-CNP) and 0.75 (ISBA-CTRIP) (Figure B3). The corresponding overall score values are similar to the GPP scores for FLUXNET data, with a multi-model mean score value of 0.62 for both ecosystem respiration and GPP (Figure 4). Note that we did not compute ecosystem respiration benchmark scores as we lack a second reference dataset.

3.2 Net ecosystem exchange

Evaluating modeled NEE against FLUXNET data shows no correlation for annual mean values (Figure B4). Annual mean FLUXNET NEE values range from -4.8 to +2.0 gC m^{-2} day^{-1}, with a mean value of -0.6 gC m^{-2} day^{-1}. Modeled values cover a smaller NEE range from -1.3 to +0.4 gC m^{-2} day^{-1} with a mean value of -0.2 gC m^{-2} day^{-1}. The apparent mismatch between modeled and observed values could be due to a variety of reasons. First, grid cell values represent a much larger region compared to eddy covariance measurements. Second, the globally gridded data are not necessarily representative of the actual meteorological conditions at the site level. Third, models do not reproduce the disturbance history of FLUXNET sites. And fourth, gap-filling observations may have reduced data quality. To address at least the last two issues, we filtered FLUXNET data for sites with mature forests and for months that have 95% of high quality data (here referred to as FLUXNETB, see section 2.2). Evaluating models against high-quality sites located in mature forests improves the correlation between models and observations, with correlation coefficients reaching up to 0.69 (Figure 5). However, the modeled NEE ranges are still substantially smaller compared to the observations. This also holds true when considering only CO_2 fluxes associated with tree PFTs (not shown, and tested for CLASSIC only due to data availability). Looking at model scores for each site shows that models perform best for sites that present modest sinks, with NEE values of -0.5 gC m^{-2} day^{-1}. The multi-model mean score improves from 0.48 to 0.55 when comparing modeled NEE against FLUXNET and FLUXNETB, respectively (Figure 4). This improvement is mainly due to an increase in the model score associated with the spatial distribution (S_{dist}). As for ecosystem respiration, we did not compute NEE benchmark scores as we lack a second reference dataset.

3.3 Net biome productivity

Inversion models estimate a net CO_2 sink with a global NBP that ranges between 1.3 PgC yr^{-1} for CarboScope (1999-2019) and CT2019 (2000-2017) and 1.9 PgC yr^{-1} for CAMS (1979-2019) (Table 3). About half of the models (7/13) lie within the NBP uncertainty range (ISBA-CTRIP, JSBACH, OCN, ORCHIDEE, ORCHIDEEv3, SDGVM, VISIT), with a multi-model mean value that is in closer agreement with CarboScope and CT2019 than with CAMS (Table Appendix B).

The zonal mean NBP of CAMS, CarboScope, and CT2019 show very little agreement, with opposing signs in multiple regions (Figure 2b). TRENDY models do not reproduce the zonal mean values of either reference dataset. The only region with some reasonable agreement between both reference datasets and models is the tendency for a carbon sink between 50°N and 65°N. Averaging NBP values across every 30 degrees latitude shows that models and reference data agree on a stronger sink in higher latitudes compared to the tropics (Figure 2c).

All three reference datasets show a very similar global seasonal cycle, with a net carbon source during the NH winter and a net carbon sink during the NH summer (Figure 6). While the seasonal cycle of the multi-model mean is in reasonable agreement with the reference data, the inter-model spread can be large. For instance, model values in the boreal region range between 0 and 2 gC m^{-2} day^{-1} during summer (Figure 6a and g). Multi-model mean scores (0.50-0.53) and benchmark scores (0.52, 0.56) are similar, with six models reaching benchmark scores (IBIS, ISAM, ISBA-CTRIP, ORCHIDEE, ORCHIDEEv3 and VISIT; Figure 4).

3.4 Vegetation carbon

The amount of vegetation carbon stored in forested regions on a global scale varies strongly among reference data, with 264.6 PgC for Geocarbon-Mokany, 310.2 PgC for

Huang2021, and 482.5 PgC for Zhang-Mokany (Table 3). As a comparison, global vegetation carbon estimates for all biomes reported by Friedlingstein et al. (2020) range from 450 to 650 PgC. This range is taken from the 5^{th} Assessment Report of the Intergovernmental Panel on Climate Change (AR5) (Ciais et al., 2013), which cites the 3^{rd} Assessment Report (AR3) (Houghton et al., 2001). The values in AR3 are based on data provided by Dixon et al. (1994) (466 PgC) and Roy et al. (2001) (654 PgC). The corresponding range for vegetation biomass in forests only is 359-539 PgC (Houghton et al., 2001), which is larger compared to the range reported in our study. The multi-model mean value (403.3-429.2 PgC) lies within the observational uncertainty range (Table 3 and Figure 2c). The biases of the individual models range between -35% and +109%, with 10/15 models that are within the uncertainty range.

The zonal mean values tend to be largest for Zhang-Mokany followed by Huang2021 and GEOCARBON-Mokay (Figure 2d). The Zhang-Mokany dataset is in stronger agreement with forest inventory data ($S_{overall} = 0.76$) than the Huang2021 ($S_{overall} = 0.69$) or the Geocarbon-Mokany dataset ($S_{overall} = 0.68$). All three tend towards a negative bias, with a larger bias for Geocarbon-Mokany (-57%) than for Huang2021 (-38%), and Zhang-Mokany (-26%), suggesting that the latter is likely to provide more accurate values, at least for regions where forest inventory data are present (Figure B5). It must be noted that this comparison is limited by the fact that the three data sets Geocarbon-Mokany, Zhang-Mokany, and FOSXue all use the same approach for estimating belowground biomass, which makes them more similar by construction.

Multi-model zonal mean values are in closer agreement with data from Zhang-Mokany compared to Huang2021 and Geocarbon-Mokany. All models tend towards a negative bias when assessed against forest inventory. Benchmark scores (0.62-0.74) and multi-model mean scores (0.60-0.69) are similar, where 6/15 models meet benchmarks when evaluated against in situ measurements (CLM5.0, ISAM, ISBA-CTRIP, JSBACH, OCN, SDGVM), and 5/15 models reach benchmarks when assessed against Geocarbon-Mokany (Figure 4).

3.5 Soil organic carbon

The global soil organic carbon pool in the top 100 cm is estimated to range between 1143 PgC (HWSD) and 2708 PgC (SG250m). The larger values in SG250 are found across all latitudes, but differences are particularly large at high latitudes (50-80°N) as well as the equator associated with differences in SE Asia (Table 3 and Figure 2e). As a comparison, the global soil carbon pool reported by Friedlingstein et al. (2020) is estimated to range from 1500 to 2400 PgC. This range is taken from AR5 (Ciais et al., 2013), and is based on a global soil carbon map developed by Batjes (1996), who estimate a soil organic carbon pool of 1462-1548 PgC in the upper 100 cm and 2376-2456 PgC in the upper 200 cm.

Models are in much closer agreement with HWSD (-3% mean bias) than with SG250m (-57% mean bias), with 5/15 models showing values that are within the observational uncertainty range (Table 3 and Figure 2d). Zonal multi-model mean values are in close agreement with HWSD, lacking the large increase of soil organic carbon at higher latitudes present in SG250m (Figure 2e). The model CLM5.0 was excluded from Figure 2e, as it produces zonal mean values that exceed 200 kgC m^{-2}, dwarfing values from all other datasets. The top three models with largest soil organic carbon stocks are CLM5.0 (3139 PgC), LPX-Bern (1838 PgC), and ISBA-CTRIP (1549 PgC), all of which include processes required for simulating carbon dynamics in permafrost regions (Table 1).

Due to the large differences between HWSD and SG250m, the benchmarking values are very small (0.33-0.42). All models but CLM5.0 therefore exceed the benchmark when assessed against HWSD. However, this result must be interpreted with caution. The large discrepancy between HWSD and SG250m suggests that the datasets have fundamental differences, possibly related to a poor representation of wetlands and permafrost soils in HWSD (Tifafi et al., 2018). It is therefore likely that SG250m is more accurate than HWSD, which implies that the difference between HWSD and SG250m overestimates the true observational uncertainty.

3.6 Leaf area index

Remotely sensed estimates of LAI yield very similar global mean values, ranging from 1.4 to 1.5 m^2 m^{-2} (Table 3 and Figure 2f). The multi-model mean value exceeds the observational uncertainty range by up to 67%, with biases from individual models between -4% and +220%. Only one model (ORCHIDEE-CNP) is within the uncertainty range, while all other models (13/14) show positive biases for all three global reference data.

Zonal mean values of annual mean LAI are very similar among all three reference datasets (Figure 2e). The multi-model zonal mean values reproduce the general pattern of the reference data, with a positive bias of up to 2 m^2 m^{-2} across most latitudes. Individual ensemble members can have very large biases of up to 7 m^2 m^{-2} at the equator. The tendency for a positive LAI bias is evident for all regions and seasons (Figure 7). The seasonal peak of maximum LAI tends to lag behind the reference data by about one month in the boreal and temperate regions. Also, the model IBIS lacks a seasonal LAI cycle in the tropics.

Comparing satellite-based LAI against in situ measurements from CEOS suggests that global reference data tend towards a negative bias ranging between -0.2 m^2 m^{-2} (-10%) for Copernicus to -0.4 m^2 m^{-2} (-19%) for MODIS when evaluated against data from CEOS. This leads to the question whether the positive LAI of TBMs described above is due to an underestimation of LAI in satellite-based reference data. Comparing modelled LAI against the same in situ data yields far greater biases for multiple models, most notably for the models IBIS (+71%), LPX-Bern (+144%), and OCN (85%; Figure 8 g, k, l). Furthermore model biases derived from globally gridded reference data and in situ data are correlated (R = 0.95) and of similar magnitude. For instance, a model with a large bias with respect to globally gridded reference LAI (LPX-Bern, 154% with respect to Copernicus) also has a large bias when assessed against in situ measurements (144% with respect to CEOS). Conversely, a model with a small bias with respect to globally gridded reference LAI (ORCHIDEE-CNP, -4%) also has a small bias when assessed against in situ data (1%). This suggests that the positive LAI bias present in some models is real, and not just due to an underestimation of LAI in satellite products. However, it must be noted that the evaluation against CEOS data is limited by the fact that sampling size varies substantially among regions, with the largest sampling density located in Europe. While none of the models reaches benchmarks for globally gridded reference LAI (0.65-0.66), 5/15 models reach the benchmark for in situ data (0.66; CLM5.0, ISBA-CTRIP, ORCHIDEE, ORCHIDEE-CNP, and ORCHIDEEv3) (Figure 4).

3.7 Latent heat flux

Global fluxes of annual mean latent heat from CLASSr and FluxCom range from 32.6 to 45.2 W m^{-2} (Table 3 and Figure 2f). The multi-model mean value, as well as the values from most individual models (14/15), lie within the observational uncertainty range. FluxCom values exceed CLASSr values across all latitudes. The inter-quartile range of models reproduces zonal patterns well, mostly within the observational uncertainty range. However, considerable inter-model spread remains in the tropics, where zonal mean latent heat fluxes range between 70 and 120 W m^{-2} at the equator, confirming previous findings from Pan et al. (2020). Multi-model mean values reproduce the seasonal cycle well, but the inter-model range is very large in the tropical parts of South America and Asia (Figure B6). The large inter-model spread is also present at the site-level, where annual mean biases across all sites range from -31% (LPX-Bern) to +20% (JSBACH) (Figure B7).

The multi model mean scores (0.67 and 0.70 when assessed against FluxCom and CLASSr, respectively) exceed the benchmark scores for globally gridded and site-level reference data (0.62-0.67). Most of the individual models reach the benchmark scores, suggesting that most models perform well given how uncertain current reference data are. One exception is JSBACH with a systematic positive bias across all regions and seasons.

3.8 Runoff and streamflow

Global mean reference runoff (CLASSr) is estimated to be 0.7 kg m^{-2} day^{-1} (Figure 3 and Figure 2g). The multi-model mean bias is -8%, with biases from individual models ranging between -55% (JSBACH) and +9% (ORCHIDEE-CNP). There is no clear tendency for models to have either positive or negative biases.

The models reproduce the zonal mean pattern of annual mean runoff reasonably well (Figure 2g). The seasonal runoff peak, however, is two months earlier compared to CLASSr (Figure B8). The time lag is present in multiple parts of the globe, including the boreal regions, tropical South America, and Europe (Figure B8).

Converting runoff to annual streamflow for the earth's 50 largest river basins and comparing values against gauge measurements from GRDC shows that models reproduce annual streamflow reasonably well (11/14 models with $R \geq 0.9$; Figure B9). However, none of the models nor the multi-model mean streamflow score of 0.71 reach the corresponding benchmark score of 0.82 (Figure 4).

3.9 Model performance

Our findings documented above show that benchmark scores vary considerably among variables, ranging from 0.33 for soil organic carbon to 0.82 for runoff. Model scores range from 0.39 to 0.71 for the same variables, which raises the question to what extent both scores are correlated. Figure 9 compares model scores against benchmark scores, where dots represent mean score values and bars show total ranges. The Figure shows that model scores and benchmark scores are positively correlated, suggesting that low model scores can result not only from model deficiencies, but also from observational uncertainties. One important exception is LAI, with model scores (0.50) that are much lower than benchmark scores (0.66 minimum) for globally gridded products. The large difference suggests that models have a great potential for improving their representation of LAI. This also applies when evaluating models against in situ LAI data from CEOS.

Another question we want to address here is to what extent model score differences are related to dynamic carbon-nitrogen (CN) interactions, permafrost, and wetlands (Table 1). There is no indication that a representation of CN interactions improves model performance. Comparing the model versions ORCHIDEE (with CN-interactions) against ORCHIDEEv3 (without CN-interactions) shows no statistically significant difference between the mean scores when considering all evaluations combined (two-sided t-test, p-value = 0.05). Comparing the mean score of all models that include CN-interactions (ten models) against the mean score of all models that lack such representation (five models) suggests that the inclusion of CN-interactions leads to statistically significant lower scores when assessing models for NBP from CT2019 (-0.03) and CAMS (-0.04). This result suggests that modeling groups may consider retuning their models when incorporating CN interactions. Models that include a representation of processes required for simulating carbon dynamics in permafrost regions (four models) tend to perform better than models that lack such representation when assessing runoff (0.02 for CLASSr and GRDC) and vegetation carbon (0.05 for FOSXue). Models that represent carbon dynamics in wetlands (three models) perform better for NBP (0.04 for CarboScope) but worse for vegetation carbon (-0.05 for ZhangMokany). Since only two models include a representation of carbon dynamics in peatlands, we cannot assess to what extent the inclusion of such processes have any statistical significance on model performance.

4 Discussion

Our study evaluates how well TRENDY models reproduce variables that drive the terrestrial carbon sink. A particular focus was to quantify what level of agreement between model output and reference data should be expected given that reference data are imperfect. Our approach accounts for observational uncertainties using two sets of skill scores. Model scores summarize the similarity between model output and reference data across multiple statistical metrics, including the bias, the centralized root mean square error, time lags of seasonal maxima or minima, inter-annual variability, as well as spatial variability and correlation. Scores range from zero to unity, where unity implies perfect agreement. Using the same statistical framework we then compute benchmark scores that quantify the similarity between independently derived reference data, which serves as an approximation of observational uncertainty. If model scores reach benchmark scores, then models perform sufficiently well, given how uncertain reference data are. For instance, comparing modeled against reference GPP from FluxCom yields a maximum model score of 0.70, suggesting that model performance is modest. However, comparing remotely sensed GPP (GOSIF) against FluxCom yields a benchmark score of 0.72, which suggests that model performance is reasonable given how uncertain reference data are.

Our results show that the disagreement between independently derived reference data are much larger than expected, with benchmark scores ranging between 0.33 for soil organic carbon, to 0.82 for annual streamflow. Comparing model scores against benchmark scores shows that both scores are positively correlated, suggesting that low model scores is often a sign of large observational uncertainty rather than poor model performance alone. For instance, model and benchmark scores are both relatively low for NBP (0.51 and 0.55, respectively) and relatively high for streamflow (0.71 and 0.82, respectively). The larger the gap between model scores and benchmark scores, the greater the potential for model improvement. For instance, this applies to LAI, with a model score of about 0.49 and a benchmark score of about 0.66 for globally gridded data. We further conclude that the lower the benchmark score, the greater the need to reduce observational uncertainty. This applies in particular to gridded reference data for soil organic carbon and inversion-based estimates for NBP.

Considering these findings, can we conclude that TRENDY models are fit for simulating the terrestrial carbon sink? Let us recall that the terrestrial carbon sink, which is here defined by the term S_{LAND} in equation 1, represents the natural carbon sink under present-day conditions for atmospheric CO_2 and climate, but pre-industrial land cover (S2 simulation). Given the counter-factual nature of S_{LAND}, we can only evaluate it indirectly by assessing NBP, and the processes that drive it, in the S3 simulation where CO_2, climate, and LULCC forcings all vary in time. The better a model performs for those variables, the greater the likelihood that its estimate of S_{LAND} is reliable. In the best case, all models, or at least the multi-model mean, would reach benchmark scores for all variables assessed in this study. While this is clearly not the case, for multiple variables (NBP, vegetation carbon, LAI, latent heat flux) there is at least one model that reaches the benchmark. In the case of GPP, none of the models reach the benchmark for globally gridded values (0.72), but some models come reasonably close (e.g. ORCHIDEE and SDGVM with 0.70). Furthermore, for GPP, vegetation carbon, and latent heat flux, the global multi-model annual mean values are within the uncertainty range of the reference data. This supports the notion that model diversity is a healthy aspect of any scientific community. Finally, the seasonal cycle of NBP across TransCom regions is reasonably consistent with results from inversion models, although the inter-model spread remains large, in particular in the boreal regions. We conclude that the performance of the TRENDY ensemble is encouraging, but that ample potential for improvements remains. Future efforts should focus on reducing the positive LAI bias across the globe, improving the representation of processes that govern soil organic carbon in high latitudes, and assessing the causes that drive the large inter-model spread of GPP amplitude in boreal regions and zonal mean GPP in the humid tropics. The potential for model improvement, however, also relies on our capability to reduce observational uncertainty. This applies in particular to globally gridded products of NBP and soil organic carbon.

Our approach leads to a new interpretation of the TRENDY model scores presented by Friedlingstein et al. (2020). Their main findings are that (i) TRENDY models show high skill scores for runoff, and to a lesser extent for vegetation biomass, GPP, and ecosystem respiration, and that (ii) skill scores are lowest for LAI and NEE, with a widest disparity among models for soil organic carbon. While our model scores are mainly consistent with these findings, our benchmark scores lead to a somewhat different interpretation. For instance, we confirm that model scores are larger for runoff than for GPP, but the difference between model and benchmark scores, and hence model performance, is approximately the same for both variables. Furthermore, the effectiveness of future model development is dependent on our ability to reduce observational uncertainties of these two variables. For soil organic carbon in particular, the observational uncertainties must be reduced substantially to provide adequate guidance for model development. If the large values in SG250m are due to a better representation of wetlands and permafrost soils compared to HWSD (Tifafi et al., 2018), then modeling groups may consider masking-out wetlands and permafrost soils when evaluating model output against HWSD (Tian, Lu, et al., 2015).

One limitation of our study is that our evaluation does not assess the CO_2 fertilization effect, which presents an important driver of S_{LAND} next to changes in climate. This could be addressed by including evaluations against Free Air CO_2 Enrichment (FACE) experiments in mature forests, which are currently in progress (Norby et al., 2016). Another limitation is that we are unable to assess how uncertainty in model inputs affects model scores as the TRENDY ensemble includes only a single set of model forcing data. However, this has been investigated by G. B. Bonan et al. (2019) and Seiler et al. (2021) for the terrestrial biosphere models CLM and CLASSIC, respectively. Both studies conclude that the uncertainties associated with climate forcing are too large to be neglected. For instance, Seiler et al. (2021) show that the global mean biases of seven out of 19 variables switches sign when forcing CLASSIC with different meteorological datasets. Such results suggest that robust model development must consider multiple forcing datasets to avoid tuning models towards a particular forcing dataset.

Future evaluations of TRENDY models would benefit from having access to aboveground vegetation carbon model output, which is currently available for some models only. Evaluating above ground rather than total vegetation carbon is an advantage because below ground vegetation carbon is difficult to measure. Furthermore, modeling groups should provide PFT-specific values for aboveground vegetation carbon and NEE to allow for a more direct evaluation against forest inventory data and eddy covariance measurements, respectively. Finally, a more comprehensive evaluation would require access to more model variables for all TRENDY models, including radiation fluxes, sensible heat flux, soil respiration, fractional area burnt, CO_2 emissions from fires, and snow water equivalent. Including those variables may help diagnosing the underlying causes of model deficiencies.

Our results demonstrate that benchmark scores facilitate the interpretation of model scores as they indicate what level of agreement between model output and reference data may be expected, and whether low model scores indeed reflect poor model performance or observational uncertainty. Our benchmark approach is not limited to TBMs or the AMBER or ILAMB statistical framework, but can be applied to any geophysical model that is evaluated against observations. We hope these results will stimulate model development that aims at reducing the uncertainties of processes that drive terrestrial carbon, water, and energy fluxes.

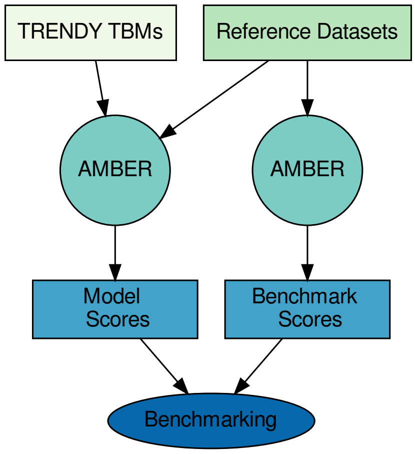

Figure 1. Conceptual diagram of benchmarking Terrestrial Biosphere Models (TBMs) using the Automated Model Benchmarking R package (AMBER). Model scores are computed by comparing model output against reference data. Benchmark scores are computed by comparing multiple reference datasets against each other. Benchmarking consists of comparing model scores against benchmark scores.

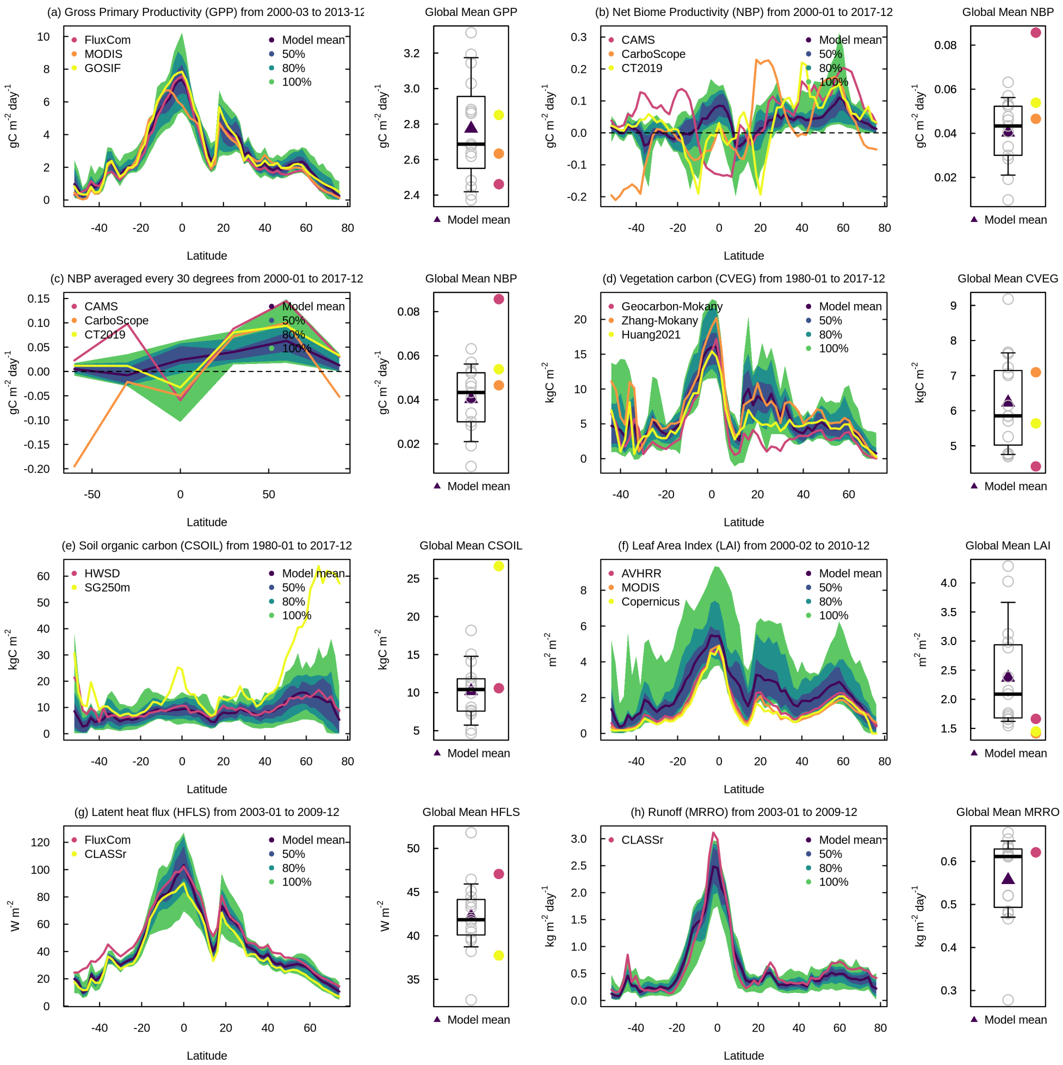

Figure 2. Zonal mean values of annual mean (a) gross primary productivity, (b) net biome productivity, (c) net biome productivity averaged every 30 degrees latitude (d) vegetation carbon, (e) soil organic carbon, (f) leaf area index, (g) latent heat flux, and (h) runoff. Red/yellow color shades denote reference data, and blue/green color shades give the mean values and percentiles of models (50%, 80%, 100%). The boxplots give the multi-model median, the inter-quartile range (box), and 80^{th} percentiles (whiskers) of global annual mean values. Triangles give the multi-model mean, and grey circles indicate results for individual models.

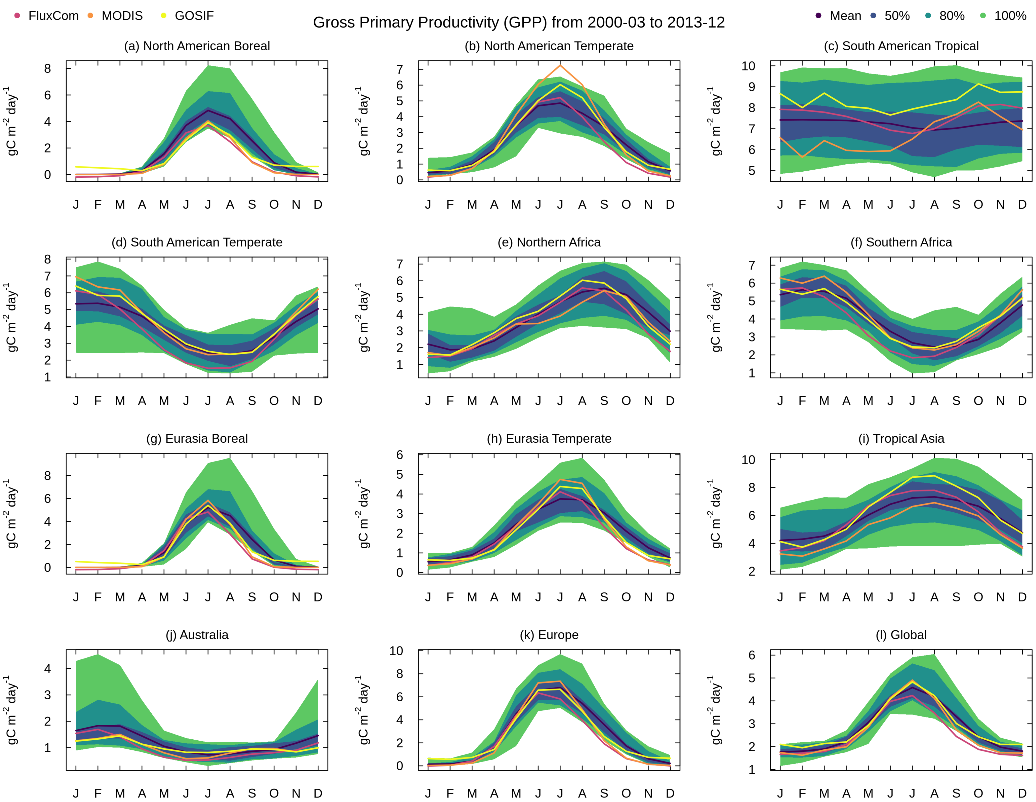

Figure 3. Climatological mean seasonal cycle of gross primary productivity for TransCom regions shown in Figure Appendix B1a. Blue/green color shades give the mean values and percentiles of models (50%, 80%, 100%).

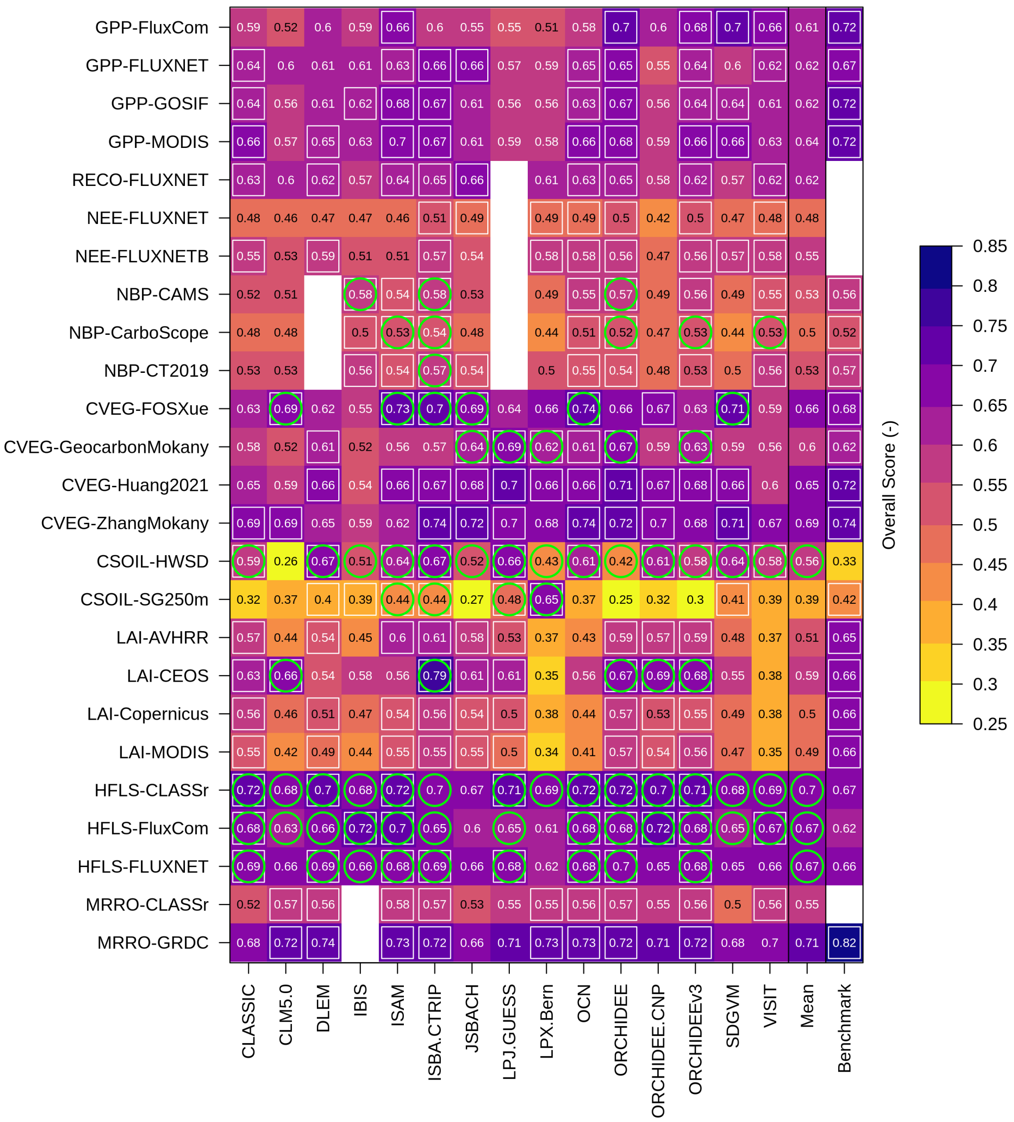

Figure 4. Model and benchmark scores, where white boxes present cases where model scores exceed the multi-model mean values and green circles denote cases where model scores exceed benchmark scores. Blank spaces indicate missing data.

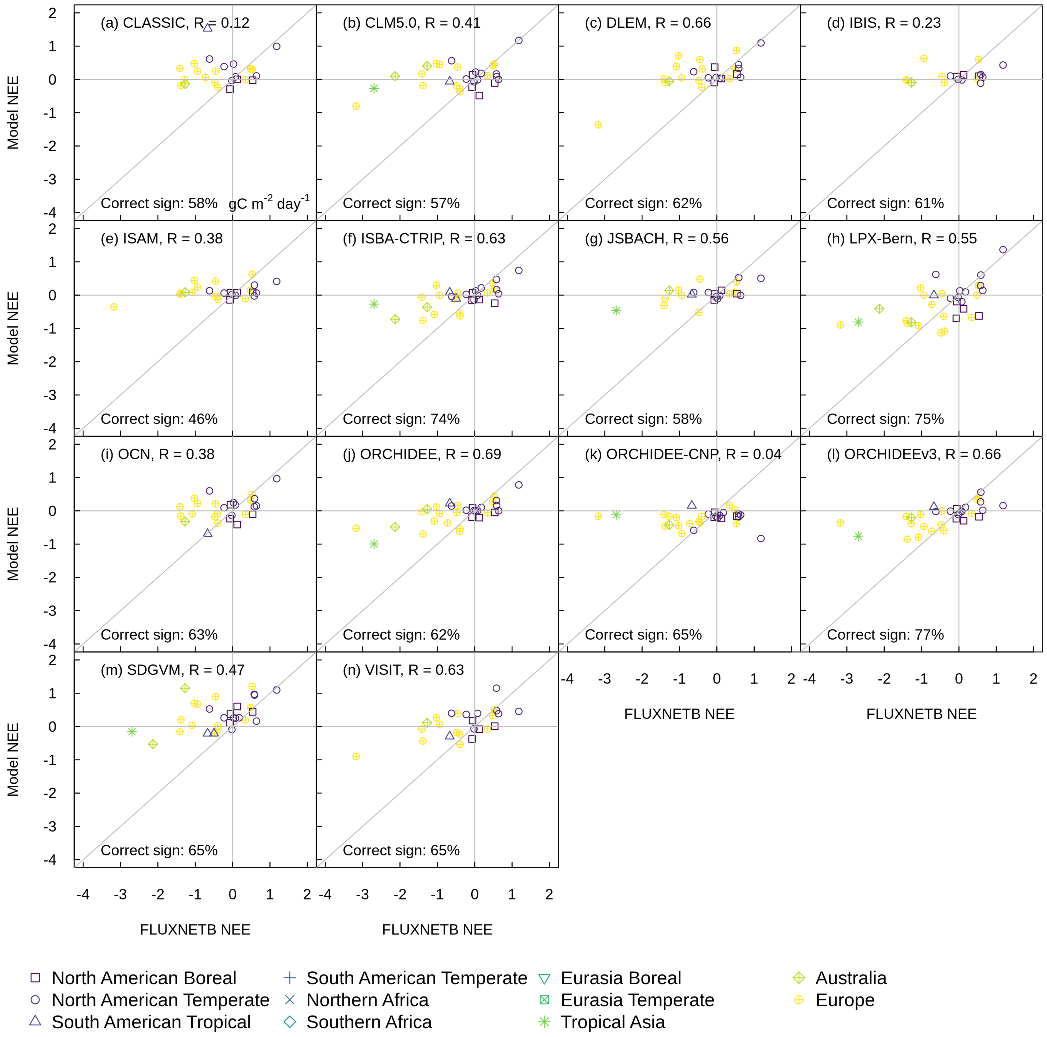

Figure 5. Evaluation of annual mean net ecosystem exchange model output against forest eddy-covariance measurements that were filtered for data quality and disturbance history in units of gC m^{-2} day^{-1}.

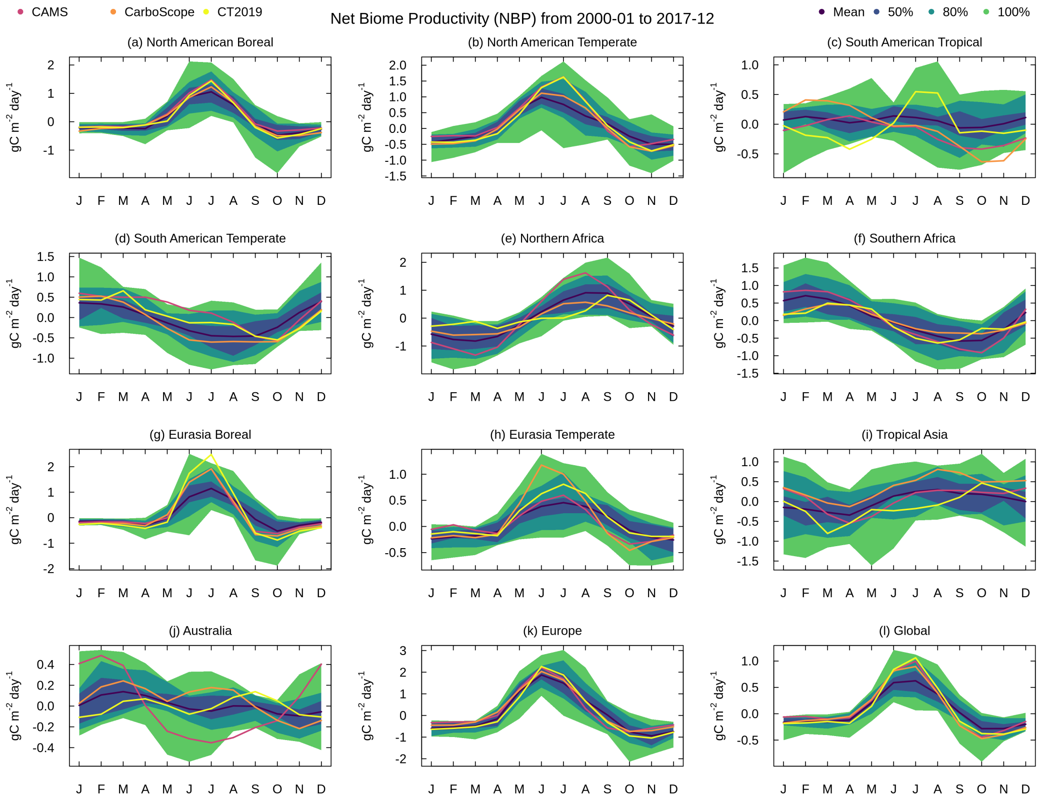

Figure 6. Same as Figure 3 but for net biome productivity.

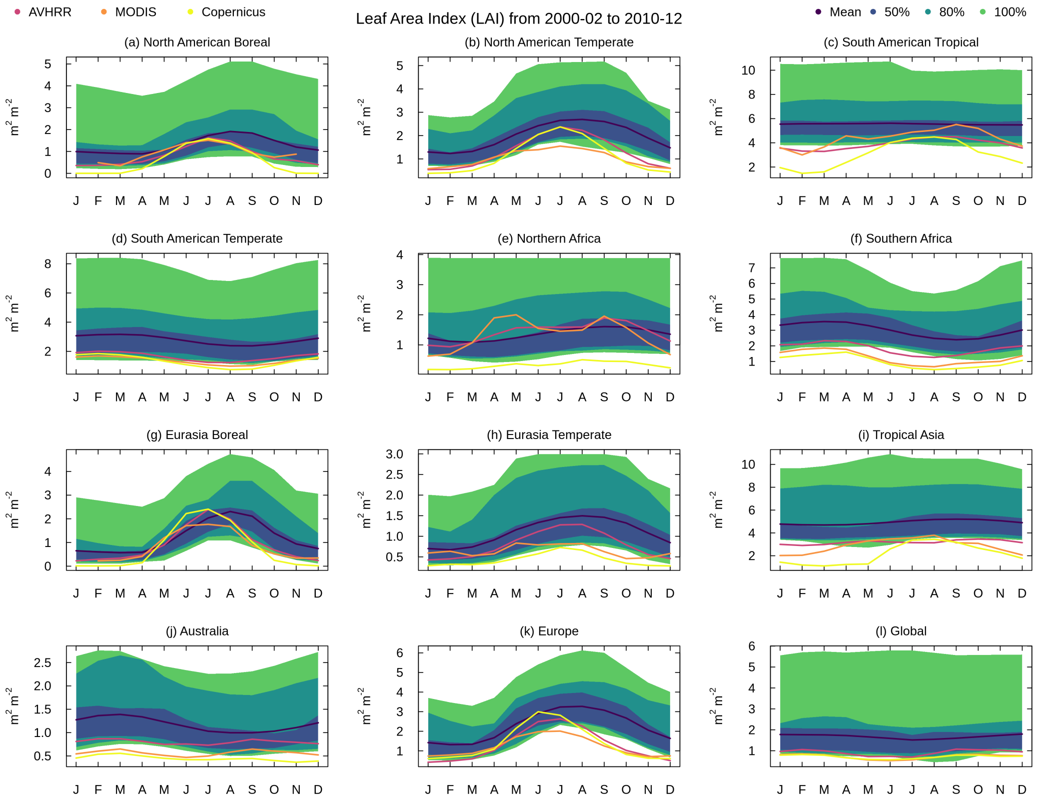

Figure 7. Same as Figure 3 but for leaf area index.

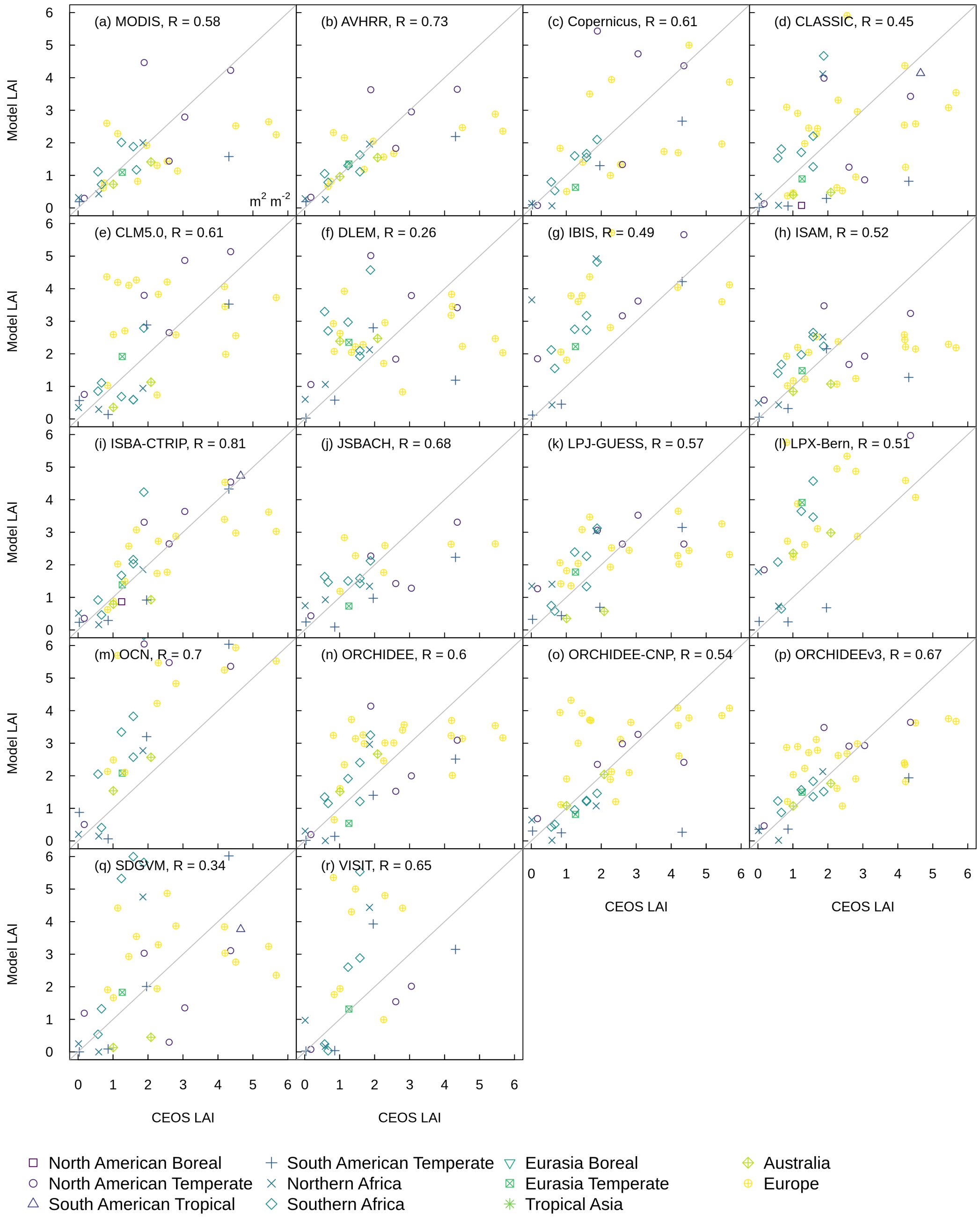

Figure 8. Evaluation of leaf area index against site-level measurements with units in m^2 m^{-2}.

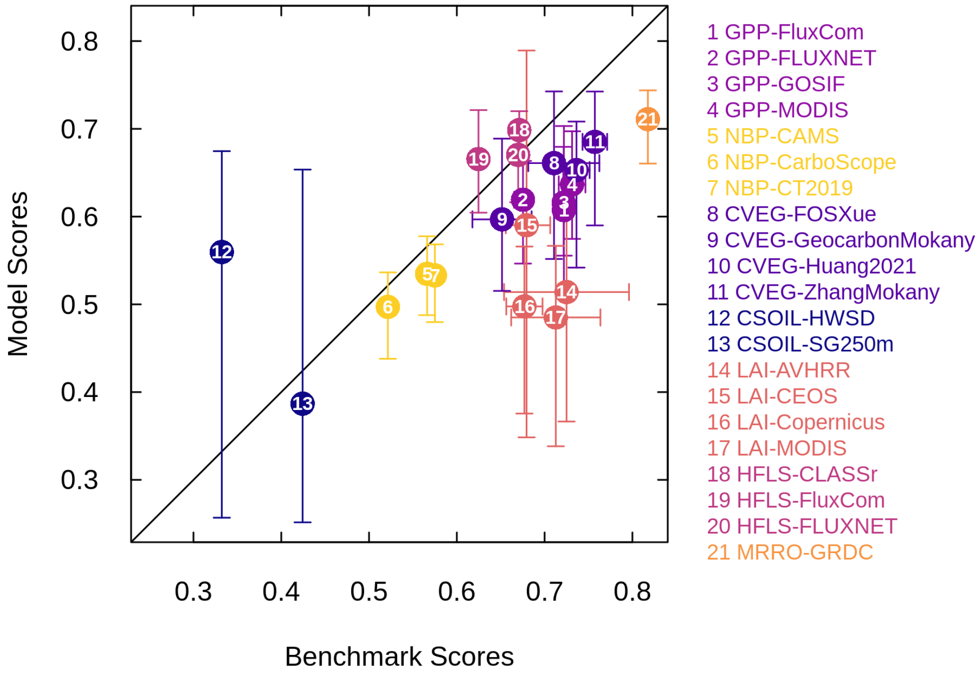

Figure 9. Model scores and benchmark scores, where dots present multi-model mean values and bars give the total range of model scores.

Table 1. TRENDY (v9) terrestrial biosphere models, their horizontal resolution in terms of degrees longitude and latitude, and whether models include representations of processes required for simulating carbon cycle dynamics related to (i) carbon-nitrogen (C-N) interaction, (ii) wetlands, (iii) peatlands, and (iv) permafrost.

Model	Resolution	C-N	Wetland	Peatland	Permafrost	Reference
CLASSIC	$1° \times 1°$	no	no	no	no	Melton et al. (2020)
CLM5.0	$1° \times 1°$	yes	no	no	yes	Lawrence et al. (2019)
DLEM	$0.5° \times 0.5°$	yes	yes	yes	no	Tian, Chen, et al. (2015)
IBIS	$1° \times 1°$	no	no	no	no	Yuan et al. (2014)
ISAM	$0.5° \times 0.5°$	yes	yes	no	yes	Meiyappan et al. (2015)
ISBA-CTRIP	$1° \times 1°$	no	no	no	yes	Delire et al. (2020)
JSBACH	$1.875° \times 1.875°$	yes	no	no	no	Reick et al. (2021)
LPJ-GUESS	$0.5° \times 0.5°$	yes	no	no	no	Smith et al. (2014)
LPX-Bern	$0.5° \times 0.5°$	yes	no	yes	yes	Lienert and Joos (2018)
OCN	$1° \times 1°$	yes	no	no	no	Zaehle and Friend (2010)
ORCHIDEE	$0.5° \times 0.5°$	no	no	no	no	Krinner et al. (2005)
ORCHIDEE-CNP	$2° \times 2°$	yes	no	no	no	Goll et al. (2017)
ORCHIDEEv3	$2° \times 2°$	yes	no	no	no	Vuichard et al. (2019)
SDGVM	$1° \times 1°$	yes	no	no	no	Walker et al. (2017)
VISIT	$0.5° \times 0.5°$	no	yes	no	no	Kato et al. (2013)

Table 2. Observation-based reference data used for model evaluation. Meanings of acronyms are provided in the Methods section.

Source	Variables	Approach (n sites)	Period	Reference
In situ measurements				
FLUXNET2015	GPP, RECO, NEE, HFLS	eddy covariance (204)	1997-2014	Pastorello et al. (2020)
FOS	CVEG	allometry (274)	1999-2018	Schepaschenko et al. (2019)
Xue	CVEG	allometry (1974)	1999-2018	Xue et al. (2017)
CEOS	LAI	transfer function (141)	1999-2017	Garrigues et al. (2008)
GRDC	MRRO	gauge records (50)	1980-2010	Dai and Trenberth (2002)
Globally gridded datasets				
MODIS	GPP	light use efficiency model	2000-2016	Zhang et al. (2017)
GOSIF	GPP	statistical model	2000-2017	Li and Xiao (2019)
FluxCom	GPP	machine learning	1980-2013	Jung et al. (2020)
CT2019	NEE	atmospheric inversion	2000-2017	Jacobson et al. (n.d.)
CAMS	NBP	atmospheric inversion	1979-2019	Agustí-Panareda et al. (2019)
CarboScope	NBP	atmospheric inversion	1999-2019	Rödenbeck et al. (2018)
GEOCARBON	CVEG	machine learning	NA	Avitabile et al. (2016), Santoro et al. (2015)
Zhang	CVEG	data fusion	2000s	Zhang and Liang (2020)
HWSD	CSOIL	soil inventory	NA	Wieder (2014) Todd-Brown et al. (2013)
SG250m	CSOIL	machine learning	NA	Hengl et al. (2017)
AVHRR	LAI	artificial neural network	1982-2010	Claverie et al. (2016)
Copernicus	LAI	artifial neural network	1999-2019	Verger et al. (2014)
MODIS	LAI	radiative transfer model	2000-2017	R. B. Myneni et al. (2002)
FluxCom	HFLS	machine learning	2001-2013	Jung et al. (2019)
CLASSr	HFLS, MRRO	blended product	2003-2009	Hobeichi et al. (2019)

Table 3. Global reference (Ref.) and multi-model mean values, with multi-model mean, minimum, and maximum relative biases, and number of models with positive (Pos.) and negative (Neg.) biases. In the absence of a reference period, model values are averaged over the 1980-2017 period.

Variable	Ref.ID	Period	Unit	Reference	Multi-model Mean	Mean Bias (%)	Minimum Bias (%)	Maximum Bias (%)	Pos.	Neg.
GPP	FluxCom	1980-2013	PgC yr^{-1}	108.9	115.0	6	-17	25	11	4
GPP	GOSIF	2000-2017	PgC yr^{-1}	123.8	116.0	-6	-27	12	4	11
GPP	MODIS	2000-2016	PgC yr^{-1}	115.2	119.3	4	-20	23	11	4
NBP	CAMS	1979-2019	PgC yr^{-1}	1.9	1.0	-46	-86	-19	0	13
NBP	CarboScope	1999-2019	PgC yr^{-1}	1.3	1.3	-1	-79	50	7	6
NBP	CT2019	2000-2017	PgC yr^{-1}	1.3	1.2	-9	-82	37	5	8
CVEG	Geocarbon-Mokany	YYYYs	PgC	264.6	403.3	52	11	109	15	0
CVEG	Zhang-Mokany	2000s	PgC	482.5	429.2	-11	-35	20	5	10
CVEG	Huang2021	NA	PgC	310.2	344.6	11	-17	53	9	6
CSOIL	HWSD	NA	PgC	1143.4	1121.1	-3	-57	146	6	9
CSOIL	SG250m	NA	PgC	2708.0	1160.9	-57	-82	9	1	14
LAI	AVHRR	1982-2010	m^2 m^{-2}	1.4	2.1	58	4	210	15	0
LAI	Copernicus	1999-2019	m^2 m^{-2}	1.4	2.0	50	-4	187	14	1
LAI	MODIS	2000-2017	m^2 m^{-2}	1.5	2.5	67	9	220	15	0
HFLS	CLASSr	2003-2009	W m^{-2}	32.6	37.0	13	-12	40	14	1
HFLS	FluxCom	2001-2013	W m^{-2}	45.2	40.1	-11	-34	10	1	14
MRRO	CLASSr	2003-2009	kg m^{-2} day^{-1}	0.7	0.6	-8	-55	9	8	6

Appendix A Automated Model Benchmarking R package (AMBER)

The Automated Model Benchmarking R package (AMBER; version 1.1.0) quantifies model performance using five scores that assess the model's bias (S_{bias}), root-mean-square-error (S_{rmse}), seasonality (S_{phase}), inter-annual variability (S_{iav}), and spatial distribution (S_{dist}). All scores are dimensionless and range from zero to one, where increasing values imply better performance. The exact definition of each skill score is provided below.

A01 Bias score (S_{bias})

The bias is defined as the difference between the time-mean values of model and reference data:

$$bias(\lambda, \phi) = \overline{v_{mod}}(\lambda, \phi) - \overline{v_{ref}}(\lambda, \phi), \tag{A1}$$

where $\overline{v_{mod}}(\lambda, \phi)$ and $\overline{v_{mod}}(\lambda, \phi)$ are the mean values in time (t) of a variable v as a function of longitude λ and latitude ϕ for model and reference data, respectively. Nondimensionalization is achieved by dividing the bias by the standard deviation of the reference data (σ_{ref}):

$$\varepsilon_{bias}(\lambda, \phi) = |bias(\lambda, \phi)|/\sigma_{ref}(\lambda, \phi). \tag{A2}$$

Note that ε_{bias} is always positive, as it uses the absolute value of the bias. For evaluations against stream flow measurements the bias is divided by the annual mean rather than the standard deviation of the reference data. This is because we assess streamflow on an annual rather than monthly basis, implying that the corresponding standard deviation is small. The same approach is applied to soil carbon and biomass, whose reference data provide a static snap shot in time. In both of these cases, $\varepsilon_{bias}(\lambda, \phi)$ becomes:

$$\varepsilon_{bias}(\lambda, \phi) = |bias(\lambda, \phi)|/\overline{v_{ref}}(\lambda, \phi). \tag{A3}$$

A bias score that scales from zero to one is calculated next:

$$s_{bias}(\lambda, \phi) = e^{-\varepsilon_{bias}(\lambda, \phi)}. \tag{A4}$$

While small relative errors yield score values close to one, large relative errors cause score values to approach zero. Taking the mean of s_{bias} across all latitudes and longitudes, denoted by a double bar over a variable, leads to the scalar score:

$$S_{bias} = \overline{\overline{s_{bias}}}(\lambda, \phi). \tag{A5}$$

A02 Root-mean-square-error score (S_{rmse})

While the bias assesses the difference between time-mean values, the root-mean-square-error ($rmse$) is concerned with the residuals of the modeled and observed time series:

$$rmse(\lambda, \phi) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} (v_{mod}(t, \lambda, \phi) - v_{ref}(t, \lambda, \phi))^2 dt}, \tag{A6}$$

where t_0 and t_f are the initial and final time step, respectively. A similar metric is the centralized $rmse$ ($crmse$), which is based on the residuals of the anomalies:

$$crmse(\lambda, \phi) = \sqrt{\frac{1}{t_f - t_0} \int_{t_0}^{t_f} [(v_{mod}(t, \lambda, \phi) - \overline{v_{mod}}(\lambda, \phi)) - (v_{ref}(t, \lambda, \phi) - \overline{v_{ref}}(\lambda, \phi))]^2 dt}. \tag{A7}$$

The $crmse$, therefore, assesses residuals that have been bias-corrected. Since we already assessed the model's bias through S_{bias}, it is convenient to assess the residuals using $crmse$ rather than $rmse$. In a similar fashion to the bias, we then compute a relative error:

$$\varepsilon_{rmse}(\lambda, \phi) = crmse(\lambda, \phi)/\sigma_{ref}(\lambda, \phi), \tag{A8}$$

scale this error onto a unit interval:

$$s_{rmse}(\lambda, \phi) = e^{-\varepsilon_{rmse}(\lambda,\phi)}, \tag{A9}$$

and compute the spatial mean:

$$S_{rmse} = \overline{\overline{s_{rmse}}}. \tag{A10}$$

A03 Phase score (S_{phase})

The skill score S_{phase} assesses how well the model reproduces the seasonality of a variable by computing the time difference ($\theta(\lambda, \phi)$) between modeled and observed maxima of the climatological mean cycle:

$$\theta(\lambda, \phi) = \max(c_{mod}(t, \lambda, \phi)) - \max(c_{ref}(t, \lambda, \phi)), \tag{A11}$$

where c_{mod} and c_{ref} are the climatological mean cycle of the model and reference data, respectively. This time difference is then scaled from zero to one based on the consideration that the maximum possible time difference is six months:

$$s_{phase}(\lambda, \phi) = \frac{1}{2}\left[1 + \cos\left(\frac{2\pi\theta(\lambda, \phi)}{365}\right)\right]. \tag{A12}$$

The spatial mean of s_{phase} then leads to the scalar score:

$$S_{phase} = \overline{\overline{s_{phase}}}. \tag{A13}$$

A04 Inter-annual variability score (S_{iav})

The skill score S_{iav} quantifies how well the model reproduces patterns of inter-annual variability. This score is based on data where the seasonal cycle (c_{mod} and c_{ref}) has been removed:

$$iav_{mod}(\lambda, \phi) = \sqrt{\frac{1}{t_f - t_0}\int_{t_0}^{t_f} (v_{mod}(t, \lambda, \phi) - c_{mod}(t, \lambda, \phi))^2 dt}, \tag{A14}$$

$$iav_{ref}(\lambda, \phi) = \sqrt{\frac{1}{t_f - t_0}\int_{t_0}^{t_f} (v_{ref}(t, \lambda, \phi) - c_{ref}(t, \lambda, \phi))^2 dt}. \tag{A15}$$

The relative error, nondimensionalization, and spatial mean are computed next:

$$\varepsilon_{iav} = |(iav_{mod}(\lambda, \phi) - iav_{ref}(\lambda, \phi))|/iav_{ref}(\lambda, \phi), \tag{A16}$$

$$s_{iav}(\lambda, \phi) = e^{-\varepsilon_{iav}(\lambda,\phi)}, \tag{A17}$$

$$S_{iav} = \overline{\overline{s_{iav}}}. \tag{A18}$$

A05 Spatial distribution score (S_{dist})

The spatial distribution score S_{dist} assesses how well the model reproduces the spatial pattern of a variable. The score considers the correlation coefficient R and the relative standard deviation σ between $\overline{v_{mod}}(\lambda, \phi)$ and $\overline{v_{ref}}(\lambda, \phi)$. The score S_{dist} increases from zero to one, the closer R and σ approach a value of one. No spatial integration is required as this calculation yields a single value:

$$S_{dist} = 2(1 + R)\left(\sigma + \frac{1}{\sigma}\right)^{-2}, \tag{A19}$$

where σ is the ratio between the standard deviation of the model and reference data:

$$\sigma = \sigma_{\overline{v_{mod}}}/\sigma_{\overline{v_{ref}}}. \tag{A20}$$

A06 Overall score ($S_{overall}$)

As a final step, scores are averaged to obtain an overall score:

$$S_{overall} = \frac{S_{bias} + 2S_{rmse} + S_{phase} + S_{iav} + S_{dist}}{1+2+1+1+1}. \tag{A21}$$

Note that S_{rmse} is weighted by a factor of two, which emphasizes its importance.

Appendix B Supportive Figures

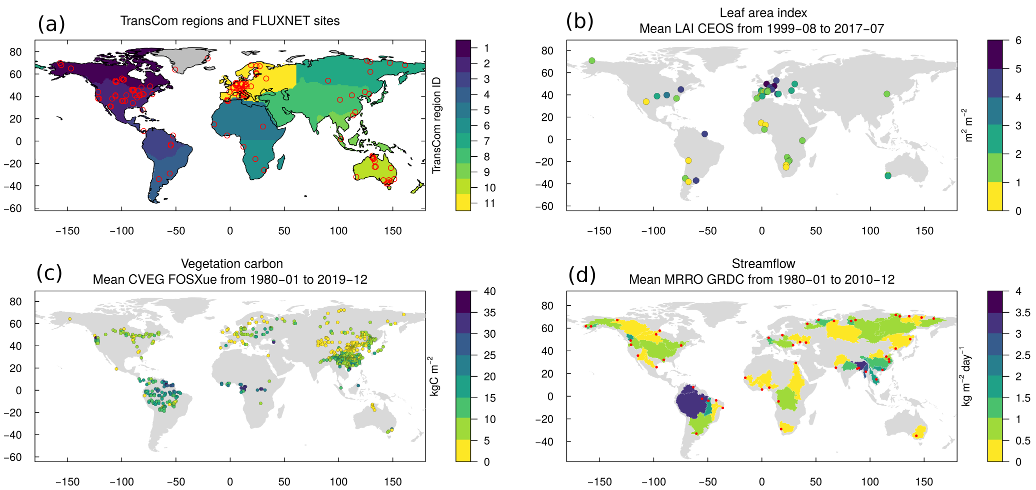

Figure B1. (a) Location of FLUXNET sites and TransCom regions (1 = North American Boreal, 2 = North American Temperate, 3 = South American Tropical, 4 = South American Temperate, 5 = Northern Africa, 6 = Southern Africa, 7 = Eurasia Boreal, 8 = Eurasia Temperate, 9 = Tropical Asia, 10 = Australia, 11 = Europe) (Gurney et al., 2004), (b) site-level measurements of leaf area index, (c) forest inventory sites, and (d) river basins with location of streamflow measurements.

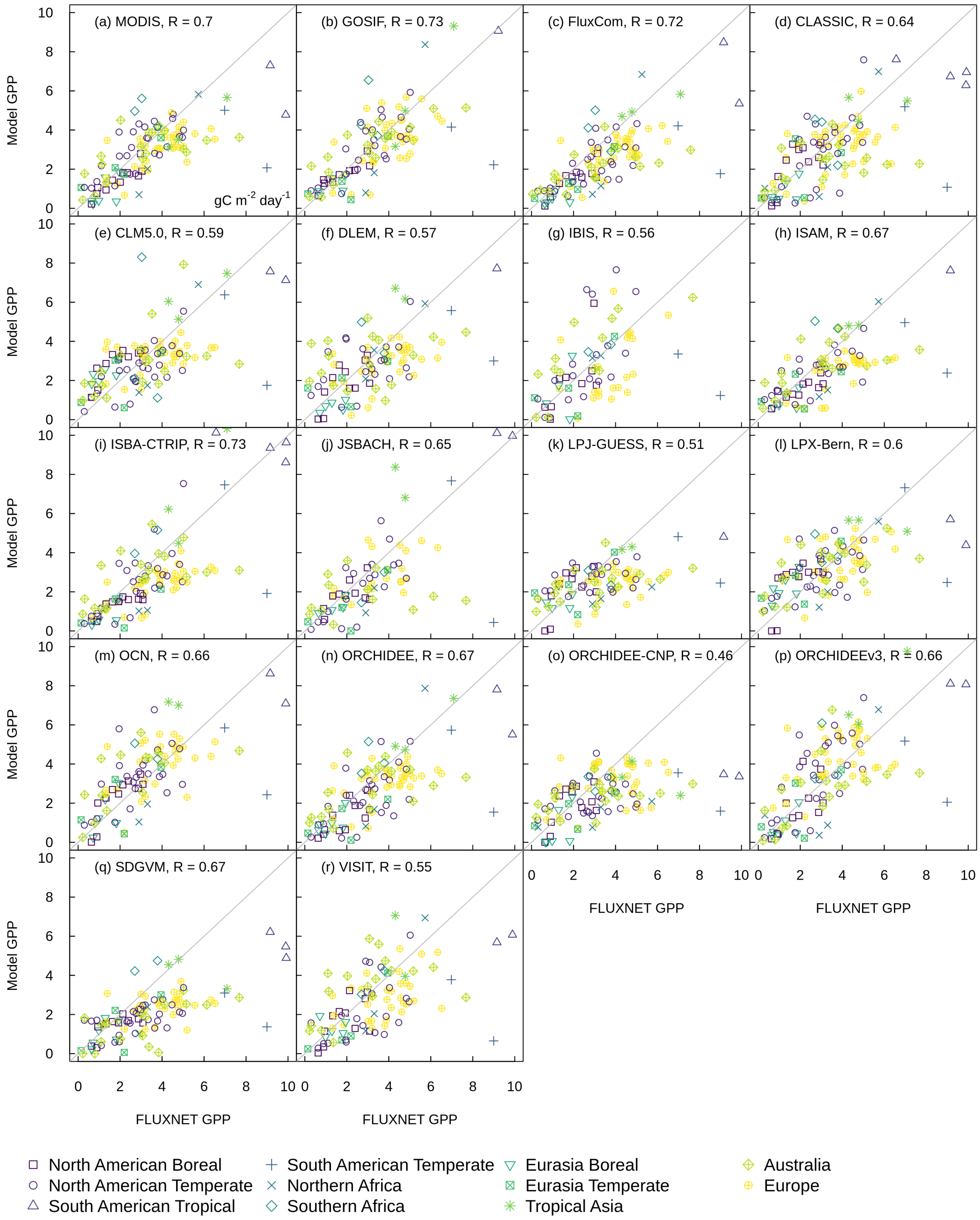

Figure B2. Evaluation of gross primary productivity against eddy covariance measurements in units of gC m^{-2} day^{-1}.

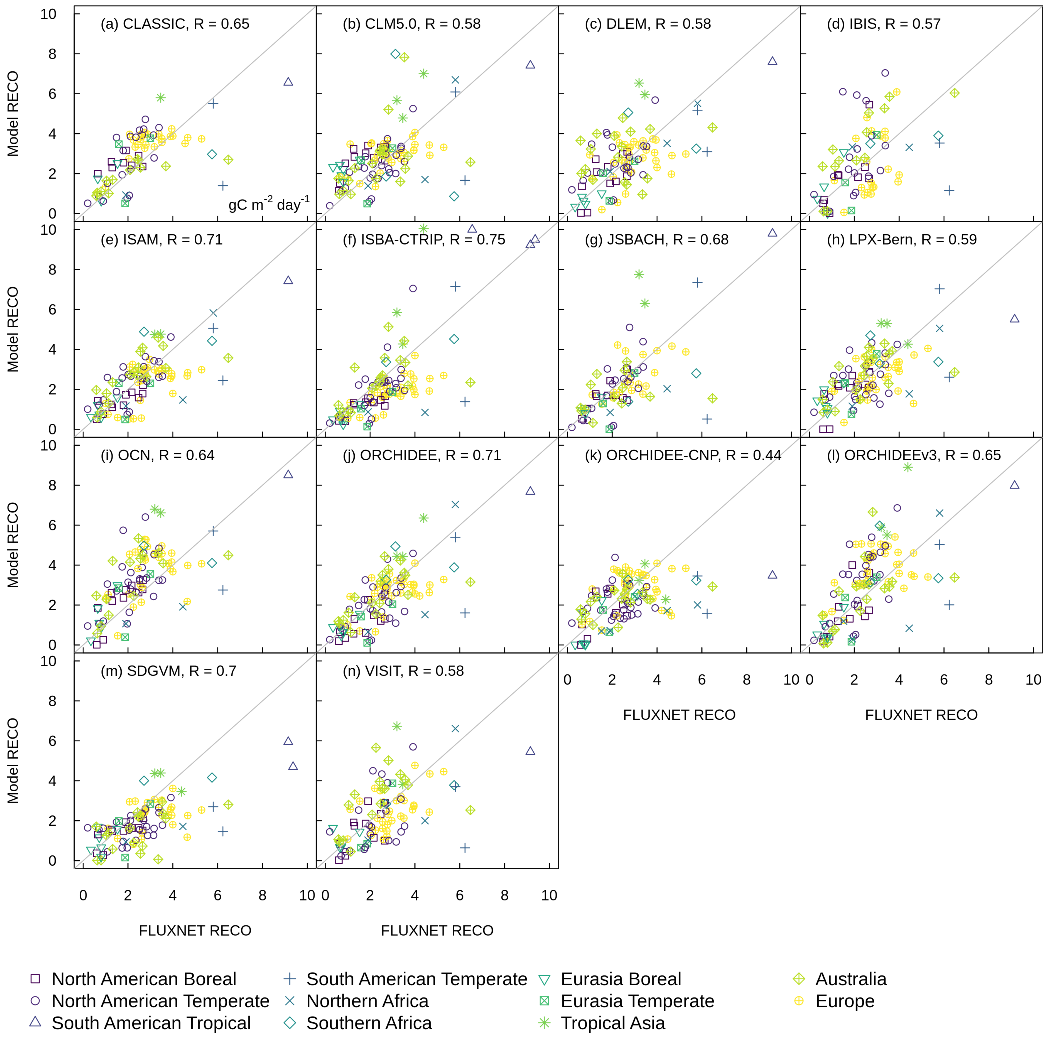

Figure B3. Evaluation of ecosystem respiration against eddy covariance measurements in units of gC m^{-2} day^{-1}.

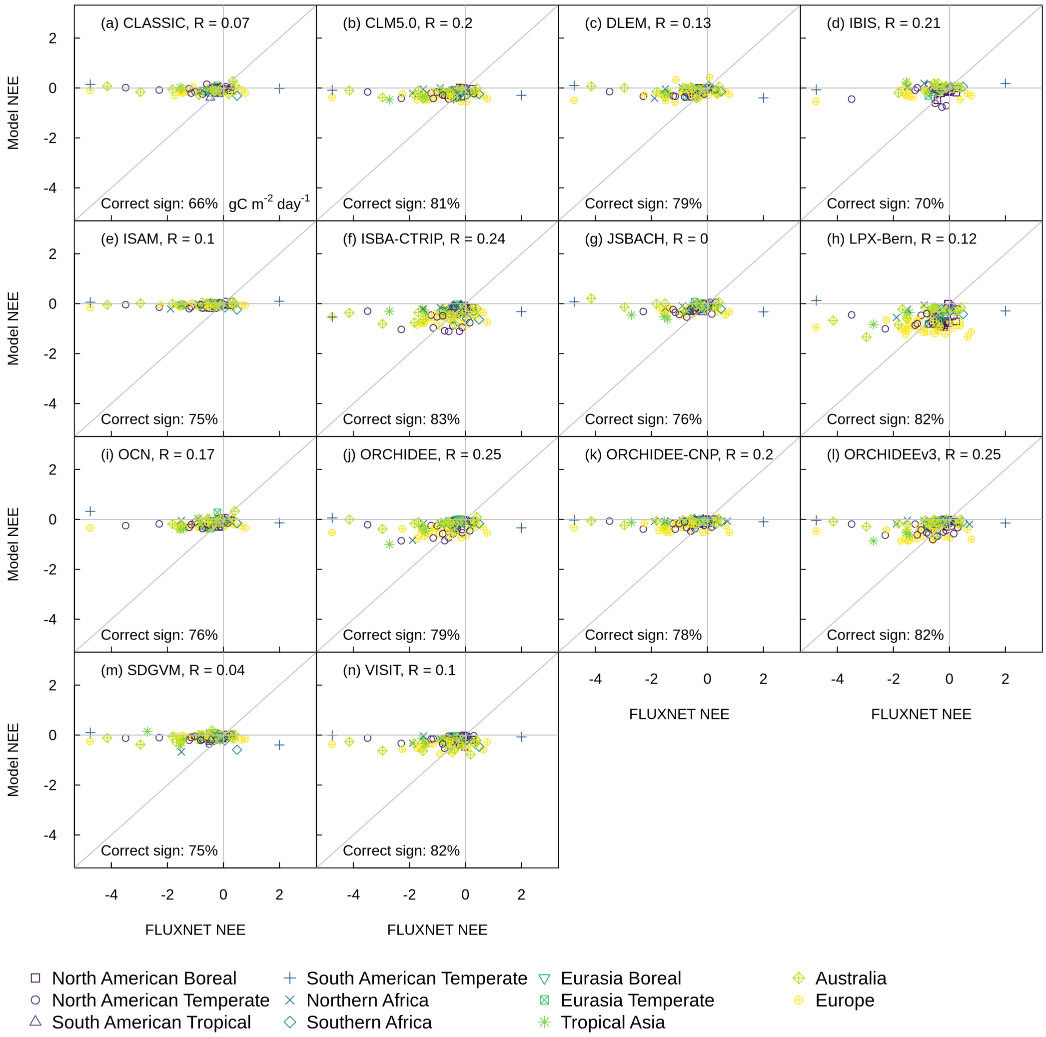

Figure B4. Evaluation of annual mean net ecosystem exchange model output against eddy-covariance measurements in units of gC m^{-2} day^{-1}.

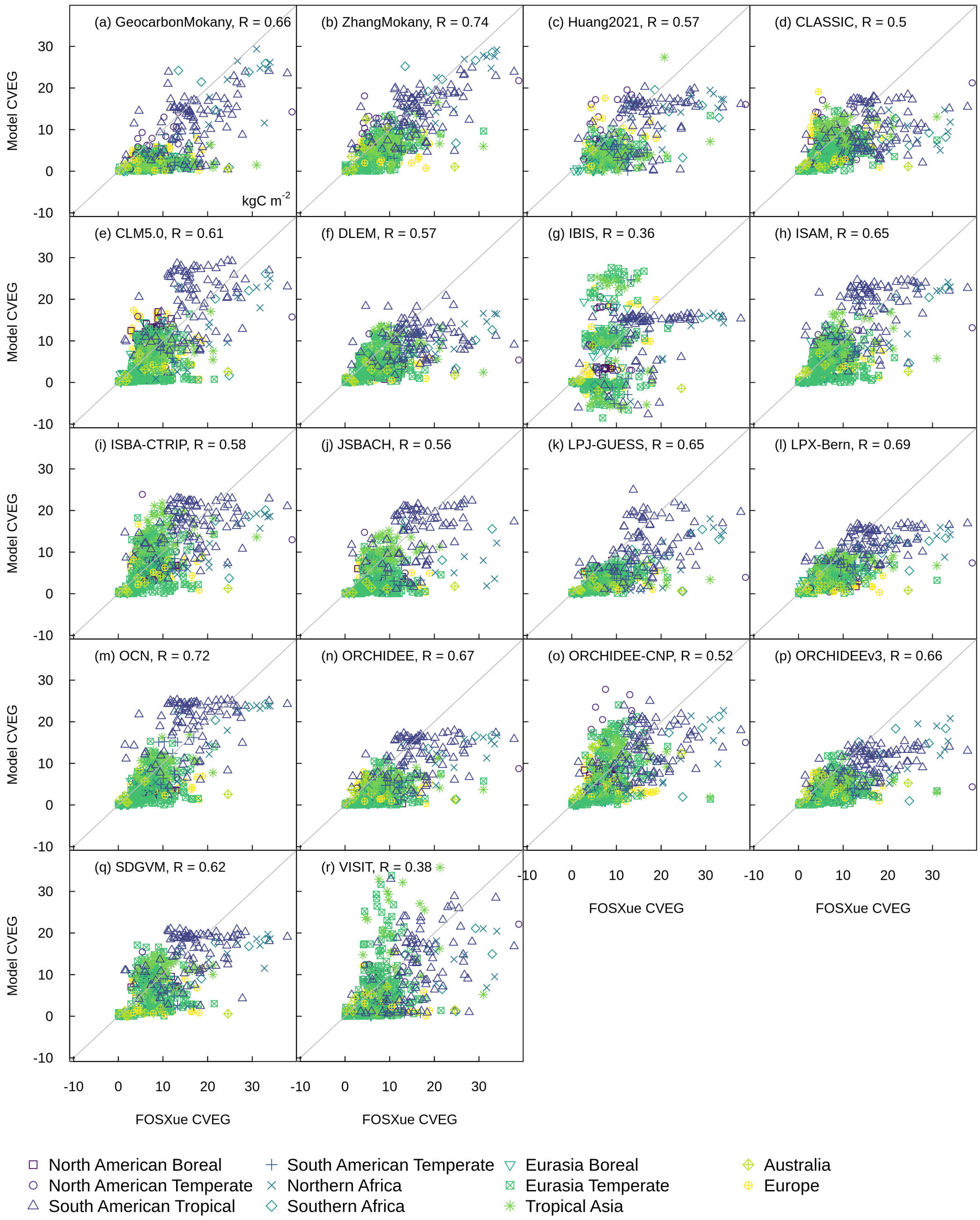

Figure B5. Evaluation of vegetation carbon against site-level measurements in units of kgC m^{-2}.

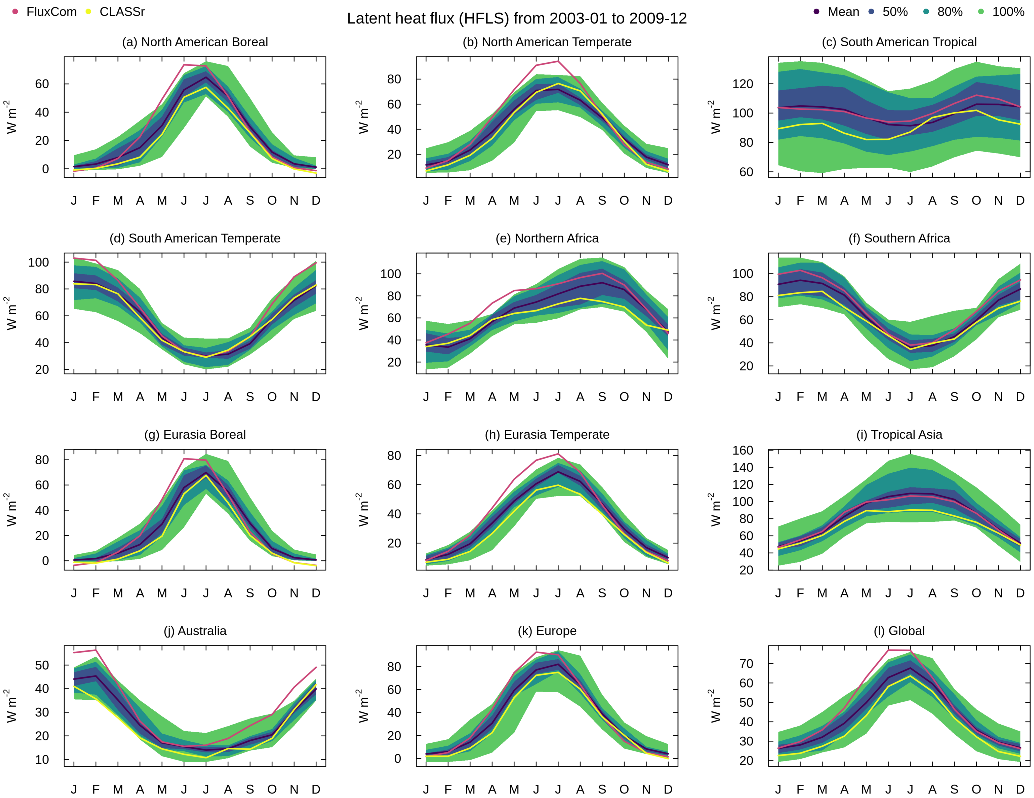

Figure B6. Same as Figure 3 but for latent heat flux.

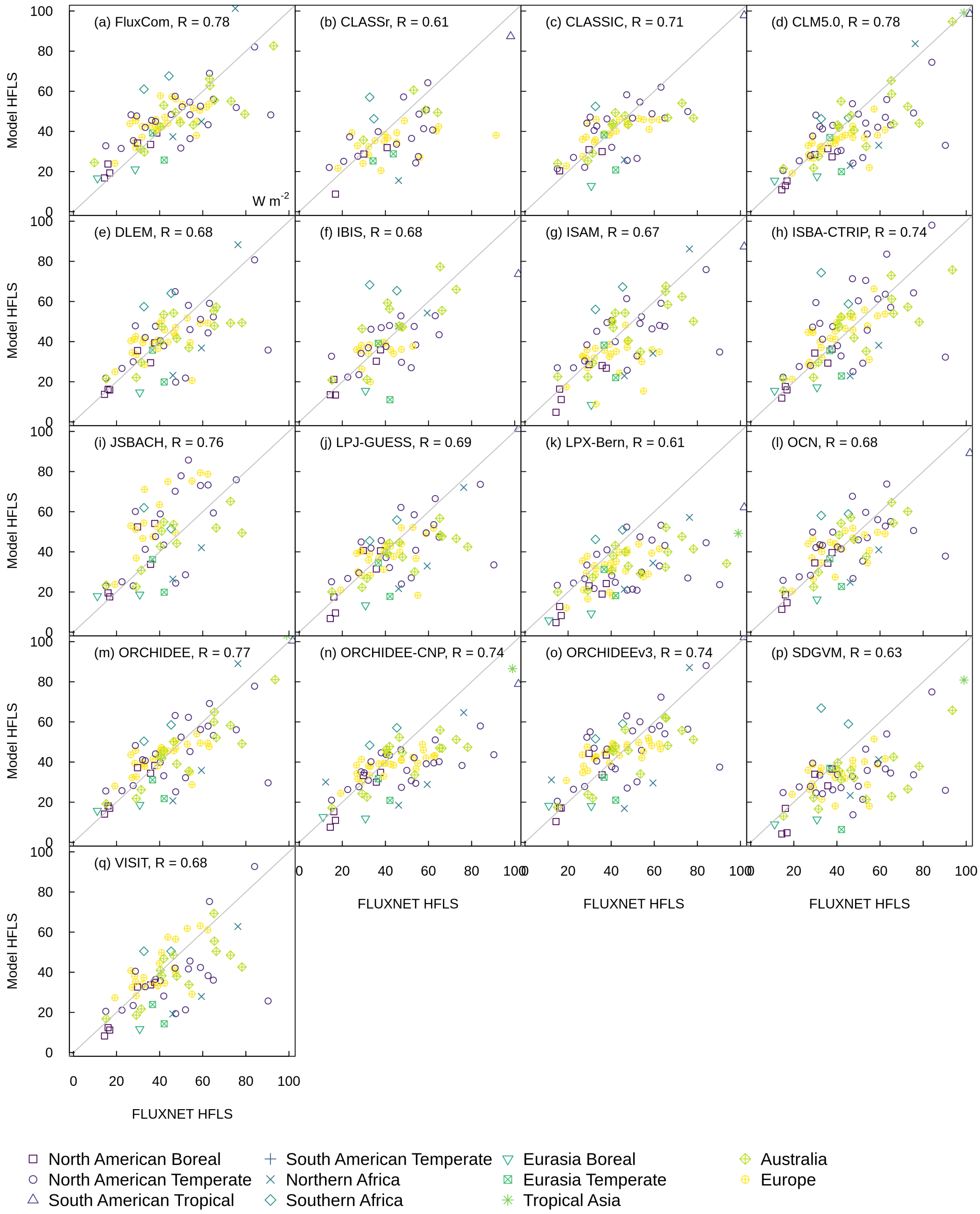

Figure B7. Evaluation of latent heat flux against eddy covariance measurements in units of W m^{-2}.

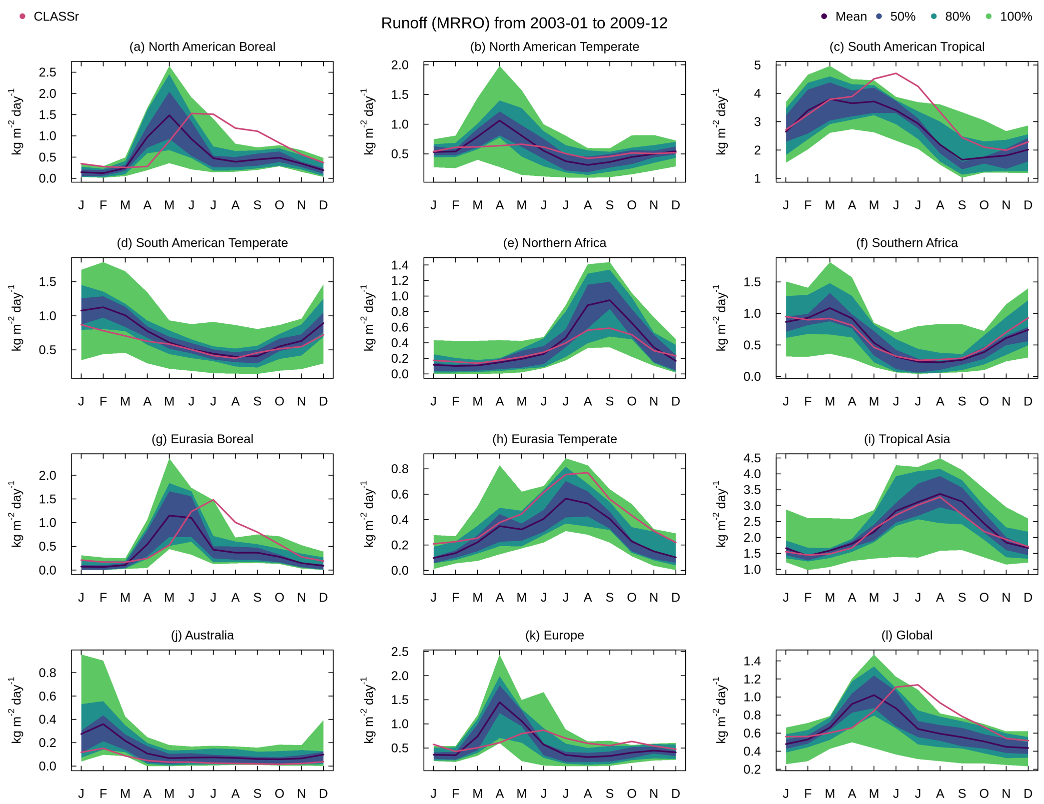

Figure B8. Same as Figure 3 but for runoff.

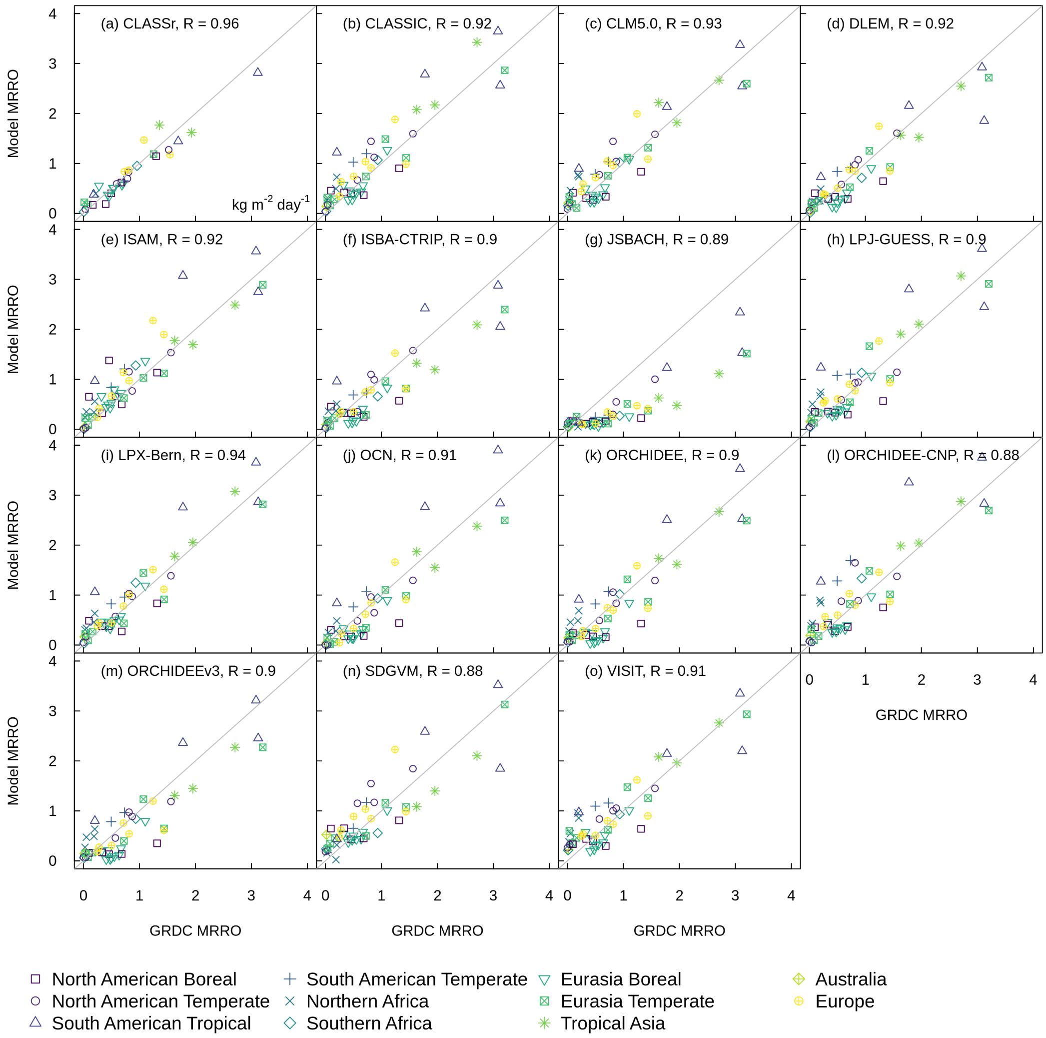

Figure B9. Evaluation of annually streamflow against gauge records in units of kg m^{-2} day^{-1}.

Acknowledgments

The authors wish to thank all groups that provided public access to the reference data listed in Table 2. The eddy covariance data that are shared by the FLUXNET community include the networks AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. 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. ORNL is managed by UT-Battelle, LLC, for the DOE under contract DE-AC05-1008 00OR22725. EJ acknowledges the European Union's Horizon 2020 research and innovation program under grant agreement no. 101003536 (ESM2025 – Earth System Models for the Future). Libo Wang compiled LAI from MODIS and Brianna Wolfe compiled LAI from Copernicus, as well as aboveground biomass in situ measurements. Mike Brady ensured that AMBER and its dependencies can be deployed across Linux platforms. Roland Séférian provided comments on an earlier version of the text. The data, scripts, code, computational environment, and instructions required for reproducing the results presented in our paper can be downloaded from `https://doi.org/10.5281/zenodo.5670387`. The full set of Figures produced by AMBER for this study can be accessed at `https://cseiler.shinyapps.io/AmberTrendy2020/` (last visited on November 22, 2021).

References

Agustí-Panareda, A., Diamantakis, M., Massart, S., Chevallier, F., Muñoz-Sabater, J., Barré, J., … Wunch, D. (2019, June). Modelling CO_2 weather – why horizontal resolution matters. *Atmos. Chem. Phys.*, *19*(11), 7347–7376.

Avitabile, V., Herold, M., Heuvelink, G. B. M., & others. (2016). An integrated pan tropical biomass map using multiple reference datasets. *Glob. Chang. Biol.*.

Baccini, A., Goetz, S. J., Walker, W. S., Laporte, N. T., Sun, M., Sulla-Menashe, D., … Houghton, R. A. (2012, January). Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. *Nat. Clim. Chang.*, *2*(3), 182–185.

Baret, F., Morissette, J. T., Fernandes, R. A., Champeaux, J. L., Myneni, R. B., Chen, J., … Nickeson, J. E. (2006, July). Evaluation of the representativeness of networks of sites for the global validation and intercomparison of land biophysical products: proposition of the CEOS-BELMANIP. *IEEE Trans. Geosci. Remote Sens.*, *44*(7), 1794–1803.

Bartholome, E., & Belward, A. S. (2005). GLC2000: a new approach to global land cover mapping from earth observation data. *Int. J. Remote Sens.*, *26*(9), 1959–1977.

Bastos, A., O'Sullivan, M., Ciais, P., & others. (2020). Sources of uncertainty in regional and global terrestrial CO2 exchange estimates. *Global*.

Batjes, N. H. (1996, June). Total carbon and nitrogen in the soils of the world. *Eur. J. Soil Sci.*, *47*(2), 151–163.

Besnard, S., Carvalhais, N., Altaf Arain, M., Black, A., de Bruin, S., Buchmann, N., … Reichstein, M. (2018, December). Quantifying the effect of forest age in annual net forest carbon balance. *Environ. Res. Lett.*, *13*(12), 124018.

Bonan, G. (2019). *Climate change and terrestrial ecosystem modeling.* Cambridge University Press.

Bonan, G. B., Lombardozzi, D. L., Wieder, W. R., Oleson, K. W., Lawrence, D. M., Hoffman, F. M., & Collier, N. (2019, October). Model structure and climate data uncertainty in historical simulations of the terrestrial carbon cycle (1850–2014). *Global Biogeochem. Cycles*, *33*(10), 1310–1326.

Chevallier, F. (2013). On the parallelization of atmospheric inversions of CO 2 surface fluxes within a variational framework. *Geoscientific Model Development*, *6*(3), 783–790.

Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., ... Thornton, P. (2013). Carbon and other biogeochemical cycles [Book Section]. In T. Stocker et al. (Eds.), *Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change* (pp. 465–570). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. Retrieved from www.climatechange2013.org doi: 10.1017/CBO9781107415324.015

Claverie, M., Matthews, J. L., Vermote, E. F., & Justice, C. O. (2016, March). A 30+ year AVHRR LAI and FAPAR climate data record: Algorithm description and validation. *Remote Sensing*, *8*(3), 263.

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., ... Randerson, J. T. (2018). The international land model benchmarking (ilamb) system: design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2731–2754.

Covey, C., AchutaRao, K. M., Fiorino, M., Gleckler, P. J., Taylor, K. E., & Wehner, M. F. (2002). Intercomparison of climate data sets as a measure of observational uncertainty. *PCMDI Rep*, *69*.

Dai, A., & Trenberth, K. E. (2002, December). Estimates of freshwater discharge from continents: Latitudinal and seasonal variations. *J. Hydrometeorol.*, *3*(6), 660–687.

De Kauwe, M. G., Keenan, T. F., Medlyn, B. E., Prentice, I. C., & Terrer, C. (2016, October). Satellite based estimates underestimate the effect of CO2 fertilization on net primary productivity. *Nat. Clim. Chang.*, *6*(10), 892–893.

Delire, C., Séférian, R., Decharme, B., Alkama, R., Calvet, J., Carrer, D., ... Tzanos, D. (2020, September). The global land carbon cycle simulated with ISBA-CTRIP: Improvements over the last decade. *J. Adv. Model. Earth Syst.*, *12*(9).

Dixon, R. K., Solomon, A. M., Brown, S., Houghton, R. A., Trexier, M. C., & Wisniewski, J. (1994, January). Carbon pools and flux of global forest ecosystems. *Science*, *263*(5144), 185–190.

Fang, H., Wei, S., Jiang, C., & Scipal, K. (2012, September). Theoretical uncertainty analysis of global MODIS, CYCLOPES, and GLOBCARBON LAI products using a triple collocation method. *Remote Sens. Environ.*, *124*, 610–621.

Fernández-Martínez, M., Sardans, J., Chevallier, F., Ciais, P., Obersteiner, M., Vicca, S., ... Peñuelas, J. (2019, January). Global trends in carbon sinks and their relationships with CO2 and temperature. *Nat. Clim. Chang.*, *9*(1), 73–79.

Fisher, R. A., & Koven, C. D. (2020, April). Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *J. Adv. Model. Earth Syst.*, *12*(4).

Forzieri, G., Duveiller, G., Georgievski, G., Li, W., Robertson, E., Kautz, M., ... Cescatti, A. (2018, May). Evaluating the interplay between biophysical processes and leaf area changes in land surface models. *J Adv Model Earth Syst*, *10*(5), 1102–1126.

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Hauck, J., Olsen, A., ... Zaehle, S. (2020, December). Global carbon budget 2020. *Earth Syst. Sci. Data*, *12*(4), 3269–3340.

Garrigues, S., Lacaze, R., Baret, F., Morisette, J. T., Weiss, M., Nickeson, J. E., ... Others (2008). Validation and intercomparison of global leaf area index products derived from remote sensing data. *Journal of Geophysical Research: Biogeosciences*, *113*(G2).

Goll, D. S., Vuichard, N., Maignan, F., Jornet-Puig, A., Sardans, J., Violette, A., ... Ciais, P. (2017, October). A representation of the phosphorus cycle for

ORCHIDEE (revision 4520). *Geosci. Model Dev.*, *10*(10), 3745–3770.

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Pak, B. C., Baker, D., ... Taguchi, S. (2004, March). Transcom 3 inversion intercomparison: Model mean results for the estimation of seasonal carbon sources and sinks. *Global Biogeochem. Cycles*, *18*(1).

Harris, I., Jones, P. D., Osborn, T. J., & others. (2014). Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 dataset. *International journal of climatology*, *34*(3), 623–642.

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., ... Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*(2), e0169748.

Hobeichi, S., Abramowitz, G., & Evans, J. (2019). Conserving land-atmosphere synthesis suite (CLASS). *J. Clim.*(2019).

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., van der Linden, P. J., Dai, X., ... Johnson, C. A. (2001). *Climate change 2001: the scientific basis.* The Press Syndicate of the University of Cambridge.

Hourdin, F., Musat, I., Bony, S., Braconnot, P., Codron, F., Dufresne, J.-L., ... Lott, F. (2006, October). The LMDZ4 general circulation model: climate performance and sensitivity to parametrized physics with emphasis on tropical convection. *Clim. Dyn.*, *27*(7-8), 787–813.

Huang, Y., Ciais, P., Santoro, M., Makowski, D., Chave, J., Schepaschenko, D., ... Piao, S. (2021). A global map of root biomass across the world's forests. *Earth System Science Data*, *13*(9), 4263–4274. Retrieved from `https://essd.copernicus.org/articles/13/4263/2021/` doi: 10.5194/essd-13-4263-2021

Huijnen, V., Williams, J., van Weele, M., van Noije, T., Krol, M., Dentener, F., ... Pätz, H.-W. (2010, October). The global chemistry transport model TM5: description and evaluation of the tropospheric chemistry version 3.0. *Geosci. Model Dev.*, *3*(2), 445–473.

Hurtt, G. C., Chini, L., Sahajpal, R., Frolking, S., Bodirsky, B. L., Calvin, K., ... Zhang, X. (2020, November). Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6. *Geosci. Model Dev.*, *13*(11), 5425–5464.

Jacobson, A. R., Schuldt, K. N., Miller, J. B., & Oda, T. (n.d.). *CarbonTracker documentation CT2019 release.* `https://gml.noaa.gov/ccgg/carbontracker/CT2019/CT2019_doc.pdf`. (Accessed: 2021-5-20)

Jacobson, A. R., Schuldt, K. N., Miller, J. B., Oda, T., Tans, P., Arlyn Andrews, ... Miroslaw Zimnoch (2020). *Carbontracker ct2019.* NOAA Earth System Research Laboratory, Global Monitoring Division. Retrieved from `https://www.esrl.noaa.gov/gmd/ccgg/carbontracker/CT2019/` doi: 10.25925/39M3-6069

JCGM. (2008). *Jcgm.* `https://www.iso.org/sites/JCGM/GUM/JCGM100/C045315e-html/C045315e_FILES/MAIN_C045315e/02_e.html`. (Accessed: 2021-10-4)

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., ... Reichstein, M. (2019, May). The FLUXCOM ensemble of global land-atmosphere energy fluxes. *Sci Data*, *6*(1), 74.

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., ... others (2020). Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the fluxcom approach. *Biogeosciences*, *17*(5), 1343–1365.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J. J., Fiorino, M., & Potter, G. L. (2002). Ncep–doe amip-ii reanalysis (r-2). *Bull. Am. Meteorol. Soc.*, *83*(11), 1631–1644.

Kato, E., Kinoshita, T., Ito, A., Kawamiya, M., & Yamagata, Y. (2013, March). Evaluation of spatially explicit emission scenario of land-use change and biomass burning using a process-based biogeochemical model. *J. Land Use Sci.*, *8*(1), 104–122.

Knyazikhin, Y., Martonchik, J. V., Myneni, R. B., Diner, D. J., & Running, S. W. (1998). Synergistic algorithm for estimating vegetation canopy leaf area index and fraction of absorbed photosynthetically active radiation from MODIS and MISR data. *J. Geophys. Res. D: Atmos.*, *103*(D24), 32257–32275.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., ... Takahashi, K. (2015). The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan*, *93*(1), 5–48.

Kondo, M., Patra, P. K., Sitch, S., & others. (2020). State of the science in reconciling top-down and bottom-up approaches for terrestrial CO2 budget. *Glob. Chang. Biol.*.

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., ... Prentice, I. C. (2005, March). A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochem. Cycles*, *19*(1).

Lasslop, G., Reichstein, M., Papale, D., Richardson, A. D., Arneth, A., Barr, A., ... Wohlfahrt, G. (2010). Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. *Glob. Chang. Biol.*, *16*(1), 187–208.

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., ... Zeng, X. (2019, December). The community land model version 5: Description of new features, benchmarking, and impact of forcing uncertainty. *J. Adv. Model. Earth Syst.*, *11*(12), 4245–4287.

Li, X., & Xiao, J. (2019, October). Mapping photosynthesis solely from Solar-Induced chlorophyll fluorescence: A global, Fine-Resolution dataset of gross primary production derived from OCO-2. *Remote Sensing*, *11*(21), 2563.

Lienert, S., & Joos, F. (2018, May). A bayesian ensemble data assimilation to constrain model parameters and land-use carbon emissions. *Biogeosciences*, *15*(9), 2909–2930.

Masarie, K. A., Peters, W., Jacobson, A. R., & Tans, P. P. (2014). ObsPack: a framework for the preparation, delivery, and attribution of atmospheric greenhouse gas measurements. *Earth Syst. Sci. Data*, *6*(2), 375–384.

Meiyappan, P., Jain, A. K., & House, J. I. (2015, September). Increased influence of nitrogen limitation on CO 2 emissions from future land use and land use change. *Global Biogeochem. Cycles*, *29*(9), 1524–1548.

Melton, J. R., Arora, V. K., Wisernig-Cojoc, E., Seiler, C., Fortier, M., Chan, E., & Teckentrup, L. (2020). Classic v1.0: the open-source community successor to the canadian land surface scheme (class) and the canadian terrestrial ecosystem model (ctem) – part 1: Model framework and site-level performance. *Geoscientific Model Development*, *13*(6), 2825–2850. Retrieved from `https://gmd.copernicus.org/articles/13/2825/2020/` doi: 10.5194/gmd-13-2825-2020

Merchant, C. J., Paul, F., Popp, T., Ablain, M., Bontemps, S., Defourny, P., ... Wagner, W. (2017, July). Uncertainty information in climate data records from earth observation. *Earth Syst. Sci. Data*, *9*(2), 511–527.

Mokany, K., Raison, R. J., & Prokushkin, A. S. (2006, January). Critical analysis of root : shoot ratios in terrestrial biomes. *Glob. Chang. Biol.*, *12*(1), 84–96.

Myneni, R., Knyazikhin, Y., & Park, T. (2015). MOD15A2H MODIS/terra leaf area index/FPAR 8-day L4 global 500 m SIN grid V006. *NASA EOSDIS Land Processes DAAC*.

Myneni, R. B., Hoffman, S., Knyazikhin, Y., Privette, J. L., Glassy, J., Tian, Y., ... Running, S. W. (2002, November). Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sens. Environ.*, *83*(1), 214–231.

Norby, R. J., De Kauwe, M. G., Domingues, T. F., Duursma, R. A., Ellsworth, D. S., Goll, D. S., … Zaehle, S. (2016, January). Model-data synthesis for the next generation of forest free-air CO2 enrichment (FACE) experiments. *New Phytol.*, *209*(1), 17–28.

Pan, S., Pan, N., Tian, H., Friedlingstein, P., Sitch, S., Shi, H., … Running, S. W. (2020, March). Evaluation of global terrestrial evapotranspiration using state-of-the-art approaches in remote sensing, machine learning and land surface modeling. *Hydrol. Earth Syst. Sci.*, *24*(3), 1485–1509.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., … Papale, D. (2020, July). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Sci Data*, *7*(1), 225.

Piao, S., Wang, X., Wang, K., Li, X., Bastos, A., Canadell, J. G., … Sitch, S. (2020, January). Interannual variation of terrestrial carbon cycle: Issues and perspectives. *Glob. Chang. Biol.*, *26*(1), 300–318.

Potter, C. S., Randerson, J. T., Field, C. B., Matson, P. A., Vitousek, P. M., Mooney, H. A., & Klooster, S. A. (1993, December). Terrestrial ecosystem production: A process model based on global satellite and surface data. *Global Biogeochem. Cycles*, *7*(4), 811–841.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., … Valentini, R. (2005, September). On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Glob. Chang. Biol.*, *11*(9), 1424–1439.

Reick, C., Gayler, V., Goll, D., Hagemann, S., Heidkamp, M., Nabel, J., … Schnur R., S., Wilkenskjeld (2021). *JSBACH 3 - the land component of the MPI earth system model: documentation of version 3.2* (Tech. Rep.). MPI für Meteorologie.

Rödenbeck, C., Zaehle, S., Keeling, R., & Heimann, M. (2018, April). How does the terrestrial carbon exchange respond to inter-annual climatic variations? a quantification based on atmospheric CO_2 data. *Biogeosciences*, *15*(8), 2481–2498.

Roy, J., Saugier, B., & Mooney, H. A. (2001). *Terrestrial global productivity.* Academic Press.

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. A., Salas, W., … Others (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. *Proceedings of the national academy of sciences*, *108*(24), 9899–9904.

Santoro, M., Beaudoin, A., Beer, C., Cartus, O., Fransson, J. E. S., Hall, R. J., … Wegmüller, U. (2015, October). Forest growing stock volume of the northern hemisphere: Spatially explicit estimates for 2010 derived from envisat ASAR. *Remote Sens. Environ.*, *168*, 316–334.

Santoro, M., Cartus, O., Carvalhais, N., Rozendaal, D. M. A., Avitabile, V., Araza, A., … Willcock, S. (2021, August). The global forest above-ground biomass pool for 2010 estimated from high-resolution satellite observations. *Earth Syst. Sci. Data*, *13*(8), 3927–3950.

Schepaschenko, D., Chave, J., Phillips, O. L., Lewis, S. L., Davies, S. J., Réjou-Méchain, M., … Zo-Bi, I. C. (2019, October). The forest observation system, building a global reference dataset for remote sensing of forest biomass. *Sci Data*, *6*(1), 198.

Seiler, C. (2020). Amber: Automated model benchmarking r package [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=amber` (R package version 1.0.3)

Seiler, C., Melton, J. R., Arora, V. K., & Wang, L. (2021, May). CLASSIC v1.0: the open-source community successor to the canadian land surface scheme (CLASS) and the canadian terrestrial ecosystem model (CTEM) – part 2: Global benchmarking. *Geoscientific Model Development*, *14*(5), 2371–2417.

Smith, B., Wårlind, D., Arneth, A., Hickler, T., Leadley, P., Siltberg, J., & Zaehle, S. (2014, November). Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model. *Biogeosci. Discuss.*, *10*(11), 18613–18685.

Tian, H., Chen, G., Lu, C., Xu, X., Hayes, D. J., Ren, W., … Wofsy, S. C. (2015). North american terrestrial CO2 uptake largely offset by CH4 and N2O emissions: toward a full accounting of the greenhouse gas budget. *Clim. Change*, *129*(3-4), 413–426.

Tian, H., Lu, C., Yang, J., Banger, K., Huntzinger, D. N., Schwalm, C. R., … Zeng, N. (2015, June). Global patterns and controls of soil organic carbon dynamics as simulated by multiple terrestrial biosphere models: Current status and future directions. *Global Biogeochem. Cycles*, *29*(6), 775–792.

Tifafi, M., Guenet, B., & Hatté, C. (2018). Large differences in global and regional total soil carbon stock estimates based on SoilGrids, HWSD, and NCSCD: Intercomparison and evaluation based on field data from USA, england, wales, and france. *Global Biogeochem. Cycles*, *32*(1), 42–56.

Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E. A. G., & Allison, S. D. (2013, March). Causes of variation in soil carbon simulations from CMIP5 earth system models and comparison with observations. , *10*(3), 1717–1736.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., … Papale, D. (2016, July). Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences*, *13*(14), 4291–4313.

van der Werf, G. R., Randerson, J. T., Giglio, L., van Leeuwen, T. T., Chen, Y., Rogers, B. M., … Kasibhatla, P. S. (2017, September). Global fire emissions estimates during 1997–2016. *Earth Syst. Sci. Data*, *9*(2), 697–720.

Verger, A., Baret, F., & Weiss, M. (2014). Near real-time vegetation monitoring at global scale. *IEEE Journal of Selected Topics in*.

Vuichard, N., Messina, P., Luyssaert, S., Guenet, B., Zaehle, S., Ghattas, J., … Peylin, P. (2019). Accounting for carbon and nitrogen interactions in the global terrestrial ecosystem model ORCHIDEE (trunk version, rev 4999): multi-scale evaluation of gross primary production. *Geoscientific Model Development*, *12*(11), 4751–4779.

Walker, A. P., Quaife, T., Bodegom, P. M., De Kauwe, M. G., Keenan, T. F., Joiner, J., … Woodward, F. I. (2017, September). The impact of alternative trait-scaling hypotheses for the maximum photosynthetic carboxylation rate (V cmax) on global gross primary production. *New Phytol.*, *215*(4), 1370–1386.

Wieder, W. (2014). *Regridded harmonized world soil database v1.2.* ORNL Distributed Active Archive Center. Retrieved from `http://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1247` doi: 10.3334/ORNLDAAC/1247

Xue, B.-L., Guo, Q., Hu, T., Wang, G., Wang, Y., Tao, S., … Zhao, X. (2017, July). Evaluation of modeled global vegetation carbon dynamics: Analysis based on global carbon flux and above-ground biomass data. *Ecol. Modell.*, *355*, 84–96.

Yuan, W., Liu, D., Dong, W., Liu, S., Zhou, G., Yu, G., … Zhao, L. (2014, May). Multiyear precipitation reduction strongly decreases carbon uptake over northern china. *J. Geophys. Res. Biogeosci.*, *119*(5), 881–896.

Zaehle, S., & Friend, A. D. (2010). Carbon and nitrogen cycle dynamics in the O-CN land surface model: 1. model description, site-scale evaluation, and sensitivity to parameter estimates. *Global Biogeochem. Cycles*, *24*(1).

Zhang, Y., & Liang, S. (2020, August). Fusion of multiple gridded biomass datasets for generating a global forest aboveground biomass map. *Remote Sensing*, *12*(16), 2559.

Zhang, Y., Xiao, X., Wu, X., Zhou, S., Zhang, G., Qin, Y., & Dong, J. (2017, October). A global moderate resolution dataset of gross primary production of vegetation for 2000–2016. *Scientific Data*, *4*(1), 170165.

Table B1. Globally summed mean values and corresponding biases

Variable	Ref. ID	Model ID	Ref.	Model	Bias	Bias (%)	Unit	Period
NBP	CAMS	CLASSIC	1.86	0.82	-1.04	-55.91	PgC yr^{-1}	1979-2017
NBP	CAMS	CLM5.0	1.90	0.68	-1.22	-64.21	PgC yr^{-1}	1979-2019
NBP	CAMS	IBIS	1.60	0.74	-0.86	-53.75	PgC yr^{-1}	1979-2019
NBP	CAMS	ISAM	1.88	0.94	-0.94	-50.00	PgC yr^{-1}	1979-2019
NBP	CAMS	ISBA-CTRIP	1.89	1.19	-0.70	-37.04	PgC yr^{-1}	1979-2019
NBP	CAMS	JSBACH	1.80	1.01	-0.79	-43.89	PgC yr^{-1}	1979-2019
NBP	CAMS	LPX-Bern	1.90	0.40	-1.50	-78.95	PgC yr^{-1}	1979-2019
NBP	CAMS	OCN	1.86	1.51	-0.35	-18.82	PgC yr^{-1}	1979-2019
NBP	CAMS	ORCHIDEE	1.90	1.46	-0.44	-23.16	PgC yr^{-1}	1979-2019
NBP	CAMS	ORCHIDEE-CNP	1.91	0.26	-1.65	-86.39	PgC yr^{-1}	1979-2019
NBP	CAMS	ORCHIDEEv3	1.91	1.34	-0.57	-29.84	PgC yr^{-1}	1979-2019
NBP	CAMS	SDGVM	1.87	1.30	-0.57	-30.48	PgC yr^{-1}	1979-2019
NBP	CAMS	VISIT	1.85	1.26	-0.59	-31.89	PgC yr^{-1}	1979-2019
NBP	CT2019	CLASSIC	1.33	1.17	-0.16	-12.03	PgC yr^{-1}	2000-2017
NBP	CT2019	CLM5.0	1.33	0.80	-0.53	-39.85	PgC yr^{-1}	2000-2018
NBP	CT2019	IBIS	1.17	0.97	-0.20	-17.09	PgC yr^{-1}	2000-2018
NBP	CT2019	ISAM	1.31	0.91	-0.40	-30.53	PgC yr^{-1}	2000-2018
NBP	CT2019	ISBA-CTRIP	1.32	1.24	-0.08	-6.06	PgC yr^{-1}	2000-2018
NBP	CT2019	JSBACH	1.32	1.23	-0.09	-6.82	PgC yr^{-1}	2000-2018
NBP	CT2019	LPX-Bern	1.32	0.62	-0.70	-53.03	PgC yr^{-1}	2000-2018
NBP	CT2019	OCN	1.34	1.83	0.49	36.57	PgC yr^{-1}	2000-2018
NBP	CT2019	ORCHIDEE	1.33	1.74	0.41	30.83	PgC yr^{-1}	2000-2018
NBP	CT2019	ORCHIDEE-CNP	1.33	0.24	-1.09	-81.95	PgC yr^{-1}	2000-2018
NBP	CT2019	ORCHIDEEv3	1.33	1.44	0.11	8.27	PgC yr^{-1}	2000-2018
NBP	CT2019	SDGVM	1.33	1.67	0.34	25.56	PgC yr^{-1}	2000-2018
NBP	CT2019	VISIT	1.32	1.79	0.47	35.61	PgC yr^{-1}	2000-2018
NBP	CarboScope	CLASSIC	1.46	1.40	-0.06	-4.11	PgC yr^{-1}	1999-2017
NBP	CarboScope	CLM5.0	1.38	0.90	-0.48	-34.78	PgC yr^{-1}	1999-2019
NBP	CarboScope	IBIS	1.18	1.07	-0.11	-9.32	PgC yr^{-1}	1999-2019
NBP	CarboScope	ISAM	1.29	0.94	-0.35	-27.13	PgC yr^{-1}	1999-2019
NBP	CarboScope	ISBA-CTRIP	1.40	1.41	0.01	0.71	PgC yr^{-1}	1999-2019
NBP	CarboScope	JSBACH	1.14	1.33	0.19	16.67	PgC yr^{-1}	1999-2019
NBP	CarboScope	LPX-Bern	1.36	0.65	-0.71	-52.21	PgC yr^{-1}	1999-2019
NBP	CarboScope	OCN	1.25	1.88	0.63	50.40	PgC yr^{-1}	1999-2019
NBP	CarboScope	ORCHIDEE	1.37	1.83	0.46	33.58	PgC yr^{-1}	1999-2019
NBP	CarboScope	ORCHIDEE-CNP	1.46	0.30	-1.16	-79.45	PgC yr^{-1}	1999-2019
NBP	CarboScope	ORCHIDEEv3	1.46	1.54	0.08	5.48	PgC yr^{-1}	1999-2019
NBP	CarboScope	SDGVM	1.30	1.73	0.43	33.08	PgC yr^{-1}	1999-2019
NBP	CarboScope	VISIT	1.27	1.88	0.61	48.03	PgC yr^{-1}	1999-2019