Geosci. Model Dev., 14, 2371–2417, 2021
https://doi.org/10.5194/gmd-14-2371-2021

Geoscientific Model Development
Open Access
EGU

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

Christian Seiler[1], **Joe R. Melton**[1], **Vivek K. Arora**[2], **and Libo Wang**[3]

[1]Climate Processes Section, Climate Research Division, Environment and Climate Change Canada, Victoria, BC, Canada
[2]Canadian Centre for Climate Modelling and Analysis, Climate Research Division, Environment and Climate Change Canada, Victoria, BC, Canada
[3]Climate Processes Section, Climate Research Division, Environment and Climate Change Canada, North York, ON, Canada

Correspondence: Christian Seiler (christian.seiler@canada.ca)

Received: 1 September 2020 – Discussion started: 28 October 2020
Revised: 13 March 2021 – Accepted: 22 March 2021 – Published: 3 May 2021

Abstract. The Canadian Land Surface Scheme Including Biogeochemical Cycles (CLASSIC) is an open-source community model designed to address research questions that explore the role of the land surface in the global climate system. Here, we evaluate how well CLASSIC reproduces the energy, water, and carbon cycle when forced with quasi-observed meteorological data. Model skill scores summarize how well model output agrees with observation-based reference data across multiple statistical metrics. A lack of agreement may be due to deficiencies in the model, its forcing data, and/or reference data. To address uncertainties in the forcing, we evaluate an ensemble of CLASSIC runs that is based on three meteorological data sets. To account for observational uncertainty, we compute benchmark skill scores that quantify the level of agreement among independent reference data sets. The benchmark scores demonstrate what score values a model may realistically achieve given the uncertainties in the observations. Our results show that uncertainties associated with the forcing and observations are considerably large. For instance, for 10 out of 19 variables assessed in this study, the sign of the bias changes depending on what forcing and reference data are used. Benchmark scores are much lower than expected, implying large observational uncertainties. Model and benchmark score values are mostly similar, indicating that CLASSIC performs well when considering observational uncertainty. Future model development should address (i) a positive albedo bias and resulting shortwave radiation bias in parts of the Northern Hemisphere (NH) extratropics and Tibetan Plateau, (ii) an out-of-phase seasonal gross primary productivity cycle in the humid tropics of South America and Africa, (iii) a lacking spatial correlation of annual mean net ecosystem exchange with site-level measurements, (iv) an underestimation of fractional area burned and corresponding emissions in the boreal forests, (v) a negative soil organic carbon bias in high latitudes, and (vi) a time lag in seasonal leaf area index maxima in parts of the NH extratropics. Our results will serve as a baseline for guiding and monitoring future CLASSIC development.

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

1 Introduction

The land surface interacts with the atmosphere through fluxes of momentum, radiation, heat, and mass, the latter including water, trace gases, and aerosols. Land surface models (LSMs) have been developed to simulate these fluxes in global climate models. Early LSMs consisted of simple aerodynamic bulk transfer formulas with prescribed surface parameters that are uniform across continents (Manabe, 1969; Manabe and Bryan, 1969). Major advancements in LSM development have occurred since then, including the incorporation of (i) vegetation effects on the surface energy balance (Dickinson, 1984), (ii) vegetation phenology (Sellers et al., 1996), (iii) ecological processes (Foley et al., 1996), and (iv) nutrient cycles (Goll et al., 2012).

The Canadian Land Surface Scheme Including Biogeochemical Cycles (CLASSIC; v1.0; Melton et al., 2020) is a state-of-the-art land surface model primarily developed by Environment and Climate Change Canada. CLASSIC is the open-source community model successor to the CLASS-CTEM modeling framework, which consists of the Canadian Land Surface Scheme (CLASS) and the Canadian Terrestrial Ecosystem Model (CTEM). CLASS and CTEM simulate physical and biogeochemical land surface processes, respectively, and together they form the land component of the Canadian Earth System Model (Swart et al., 2019).

Assessing model performance provides invaluable guidance for future model development. Strategies for LSM evaluation have been developed in a number of collaborative research projects, such as the Project for Intercomparison of Land Surface Parameterization Schemes (PILPS) (Henderson-Sellers et al., 1993), the Global Land-Atmosphere Coupling Experiment (GLACE) (Koster et al., 2006), the Protocol for the Analysis of Land Surface Models (PALS), the Land Surface Model Benchmarking Evaluation Project (PLUMBER) (Best et al., 2015), and the International Land Model Benchmarking (ILAMB) project (Collier et al., 2018). The ILAMB approach summarizes model performance across multiple statistical metrics using a dimensionless skill score that ranges from zero to unity, where higher values imply better performance (see Sect. 2 for details).

The ILAMB framework has been adopted by the Global Carbon Project to evaluate the performance of LSMs used for estimating global carbon budgets (Friedlingstein et al., 2019). The project assessed 15 LSMs with respect to gross primary productivity (GPP), ecosystem respiration, net ecosystem exchange (NEE), biomass, soil carbon, leaf area index (LAI), evapotranspiration, and runoff. For CLASS-CTEM, which was one of the 15 models, performance was considerably weaker for LAI and NEE than for other variables. While a low LAI score was also found for most other LSMs, the NEE score was below the multi-model mean value. The variables for which CLASS-CTEM performed best and worst with respect to the multi-model mean were GPP and biomass, respectively. However, skill scores must be interpreted with caution. For instance, low score values may actually reflect uncertainties in model inputs and/or reference data rather than deficiencies in the model. A robust analysis should therefore account for uncertainties associated with the forcing and reference data.

Lawrence et al. (2019) applied ILAMB to quantify how model development affected model performance for the Community Land Model (CLM). While the authors found a general improvement from CLM version 4 to version 5, they also identified variables where model performance would improve or worsen, depending on the choice of reference data. A second source of uncertainty identified by the authors was related to the choice of meteorological forcing. Driving CLM with three different data sets (GSWP3, CRUNCEPv7, and WATCH/WFDEI) showed that model performance was generally better when forcing CLM with the Global Soil Wetness Project forcing data set (GSWP3), and that the uncertainty associated with the forcing is too large to be neglected (Bonan et al., 2019).

A first evaluation of CLASSIC has been presented by Melton et al. (2020), who ran the model for 31 FLUXNET sites using locally observed meteorological data and vegetation composition. Model outputs were compared against observed net surface radiation (RNS), latent heat flux (HFLS), sensible heat flux (HFSS), GPP, ecosystem respiration (RECO), and NEE. Results showed that CLASSIC reproduced the RNS, HFLS, HFSS, GPP, and RECO observations reasonably well. Reproducing NEE fluxes proved to be more challenging. While for some sites modeled NEE fluxes were in good agreement with measurements (e.g. site FI-Hyy; Figs. S22 and S23 in Melton et al., 2020), for most sites NEE differed considerably from observations. Assessing the annual mean NEE value for each station showed that modeled and observed annual mean values were not correlated across stations. The poor model performance for NEE may be related to the fact that the model does not account for the disturbance history nor the climate history of a given site. A comparison between results presented in the present study and Melton et al. (2020) is provided in the Discussion section.

This present paper provides a comprehensive assessment of model performance on a global scale. Our assessment is based on ILAMB's statistical framework, which we implemented in a new R package referred to as the Automated Model Benchmarking R package (AMBER) (Seiler, 2020). The development of AMBER allowed us to tailor the ILAMB approach to CLASSIC model output. The resulting benefits include (i) a seamless data ingestion that does not require any pre-processing, (ii) the ability to evaluate CLASSIC in different simulation modes (i.e. global runs on a regular grid, regional runs on a rotated grid, and site-level runs), (iii) full control on how the statistical framework is implemented, and (iv) the addition of new functionalities. Our analysis pays

special attention to how model performance is affected by uncertainties in the forcing and reference data.

The remainder of the paper is organized as follows. Section 2 describes the forcing data, reference data, model structure, and benchmarking approach. Section 3 assesses model performance for an ensemble of three model runs that differ in their meteorological forcing. For most variables, each ensemble member is compared against more than one observation-based reference data set. For each variable, we assess global, zonal, and seasonal biases, as well as root-mean-square errors, time lags in seasonal peaks, interannual variability, spatial patterns, and corresponding skill scores. Furthermore, we compare model skill scores against benchmark scores, which are based on reference data only, to assess the extent to which score values are affected by observational uncertainty. Section 4 elaborates on our principal findings, with a particular focus on LAI and NEE, and the importance of in situ measurements.

2 Methods

2.1 Canadian Land Surface Scheme Including Biogeochemical Cycles (CLASSIC)

CLASSIC is the open-source community model successor to the CLASS-CTEM modeling framework, which consists of the Canadian Land Surface Scheme (CLASS) and the Canadian Terrestrial Ecosystem Model (CTEM). CLASS and CTEM are two fully integrated model components that simulate physical and biogeochemical land surface processes, respectively. A brief outline of these two components is provided next.

CLASSIC's physical component (CLASS) divides the land surface into four possible subareas, namely bare ground, ground covered by snow, ground covered by canopy, and snow covered by canopy. The model solves the energy balance for the canopy layer and underlying ground, separately. The canopy consists of a single layer and is for the results presented here composed of four plant functional types (PFTs), namely needleleaf trees, broadleaf trees, grasses, and crops. The model output evaluated here is from the offline mode in which we use 20 soil layers ranging from 10 layers of 0.1 m thickness, gradually increasing to a 30 m thick layer for a total ground depth of over 61 m. Water fluxes are calculated for soil layers within the permeable soil depth of the ground column but not the underlying bedrock layers, whereas temperatures are calculated for both soil and bedrock layers. Also calculated are the temperature, mass, albedo, and density of the single-layer snowpack, the temperature and interception and storage of rain and snow on the vegetation canopy, and the temperature and depth of ponded water on the ground surface.

CLASSIC's biogeochemical component (CTEM) is a dynamic vegetation model coupled to CLASS. CLASS provides CTEM physical land surface information, including soil moisture, soil temperature and net radiation. CTEM uses this information to simulate photosynthesis in response to atmospheric CO_2 concentration. CTEM considers three live vegetation components (leaves, stem, and roots) and two dead carbon pools (litter and soil). These five carbon pools are tracked, in the default CLASSIC configuration used here, for nine PFTs that map directly onto the four PFTs used by CLASS. Needleleaf trees are divided into their deciduous and evergreen types, broadleaf trees are divided into cold and drought deciduous and evergreen types, and crops and grasses are divided based on their photosynthetic pathways into C_3 and C_4 versions. In this study, the spatial distribution of PFTs has been prescribed to reduce the number of possible causes for model biases. The prognostic carbon masses of leaf, stem, and root simulated by CTEM are used to calculate the structural vegetation characteristics required by CLASS. The main processes simulated by CTEM include photosynthesis, canopy conductance, tissue turnover, allocation of carbon, and phenology (Arora and Boer, 2005b); dynamic root distribution (Arora and Boer, 2003); maintenance, growth, and heterotrophic respiration (Melton et al., 2015); wildfires (Arora and Boer, 2005a; Arora and Melton, 2018); competition for space between PFTs (Arora and Boer, 2006; Melton and Arora, 2016), and land use change (Arora and Boer, 2010). For the purpose of this study, CLASSIC's recently added nitrogen cycle has been turned off (Asaadi and Arora, 2021). A more detailed description of the model and its history is documented in Melton et al. (2020). Details on the computation of radiation fluxes, surface albedo, heat fluxes, LAI, GPP, ecosystem respiration, and biomass are provided in the Appendix.

2.2 Meteorological forcing

Meteorological inputs required by CLASSIC include surface downwelling shortwave (SW) and longwave (LW) radiation, near-surface air temperature, precipitation, near-surface specific humidity, surface pressure, and near-surface horizontal wind speed. Model outputs are potentially sensitive to uncertainties in the meteorological forcing. To account for this uncertainty, we drive CLASSIC with three different quasi-observed meteorological data sets, namely the Climate Research Unit – National Centers for Environmental Prediction version 8 (CRUNCEP; 1901–2016; Viovy, 2018), the Climate Research Unit – Japanese 55-year Reanalysis version 2.0 (CRUJRAv2; 1901–2017; Harris et al., 2014; Kobayashi et al., 2015), and the Global Soil Wetness Project Phase 3 (GSWP3) – WFDE5 over land merged with ERA5 over the ocean (W5E5) (GSWP3W5E5; 1901–2016; ISIMIP, 2020). All meteorological forcings are disaggregated from 6-hourly to half-hourly time steps, following the methodology by Melton and Arora (2016).

Zonal mean values suggest that all three forcings are relatively consistent (Fig. A1). The corresponding global mean

Table 1. Observation-based reference data used for model evaluation. Meanings of acronyms are provided in Sect. 2. N/A: not applicable.

Source	Variables	Approach (*n* sites)	Period	Reference
In situ measurements				
FLUXNET2015	RNS, HFLS, HFSS, HFG GPP, RECO, NEE	eddy covariance (204)	1997–2014	Pastorello et al. (2017)
GRDC	MRRO	gauge records (50)	1980–2010	Dai and Trenberth (2002)
Mortimer	SNW	gravimetry (3271)	1970–2017	Mortimer et al. (2020)
CEOS	LAI	transfer function (141)	1999–2017	Garrigues et al. (2008)
ORNL DAAC	LAI	maximum LAI (2653)	1932–2011	Iio and Ito (2014)
FOS	AGB	allometry (274)	1999–2018	Schepaschenko et al. (2019)
Globally gridded data sets				
MODIS	ALBS	BRDF	2000–2014	Strahler et al. (1999)
	GPP	light use efficiency model	2000–2016	Zhang et al. (2017)
	LAI	radiative transfer model	2000–2017	Myneni et al. (2002)
AVHRR	LAI	artificial neural network	1982–2010	Claverie et al. (2016)
CERES	ALBS, RSS, RLS, RNS	radiative transfer model	2000–2012	Kato et al. (2013)
GEWEXSRB	ALBS, RSS, RLS, RNS	radiative transfer model	1984–2007	Stackhouse et al. (2011)
GOSIF	GPP	statistical model	2000–2017	Li and Xiao (2019)
FluxCom	RNS, HFLS, HFSS	machine learning	2001–2013	Jung et al. (2019)
FluxCom	GPP, RECO, NEE	machine learning	1980–2013	Jung et al. (2020)
CLASSr	RNS, HFLS, HFSS, HFG, MRRO	blended product	2003–2009	Hobeichi et al. (2019)
ESA	MRSLL	land surface model	1979–2017	Liu et al. (2011)
ECCC	SNW	blended product	1981–2017	Mudryk (2020)
Brown	SNW	temperature index model	1981–2017	Brown et al. (2003)
Crocus	SNW	snowpack model	1981–2017	Brun et al. (2013)
GS3	SNW	passive microwave	1981–2017	Takala et al. (2011)
MERRA-2	SNW	reanalysis	1981–2017	Gelaro et al. (2017)
GFED4S	BURNT	burned-area mapping	2001–2015	Giglio et al. (2010)
ESACCI	BURNT	burned-area mapping	2001–2017	Chuvieco et al. (2018)
CT2019	NEE, FIRE	inversion model	2000–2017	Jacobson et al. (2020)
GEOCARBON	AGB	machine learning	N/A	Avitabile et al. (2016), Santoro et al. (2015)
Saatchi	CVEG	machine learning	2000	Saatchi et al. (2011)
HWSD	CSOIL	soil inventory	N/A	Wieder (2014) Todd-Brown et al. (2013)

values for CRUJRAv2, GSWP3W5E5, and CRUNCEP are 182, 178, and 183 $W\,m^{-2}$ for downwelling SW radiation, 304, 309, and 300 $W\,m^{-2}$ for downwelling LW radiation, 10 °C for near-surface air temperature (all three forcings), 59, 64, and 60 mm per month precipitation, 6.9, 7.4, and 6.9 × 10^{-3} $kg\,kg^{-1}$ for specific humidity, 947, 948, and 950 hPa surface pressure, and 2.6, 3.3, and 2.6 $m\,s^{-1}$ horizontal wind speed, respectively. With the exception of surface pressure, the data from CRUJRAv2 and GSWP3W5E5 tend to be more similar compared to CRUNCEP.

2.3 Simulation protocol

Our simulations are based on the trends in the land carbon cycle project (TRENDY) simulation protocol with time-varying CO_2, climate, and land use (S3) (Friedlingstein et al., 2019). The protocol consists of a spinup for the year 1700 and two transient runs for the periods 1701–1900 and 1901–2016, respectively. The spinup uses a constant atmospheric CO_2 concentration of 276.59 ppm, climate data from the early decades of the 20th century (i.e. 1901–1920), and a land use map with constant crops and pasture distribution. The first transient run uses the same climate as for the spinup but time-varying CO_2 concentrations and land use for the 1701–1900 period. The second transient run uses time-varying CO_2, climate, and land use for the 1901–2017 period.

2.4 Reference data

Model outputs are evaluated against in situ observations and globally gridded observation-based reference data (Table 1). The reference database covers 19 variables that form part of the energy, water, and carbon cycle. While most data sets consist of monthly mean values, some are assessed on

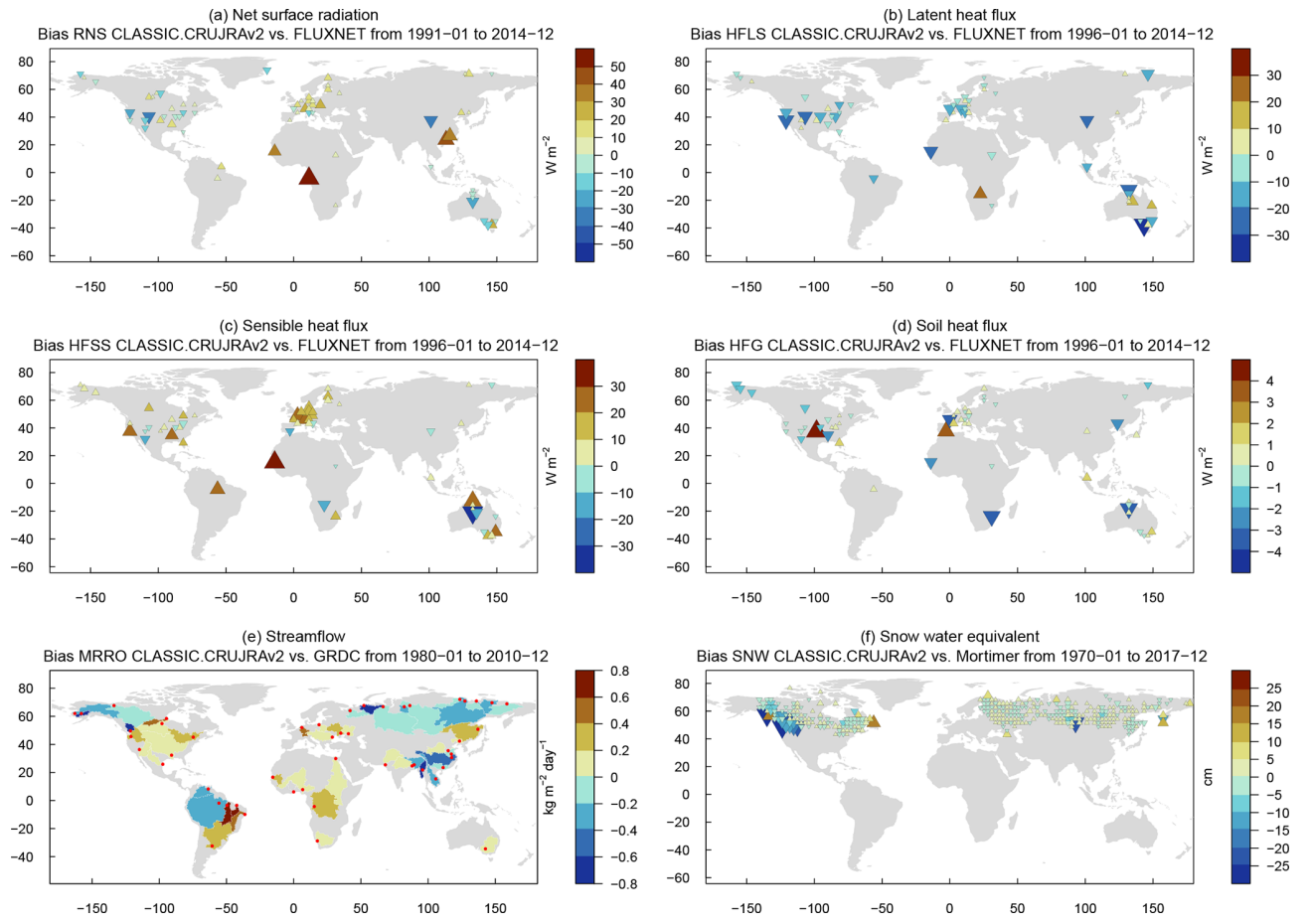

Figure 1. Biases when comparing CLASSIC forced with CRUJRAv2 against in situ measurements of **(a)** net surface radiation, **(b)** latent heat flux, **(c)** sensible heat flux, **(d)** soil heat flux, **(e)** streamflow, and **(f)** snow water equivalent. Measurements located in the same model grid cell were merged. Upward- and downward-pointing triangles indicate positive and negative biases, respectively.

an annual timescale (streamflow) or present a snapshot in time (vegetation biomass, organic soil carbon mass, and in situ LAI measurements provided by the Oak Ridge National Laboratory). All gridded reference data are spatially interpolated to the spatial resolution of a CLASSIC grid cell (2.8125° × 2.8125°). In situ observations that fall into the same model grid cell are averaged prior to the comparison against model output. In most cases, the evaluation is based on the grid cell value, which reflects the average conditions across all land cover types located in a grid cell. For above-ground biomass, however, a comparison between the grid cell value and forest inventory data is not ideal, as the latter provides biomass estimates for trees only. We therefore only consider biomass simulated for trees when evaluating model results against forest inventory data. Details on each reference data set are provided next.

2.4.1 In situ observations

The FLUXNET2015 database includes 204 eddy covariance sites with measurements overlapping during the years 1997 and 2014 (Pastorello et al., 2017) (Table 1). The corresponding variables include RNS, HFLS, HFSS, soil heat flux (HFG) (Fig. 1a–d), GPP, RECO, and NEE (Fig. 2a–c).

Streamflow gauge records are provided by the Global Runoff Data Centre (GRDC) for the world's 50 largest basins (Dai and Trenberth, 2002) (Fig. 1e). Measurements were made sometime between 1980 and 2010, depending on the basin.

Mortimer et al. (2020) compiled manual gravimetric measurements of snow water equivalent (SNW) collected from 3271 sites located between 42 and 83° N for the 1970–2017 period (Fig. 1f).

Above-ground biomass (AGB) measurements were provided by the Forest Observation System (FOS) who compiled measurements from 274 permanent research plots (Schepaschenko et al., 2019) (Fig. 2d). Measurements are based on allometric relations and were made between 1999 and 2018, depending on the inventory plot.

LAI observations are taken from the Committee on Earth Observation Satellites (CEOS) (Garrigues et al., 2008) and

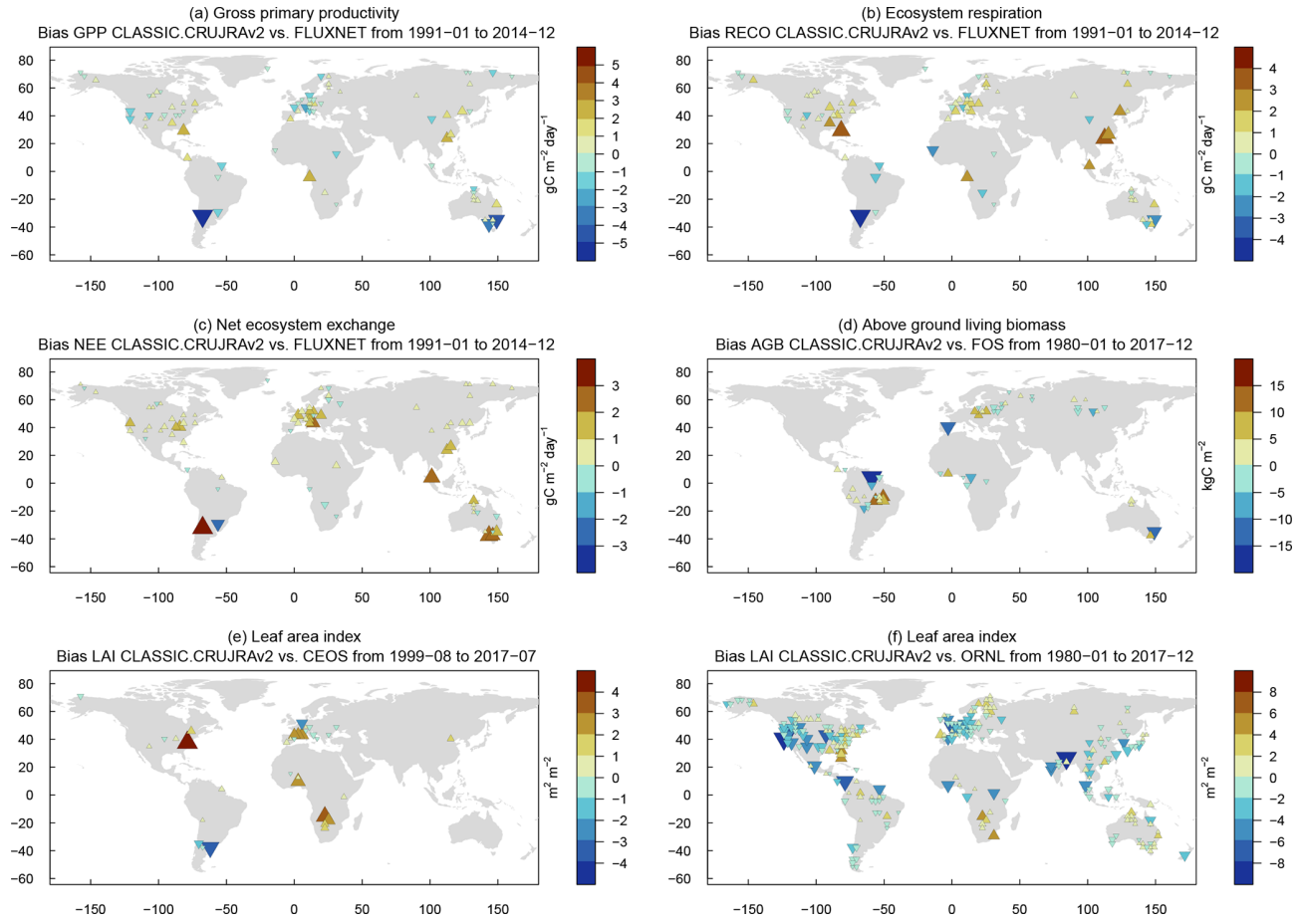

Figure 2. Same as Fig. 1 but for **(a)** gross primary productivity, **(b)** ecosystem respiration, **(c)** net ecosystem exchange, **(d)** above-ground biomass, **(e)** leaf area index, and **(f)** maximum leaf area index.

the Oak Ridge National Laboratory (ORNL) (Iio and Ito, 2014) (Fig. 2e and f). The CEOS database consists of 141 sites with measurements during the 1999–2017 period. 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. The ORNL data set contains a total of 2653 measurements compiled from literature that was published between 1932 and 2012. The values present site-specific maximum ground LAI measurements.

2.4.2 Globally gridded reference data

The Moderate Resolution Imaging Spectroradiometer (MODIS) provides globally gridded data sets for white-sky surface albedo (2000–2014), GPP (2000–2016), and LAI (2000–2017). The white-sky albedo (MCD43C3 v006) integrates the bidirectional reflectance distribution function (BRDF) over all viewing and irradiance directions. The resulting albedo values are properties of the surface that do not depend on the state of the atmosphere (Strahler et al., 1999). Only data with quality flags of less than 3 and solar zenith angles of less than 70° were considered. Evaluating MODIS V6 surface albedo against in situ measurements shows that biases for the high-quality data (full inversion) range from −0.0061 in needleleaf forests to 0.0023 in deserts (Wang et al., 2018).

MODIS-based GPP is provided by Zhang et al. (2017) for the period 2000–2017. The data set 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 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).

MODIS LAI (MOD15A2H, Collection 6; 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). The inversion approach causes LAI to be very sensitive to uncertainties in surface reflectance. Those uncertainties may occur due to calibration errors, residual at-

mospheric and cloud contamination, background underneath the canopy, topography, view-illumination geometry effects, and reflectance saturation in dense canopies (Garrigues et al., 2008). MODIS LAI uncertainties of annual mean values range from 23 % to 38 % across biomes, with a global mean value of 27 % for MODIS LAI Collection 5 (Fang et al., 2012).

A second set of globally gridded LAI was provided by Claverie et al. (2016) for the period 1982–2010. This data set is based on an artificial neural network (ANN) that relates LAI to surface reflectance from the Advanced Very High Resolution Radiometer (AVHRR). The ANN was calibrated with LAI from MODIS (MCD15A2) and in situ LAI measurements 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).

The Clouds and the Earth's Radiant Energy System (CERES) provides globally gridded SW and LW radiation fluxes at the surface under all-sky conditions for the period 2000–2012 (Kato et al., 2013). The data set is based on a radiative transfer model that uses satellite-retrieved surface, cloud, and aerosol properties as inputs. The values of these inputs are adjusted within their uncertainty ranges to obtain irradiance values at the top of the atmosphere that are consistent with observations.

The Global Energy and Water Exchanges Surface Radiation Budget (GEWEXSRB) release 3.0 (Stackhouse et al., 2011) provides globally gridded SW and LW radiation fluxes at the surface under all-sky conditions for the period 1984–2007. The data set is based on a radiative transfer model developed by Pinker and Laszlo (1992), which is driven with satellite observations (cloud parameters and ozone fields) and reanalysis data (Stackhouse et al., 2011). The algorithm scales spectral surface albedos of 12 surface types in such a way that the resulting clear-sky composite broadband top-of-the-atmosphere albedo is consistent with observations from the International Satellite Cloud Climatology Project (Zhang et al., 2019). Surface fluxes have been evaluated against ground-based measurements from sources that include the Baseline Surface Radiation Network and the Global Energy Balance Archive.

The GOSIF data provided by Li and Xiao (2019) are based on linear correlations between solar-induced chlorophyll fluorescence (SIF) soundings from the Orbiting Carbon Observatory-2 (OCO-2) and GPP measurements from 91 eddy covariance measurements sites from FLUXNET.

The FluxCom data set is an ensemble of globally gridded energy fluxes (HFLS, HFSS, and RNS from 2001 to 2013) and carbon fluxes (GPP, NEE, and RECO from 1980 to 2013). Using a range of machine learning methods, FluxCom upscales FLUXNET observations, where remote sensing data and meteorological data serve as global predictors. The former includes land surface temperature (LST; MOD11A226), land cover (MCD12Q127), fraction of absorbed photosynthetically active radiation by a canopy (fPAR; MOD15A228), and BRDF-corrected reflectances (MCD43B429) from MODIS. Meteorological data employed by FluxCom were taken from the WATCH Forcing Data ERA-Interim (WFDEI), Global Soil Wetness Project 3 (GSWP), CRUNCEPv837, Global Precipitation Climatology Project (GPCP), and CERES. The FluxCom values used in the present study are the median values computed over the FluxCom ensembles.

The Conserving Land-Atmosphere Synthesis Suite (CLASSr) provides globally gridded estimates of net surface radiation, latent heat flux, sensible heat flux, soil heat flux, and runoff for the 2003–2009 period (Hobeichi et al., 2019). 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. Net surface radiation, latent heat flux, and sensible heat flux are based on blending data from remote sensing, reanalysis, and land surface models. Soil heat flux estimates are based on reanalysis and land surface models, and runoff is derived from land surface models and hydrological models. The original acronym for the Conserving Land-Atmosphere Synthesis Suite is CLASS. However, in the present paper, we use the acronym CLASSr instead to avoid confusion with the Canadian Land Surface Scheme acronym.

Globally gridded soil moisture estimates provided by Liu et al. (2011) are based on the retrieval characteristics of passive (Advanced Microwave Scanning Radiometer – Earth Observing System; AMSR-E) and active (Advanced Scatterometer; ASCAT) microwave satellite data. Remotely sensed estimates are rescaled using the Noah land surface model (Niu et al., 2011) and evaluated with in situ measurements from the International Soil Moisture Network (Dorigo et al., 2011). Values represent conditions of the first 10 cm of the top soil layer.

Gridded estimates of snow water equivalent for the Northern Hemisphere (NH) are provided by Mudryk (2020). The blended data set is derived from four sources, including a satellite passive microwave radiometer data set (GlobSnow GS3; Takala et al., 2011), reanalysis data (Modern-Era Retrospective analysis for Research and Applications, version 2 – MERRA-2; Gelaro et al., 2017), a temperature index model by Brown et al. (2003), and a physical snowpack model referred to as Crocus (Brun et al., 2013). The present study evaluates model output against the blended product, referred to as ECCC, as well as its four sources.

Global maps of monthly burned area were taken from the fourth generation of the Global Fire Emissions Database (GFED4; Giglio et al., 2010). The data are based on the 500 m Collection 5.1 MODIS direct broadcast (DB) burned-area product (MCD64A1) using the MODIS DB burned-area mapping algorithm by Giglio et al. (2009). The data set has been evaluated against high-resolution Landsat imagery for

locations in southern Africa, Siberia, central Asia, Alaska, and the western conterminous United States.

A second data set of monthly burned area is provided by the Fire Disturbance Project of the European Space Agency's Climate Change Initiative program (Chuvieco et al., 2018). This data set was generated from the MODIS red and near-infrared reflectances (MOD09GQ and MCD14ML, both from Collection 6) and thermal anomaly data (MCD14ML). The employed algorithm identifies those grid cells where the presence of wildfires is most evident and then searches for wildfires in close proximity to those grid cells.

CarbonTracker (CT2019) is an inversion model that attempts to reproduce observed atmospheric CO_2 concentrations by adjusting CO_2 fluxes at the surface. The initial states of terrestrial fluxes are obtained from the Carnegie–Ames–Stanford approach (CASA) biogeochemical model (Potter et al., 1993). Fluxes associated with biomass burning are taken from GFED4S. Observations consist of air samples from 460 sites provided by the GLOBALVIEW+ data product version 5.0 (Masarie et al., 2014). While CT2019 is continuously updated, the period considered in this study ranges from 2000 to 2017.

Globally gridded above-ground biomass estimates are provided by GEOCARBON, which is based on a boreal forest biomass map by Santoro et al. (2015) and two pan-tropical biomass maps from Avitabile et al. (2016) and Saatchi et al. (2011). The GEOCARBON map covers only areas that are dominated by trees in the Global Land Cover 2000 map (GLC2000; Bartholome and 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. Saatchi et al. (2011) upscaled in situ measurements using a machine learning algorithm (maximum entropy approach) and satellite imagery, including the MODIS normalized difference vegetation index (NDVI) and LAI products, global quick scatterometer (QSCAT), and SRTM digital elevation data. Note that we also used the Saatchi et al. (2011) data on their own to evaluate vegetation carbon in the tropics.

The Harmonized World Soil Database (HWSD) provided by the Food and 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 data were processed by Todd-Brown et al. (2013), who computed soil carbon stocks from bulk densities and organic carbon concentrations given in the HWSD for the top 1 m of soil at a $0.5° \times 0.5°$ resolution.

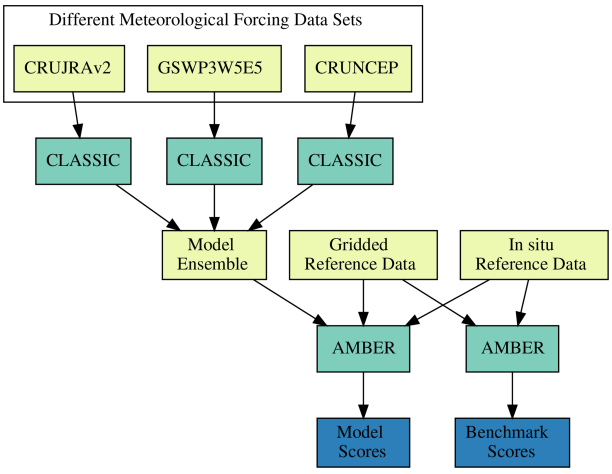

Figure 3. Experimental setup, where yellow boxes refer to data, green boxes to algorithms, and blue boxes to analysis output.

2.5 Automated Model Benchmarking R package (AMBER)

The Automated Model Benchmarking R package (AMBER; version 1.1.0) quantifies model performance using a skill score system originally developed by Collier et al. (2018). The method employs five scores that assess the model's bias (S_{bias}), root-mean-square error (S_{rmse}), seasonality (S_{phase}), interannual variability (S_{iav}), and spatial distribution (S_{dist}). 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 0 to 1, 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}).

Figure 3 provides an overview of our experimental setup. As a first step, we create a small model ensemble by driving CLASSIC with three different meteorological data sets. The purpose of the ensemble is to determine the sensitivity of model performance to uncertainties in the meteorological forcing. Next, we evaluate each ensemble member using AMBER. The resulting model scores express how well model and reference data agree. A low score, however, does not necessarily imply poor model performance, as the lack of agreement may also be due to uncertainties in the forcing data and/or reference data. To assess to what extent model scores are affected by observational uncertainties, we also run AMBER with reference data only. The resulting benchmark scores quantify the level of agreement among different reference data sets. Comparing model scores against benchmark scores then shows how well CLASSIC performs relative to the performance of a reference data set. The remaining

part of this section documents the details of the skill scores employed by AMBER.

2.5.1 Bias score (S_{bias})

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

$$\text{bias}(\lambda,\phi) = \overline{v_{\text{mod}}}(\lambda,\phi) - \overline{v_{\text{ref}}}(\lambda,\phi), \tag{1}$$

where $\overline{v_{\text{mod}}}(\lambda,\phi)$ and $\overline{v_{\text{mod}}}(\lambda,\phi)$ 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}):

$$\varepsilon_{\text{bias}}(\lambda,\phi) = |\text{bias}(\lambda,\phi)| \,/\, \sigma_{\text{ref}}(\lambda,\phi). \tag{2}$$

Note that $\varepsilon_{\text{bias}}$ is always positive, as it uses the absolute value of the bias. For evaluations against streamflow 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 snapshot in time. In both of these cases, $\varepsilon_{\text{bias}}(\lambda,\phi)$ becomes

$$\varepsilon_{\text{bias}}(\lambda,\phi) = |\text{bias}(\lambda,\phi)| \,/\, \overline{v_{\text{ref}}}(\lambda,\phi). \tag{3}$$

A bias score that scales from 0 to 1 is calculated next:

$$s_{\text{bias}}(\lambda,\phi) = e^{-\varepsilon_{\text{bias}}(\lambda,\phi)}. \tag{4}$$

While small relative errors yield score values close to 1, large relative errors cause score values to approach 0 (Fig. 4a). Taking the mean of s_{bias} across all latitudes and longitudes, denoted by a double bar over a variable, leads to the scalar score:

$$S_{\text{bias}} = \overline{\overline{s_{\text{bias}}}}(\lambda,\phi). \tag{5}$$

2.5.2 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:

$$\text{rmse}(\lambda,\phi) = \sqrt{\frac{1}{t_{\text{f}}-t_0}\int_{t_0}^{t_{\text{f}}} (v_{\text{mod}}(t,\lambda,\phi) - v_{\text{ref}}(t,\lambda,\phi))^2 \text{d}t}, \tag{6}$$

where t_0 and t_{f} are the initial and final time steps, respectively (Fig. A2a and c). A similar metric is the centralized rmse (crmse), which is based on the residuals of the anomalies:

$$\text{crmse}(\lambda,\phi) = \sqrt{\frac{1}{t_{\text{f}}-t_0}\int_{t_0}^{t_{\text{f}}} [(v_{\text{mod}}(t,\lambda,\phi) - \overline{v_{\text{mod}}}(\lambda,\phi)) - (v_{\text{ref}}(t,\lambda,\phi) - \overline{v_{\text{ref}}}(\lambda,\phi))]^2 \text{d}t}. \tag{7}$$

The crmse therefore assesses residuals that have been bias corrected (Fig. A2c and e). 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:

$$\varepsilon_{\text{rmse}}(\lambda,\phi) = \text{crmse}(\lambda,\phi) \,/\, \sigma_{\text{ref}}(\lambda,\phi), \tag{8}$$

scale this error onto a unit interval (Fig. 4a):

$$s_{\text{rmse}}(\lambda,\phi) = e^{-\varepsilon_{\text{rmse}}(\lambda,\phi)}, \tag{9}$$

and compute the spatial mean:

$$S_{\text{rmse}} = \overline{\overline{s_{\text{rmse}}}}. \tag{10}$$

2.5.3 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:

$$\theta(\lambda,\phi) = \max(c_{\text{mod}}(t,\lambda,\phi)) - \max(c_{\text{ref}}(t,\lambda,\phi)), \tag{11}$$

where c_{mod} and c_{ref} are the climatological mean cycle of the model and reference data, respectively (Fig. A2b). This time difference is then scaled from 0 to 1 based on the consideration that the maximum possible time difference is 6 months (Fig. 4b):

$$s_{\text{phase}}(\lambda,\phi) = \frac{1}{2}\left[1 + \cos\left(\frac{2\pi\theta(\lambda,\phi)}{365}\right)\right]. \tag{12}$$

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

$$S_{\text{phase}} = \overline{\overline{s_{\text{phase}}}}. \tag{13}$$

2.5.4 Interannual variability score (S_{iav})

The skill score S_{iav} quantifies how well the model reproduces patterns of interannual variability. This score is based on data where the seasonal cycle (c_{mod} and c_{ref}) has been removed (Fig. A2d and f):

$$\text{iav}_{\text{mod}}(\lambda,\phi) = \sqrt{\frac{1}{t_{\text{f}}-t_0}\int_{t_0}^{t_{\text{f}}} (v_{\text{mod}}(t,\lambda,\phi) - c_{\text{mod}}(t,\lambda,\phi))^2 \text{d}t}, \tag{14}$$

$$\text{iav}_{\text{ref}}(\lambda,\phi) = \sqrt{\frac{1}{t_{\text{f}}-t_0}\int_{t_0}^{t_{\text{f}}} (v_{\text{ref}}(t,\lambda,\phi) - c_{\text{ref}}(t,\lambda,\phi))^2 \text{d}t}.$$

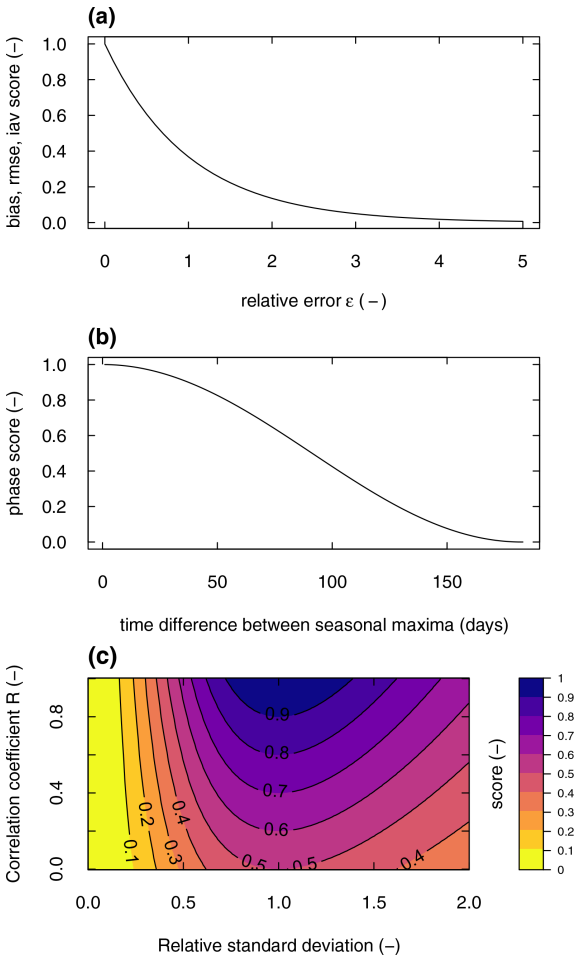

Figure 4. Score functions for **(a)** bias, root-mean-square error, and interannual variability, **(b)** seasonality, and **(c)** spatial distribution.

(15)

The relative error, nondimensionalization, and spatial mean are computed next (Fig. 4a):

$$\varepsilon_{\mathrm{iav}} = |(\mathrm{iav}_{\mathrm{mod}}(\lambda,\phi) - \mathrm{iav}_{\mathrm{ref}}(\lambda,\phi))| \big/ \mathrm{iav}_{\mathrm{ref}}(\lambda,\phi), \tag{16}$$

$$s_{\mathrm{iav}}(\lambda,\phi) = e^{-\varepsilon_{\mathrm{iav}}(\lambda,\phi)}, \tag{17}$$

$$S_{\mathrm{iav}} = \overline{\overline{s_{\mathrm{iav}}}}. \tag{18}$$

2.5.5 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 $\overline{v_{\mathrm{mod}}}(\lambda,\phi)$ and $\overline{v_{\mathrm{ref}}}(\lambda,\phi)$. The score S_{dist} increases from 0 to 1, the closer R and σ approach a value of 1 (Fig. 4c). No spatial integration is required, as this calculation yields a single value:

$$S_{\mathrm{dist}} = 2(1+R)\left(\sigma + \frac{1}{\sigma}\right)^{-2}, \tag{19}$$

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

$$\sigma = \sigma_{\overline{v_{\mathrm{mod}}}} \big/ \sigma_{\overline{v_{\mathrm{ref}}}}. \tag{20}$$

2.5.6 Overall score (S_{overall})

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

$$S_{\mathrm{overall}} = \frac{S_{\mathrm{bias}} + 2S_{\mathrm{rmse}} + S_{\mathrm{phase}} + S_{\mathrm{iav}} + S_{\mathrm{dist}}}{1+2+1+1+1}. \tag{21}$$

Note that S_{rmse} is weighted by a factor of 2, which emphasizes its importance.

3 Results

We start our evaluation with a qualitative overview of how the choice of meteorological forcing (CRUJRAv2, GSWP3W5E5, CRUNCEP) and reference data affect global mean biases. Our results show that for 10 out of 19 variables assessed in this study, the sign of the bias changes depending on what forcing and reference data are used (Table 2). Assuming that global values are reasonably accurate if (i) the sign of the bias varies depending on the choice of forcing and/or reference, or (ii) the global mean bias is reasonably small ($\leq 5\,\%$), we can divide our variables into four groups. The first group includes variables that show accurate global mean values for both gridded and in situ reference data. This includes the following variables: net surface radiation, sensible heat flux, snow water equivalent, and runoff (Table 2). The second group consists of variables that lack in situ observations but show accurate values when assessed against gridded reference data. This involves the following variables: net LW radiation, soil moisture, fractional area burned, net SW radiation, and vegetation carbon in the tropics. The third group includes variables that show accurate values for either globally gridded data (latent heat flux, GPP, NEE) or in situ reference data (ecosystem respiration, above-ground biomass, and LAI). For this group of variables, it will be important to assess the differences between evaluations based on gridded and in situ data in more detail. The fourth group includes variables with systematic biases across all evaluations, which includes surface albedo, emissions from wildfires, soil carbon, and soil heat flux. The following sections evaluate biases and other statistical metrics for each variable in detail.

3.1 Net surface radiation

RNS biases range from positive to negative values, depending on the choice of forcing (CRUJRAv2, GSWP3W5E5,

Table 2. Qualitative assessment of global mean biases. Variables for which positive and negative mean biases are found are denoted with ±. Biases that have a consistent sign across all evaluations but are smaller than ±5 % are marked with ≤ 5 %. Larger systematic biases are marked with + or −. Numbers provided for each of the in situ reference data refer to the number of grid cells involved.

Variable	Acronym	Grid	In situ	Gridded reference data	In situ reference data (n)
Net surface radiation	RNS	±	±	CERES, CLASSr, FluxCom, GEWEXSRB	FLUXNET(73)
Sensible heat flux	HFSS	±	±	CLASSr, FluxCom	FLUXNET (65)
Snow water equivalent	SNW	±	±	Brown, Crocus, GS3, MERRA-2	Mortimer (363)
Runoff	MRRO	≤ 5 %	±	CLASSr	GRDC (50 basins)
Net LW radiation	RLS	±		CERES, GEWEXSRB	
Soil moisture	MRSLL	±		ESA	
Fractional area burned	BURNT	±		ESACCI, GFED4S	
Net SW radiation	RSS	≤ 5 %		CERES, GEWEXSRB	
Vegetation carbon in the tropics	CVEG	≤ 5 %		Saatchi	
Latent heat flux	HFLS	±	−	CLASSr, FluxCom	FLUXNET (65)
Gross primary productivity	GPP	±	−	FluxCom, GOSIF, MODIS	FLUXNET (89)
Net ecosystem exchange	NEE	≤ 5 %	+	FluxCom, CT2019	FLUXNET (91)
Ecosystem respiration	RECO	+	≤ 5 %	FluxCom	FLUXNET (88)
Above-ground biomass	AGB	−	≤ 5 %	GEOCARBON	FOS (56)
Leaf area index	LAI	+	±	AVHRR, MODIS	CEOS (36), ORNL (330)
Surface albedo	ALBS	+		CERES, GEWEXSRB, MODIS	
Emissions from wildfires	FIRE	−		CT2019	
Soil carbon	CSOIL	−		HWSD	
Soil heat flux	HFG	−	−	CLASSr	FLUXNET (60)

CRUNCEP) and reference data (CERES, CLASSr, FluxCom, GEWEXSRB, and FLUXNET). For instance, global mean biases range from −15 % ($-12.07\,\mathrm{W\,m^{-2}}$) when forcing CLASSIC with CRUNCEP and evaluating results against CERES (2000–2012), to +19 % ($+12.24\,\mathrm{W\,m^{-2}}$) for CRUJRAv2 and CLASSr (2003–2009) (Fig. 5). Evaluating model output against in situ measurements by FLUXNET yields mean biases that range from −8 % ($-6.39\,\mathrm{W\,m^{-2}}$ for CRUNCEP) to +2 % ($+1.99\,\mathrm{W\,m^{-2}}$ for CRUJRAv2) (Figs. 5 and 1a).

Zonal mean values in net surface radiation are in close agreement with the corresponding reference values, with biases that range from −20 to $+20\,\mathrm{W\,m^{-2}}$ (Fig. 6). The same figure shows that net surface radiation biases in the NH extratropics are negative when evaluating CLASSIC against CERES, FluxCom, and GEWEXSRB, and positive when choosing CLASSr as a reference data set. Considering all three model runs and four gridded reference data sets shows that for most grid cells annual mean biases can be both negative and positive, depending on which model run is compared against which reference data (Fig. 7a). The strongest annual mean bias is located in the Tibetan Plateau with values below $-30\,\mathrm{W\,m^{-2}}$. Hovmöller diagrams show that CLASSIC captures the observed seasonality well across all latitudes (Fig. 8). Monthly biases range from negative to positive values across most latitudes and months, depending on the choice of meteorological forcing and reference data. This suggests that model values are reasonably accurate across space and time when considering the uncertainties associated with the forcing and all four globally gridded reference data.

Other statistical metrics confirm that CLASSIC reproduces net surface radiation reasonably well. The relative measure of the centralized root-mean-square error ($\varepsilon_{\mathrm{crmse}}$) ranges from 0.30 to 0.49 (unitless), which is lower than for most other variables (Fig. 5). Differences in the timing of modeled and observed seasonal peaks (θ) are less than 1 month. Modeled interannual variability is lower compared to CERES, CLASSr, GEWEXSRB, and FLUXNET, and larger than FluxCom (not shown). The relative interannual variability error ($\varepsilon_{\mathrm{iav}}$) is reasonably small (0.30 to 0.76, unitless). CLASSIC reproduces the spatial variability of annual mean values well, where the ratio of model and reference standard deviation in space (σ) is close to unity (1.00 to 1.19). Also, the spatial correlation coefficients (R) range from 0.86 to 0.97. Averaging the resulting scores across all three ensemble members shows that S_{overall} ranges from 0.74 to 0.80 (unitless), depending on the reference data (Fig. 9a). The choice of meteorological forcing affects scores by less than 0.1 (Fig. 9b). Scores tend to be higher when forcing CLASSIC with GSWP3W5E5 (Fig. 9c).

3.2 Net shortwave and longwave radiation

Results for net SW radiation (RSS) and LW radiation (RLS) are mainly consistent with findings for net surface radiation

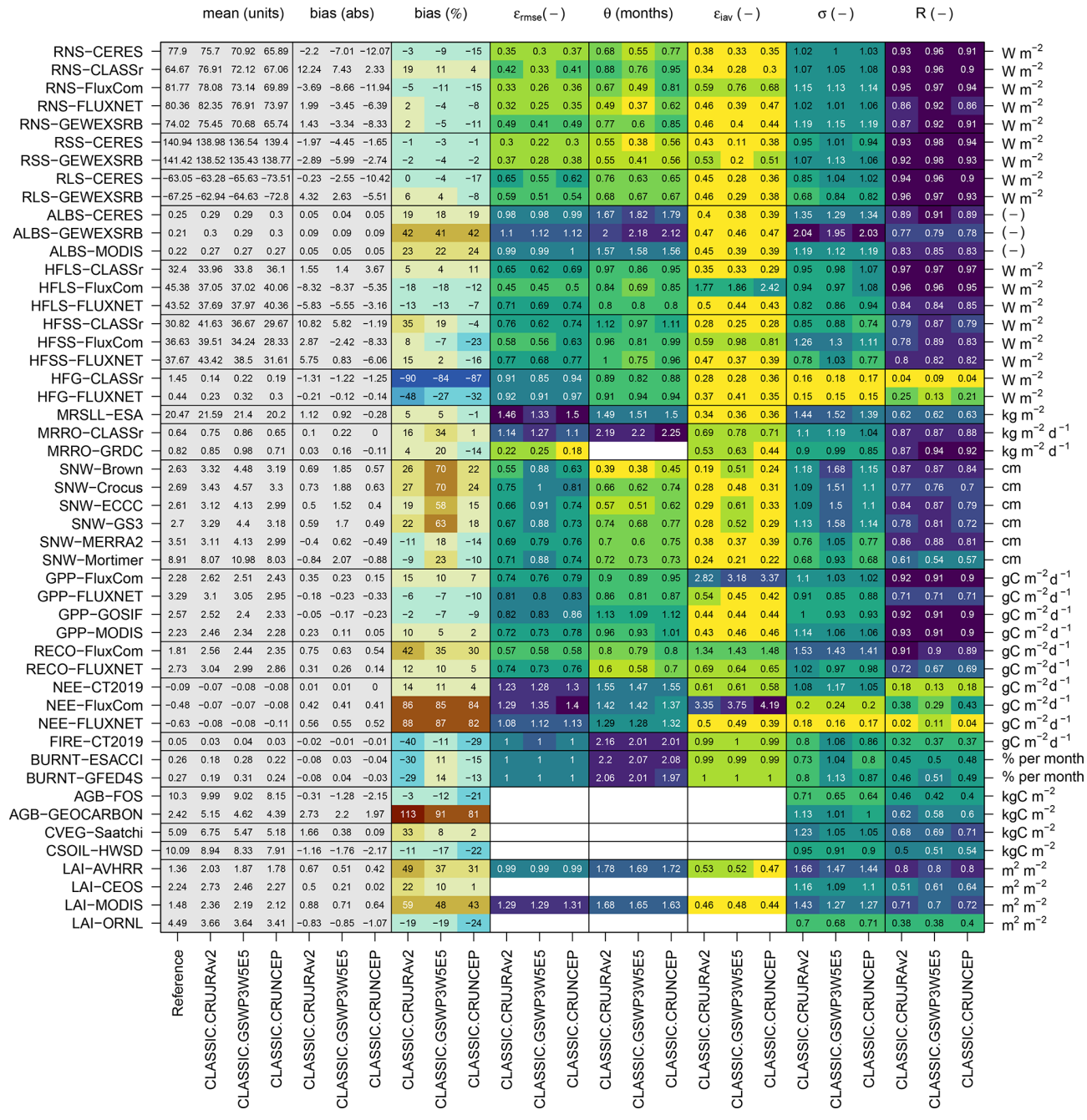

Figure 5. Global mean values of statistical metrics defined in Sect. 2. The color ranges span the minimum and maximum for each metric. The physical units of the mean and absolute bias are given by the vertical axis on the right-hand side.

described above. Net SW and LW radiation values produced by CLASSIC are reasonably accurate when considering uncertainties associated with meteorological forcing and reference data (CERES and GEWEXSRB). Global mean biases for net SW radiation range from −1 % (−1.65 W m^{-2}) to −4 % (−5.99 W m^{-2}) (Fig. 5). For net LW radiation, global mean biases range from −17 % (−10.42 W m^{-2}) to +6 % (+4.32 W m^{-2}). CLASSIC captures the zonal mean values and seasonal patterns of both variables well (Figs. 6, A4, and A5). However, CLASSIC also shows a systematic negative SW radiation bias in the NH extratropics from December to May (up to −40 W m^{-2}) (Fig. A4), which is consistent with a surface albedo bias as discussed further below. Results for other statistical metrics and corresponding scores are similar to the ones described for net surface radiation. For net SW radiation, ε_{crmse} range between 0.22 and 0.38 (unitless),

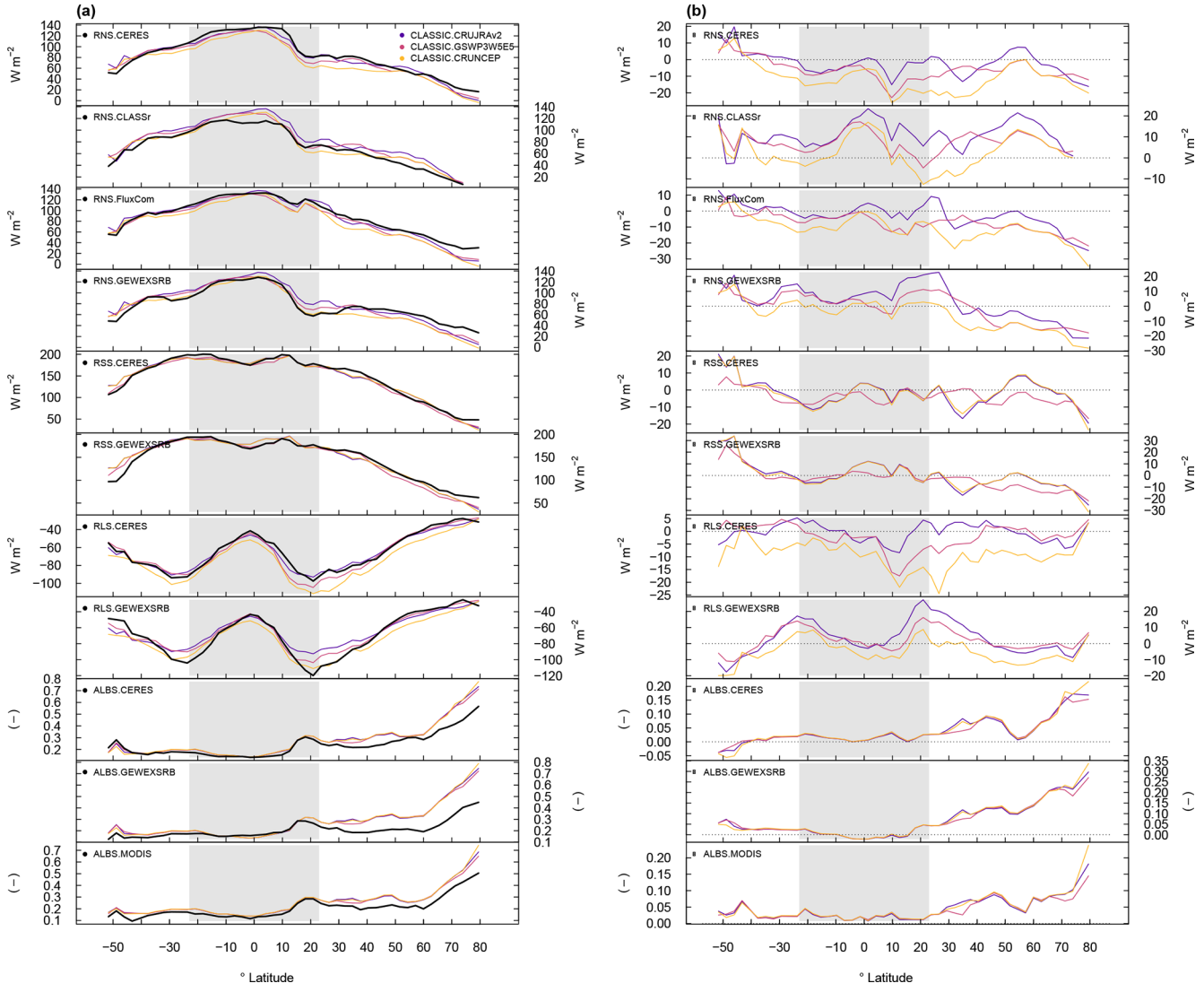

Figure 6. **(a)** Zonal mean values and **(b)** corresponding biases for variables related to the radiation cycle.

θ ranges from 0.38 to 0.56 months, ε_{iav} ranges from 0.11 to 0.53 (unitless), σ ranges from 0.94 to 1.13 (unitless), and R ranges from 0.92 to 0.98 (Fig. 5). For net LW radiation, values in $\varepsilon_{\text{crmse}}$ and θ are considerably higher than for net SW radiation, where $\varepsilon_{\text{crmse}}$ ranges from 0.51 to 0.65 (unitless) and θ ranges from 0.63 to 0.76 months. The overall scores are higher for net SW radiation ($S_{\text{overall}} \approx 0.80$) than for net LW radiation ($S_{\text{overall}} = 0.72$) (Fig. 9). The choice of the forcing has a considerable impact on how well the model captures the interannual variability of both net SW and LW radiation. The difference between the maximum and minimum interannual variability scores (S_{iav}) ranges from 0.10 to 0.21, depending on the reference data (Fig. 9b). For net LW radiation, the choice of the forcing has also a considerable impact on the bias score (ΔS_{bias}: 0.17–0.2). CLASSIC performs best when forcing the model with GSWP3W5E5 (Fig. 9c).

3.3 Albedo

Surface albedo (ALBS) shows a positive global mean bias across all three ensemble members and reference data sets (CERES, GEWEXSRB, and MODIS), ranging from +18 % to +42 % (+0.04 to +0.09, unitless) (Fig. 5). This bias occurs mainly in the NH extratropics, with increasing values towards higher latitudes (+0.2; Figs. 6 and 7b). The strongest annual mean bias is located in the Tibetan Plateau with values above 0.3. In the NH extratropics, the bias is most evident during the cold season (+0.4 between 60 and 80° N from December to January) (Fig. A6). Larger biases (+0.5) are present north of 80° N, but this involves only few grid cells. The timing of the bias suggests a possible relation with snow and/or the large solar zenith angle. The performance with respect to other statistical metrics tends to be lower than for net SW radiation. Values for $\varepsilon_{\text{crmse}}$ range between

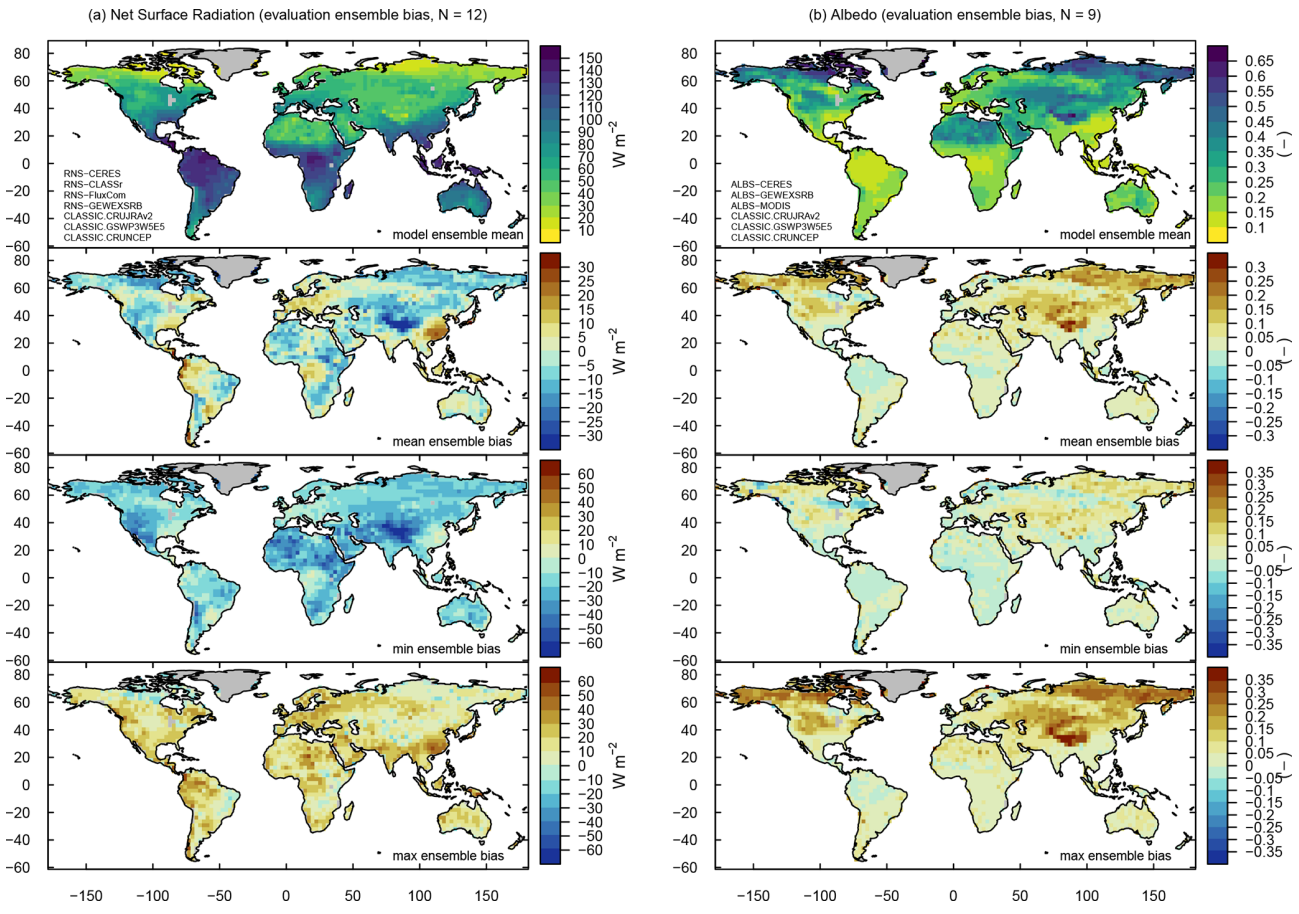

Figure 7. Model ensemble mean values and respective biases for **(a)** net surface radiation and **(b)** surface albedo. The mean, minimum, and maximum ensemble biases refer to the mean, minimum, and maximum values across all evaluations, respectively. For instance, a minimum bias that is negative implies that at least one evaluation yields a negative bias.

0.98 and 1.12 (unitless), θ ranges from 1.67 to 2.18 months, $\varepsilon_{\mathrm{iav}}$ ranges from 0.38 to 0.48 (unitless), σ ranges from 1.12 to 2.04 (unitless), and R ranges from 0.77 to 0.91 (Fig. 5). The corresponding overall score S_{overall} ranges between 0.48 and 0.59, depending on the reference data (Fig. 9). The S_{overall} values are not very sensitive to the choice of forcing ($\Delta S_{\mathrm{overall}} = 0.01$).

3.4 Heat fluxes

CLASSIC reproduces the partitioning of net surface radiation into sensible (HFSS), latent (HFLS), and soil heat flux (HFG) reasonably well. Global mean values in sensible heat flux simulated by CLASSIC are likely to be reasonable, with biases that range from negative to positive values, depending on the meteorological forcing ($-23\,\%$ to $+35\,\%$, -8.33 to $+10.82\,\mathrm{W\,m^{-2}}$) (Figs. 5 and 10). Driving CLASSIC with CRUNCEP causes biases to be negative for all three reference data sets, while the opposite is true when running the model with CRUJRAv2. For latent heat flux, the sign of the bias is consistent across all ensemble members, with negative biases when compared against FluxCom ($-12\,\%$ to $-18\,\%$, -5.35 to $-8.32\,\mathrm{W\,m^{-2}}$) and FLUXNET ($-7\,\%$ to $-13\,\%$, -3.16 to $-5.83\,\mathrm{W\,m^{-2}}$) and reasonably small positive biases when evaluated against CLASSr ($+5\,\%$ to $+11\,\%$, $+1.55$ to $+3.67\,\mathrm{W\,m^{-2}}$). Biases in soil heat flux are negative across all ensemble members and both reference data sets (CLASSr and FLUXNET) ($-32\,\%$ to $-90\,\%$). However, it must be noted that the corresponding absolute values are small (-0.14 to $-1.31\,\mathrm{W\,m^{-2}}$), implying that the impact of annual mean soil heat flux biases on the surface energy balance is minor. The relative partitioning of net surface radiation into sensible and latent heat fluxes is well reproduced, with reasonably small positive and negative biases, depending on the meteorological forcing (Fig. 10b). A more detailed assessment of each heat flux is provided next.

CLASSIC reproduces the zonal mean patterns of sensible and latent heat flux well, with zonal mean biases of about $\pm 20\,\mathrm{W\,m^{-2}}$ (Fig. 11). For sensible heat flux, zonal mean biases range from positive to negative values across all latitudes, depending on the meteorological forcing (Figs. 11 and A7a). For latent heat flux, biases are primarily positive

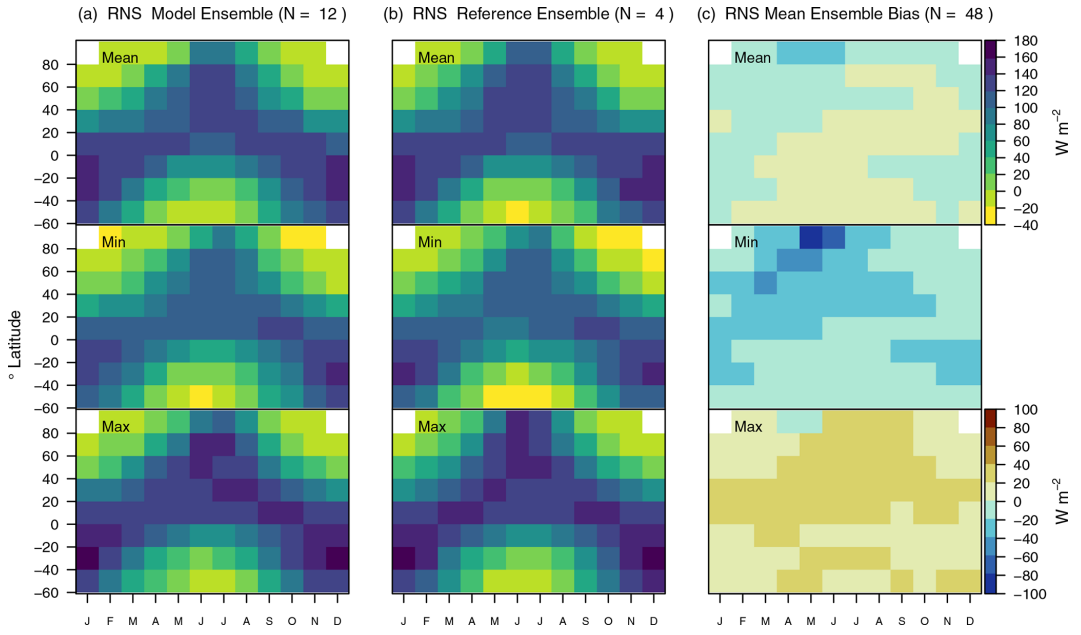

Figure 8. Net surface radiation computed from three model runs (CLASSIC driven with CRUJRAv2, GSWP3W5E5, and CRUNCEP) and four reference data sets (CERES, CLASSr, FluxCom, and GEWEXSRB). The top, center, and bottom rows provide ensemble mean, minimum, and maximum values for **(a)** CLASSIC, **(b)** reference data, and **(c)** corresponding biases. The mean, minimum, and maximum ensemble biases refer to the mean, minimum, and maximum values across all evaluations, respectively. For instance, a minimum bias that is negative implies that at least one evaluation yields a negative bias.

when compared against CLASSr and negative for FluxCom, regardless of the forcing (Figs. 11 and A7b). Zonal mean patterns of soil heat flux data from CLASSr are not well captured, with biases that become increasingly negative towards higher latitudes ($-4\,\mathrm{W\,m^{-2}}$ in the Arctic) (Figs. 11 and A8a).

CLASSIC captures the seasonal cycle of all three heat fluxes reasonably well (Figs. A9, A10, and A11). The negative latent heat flux biases are stronger during the warm season in both hemispheres. For soil heat fluxes, CLASSIC tends towards a more intense seasonal cycle in the extratropics, with positive biases in the warm and negative biases in the cold season.

Values of $\varepsilon_{\mathrm{rmse}}$ are similar among sensible and latent heat flux (0.45 to 0.77, unitless) and higher for soil heat flux (0.85 to 0.97) (Fig. 5). For all three heat fluxes, time lags in seasonal peaks are close to 1 month. The interannual variability of sensible and latent heat fluxes tends to be smaller compared to the CLASSr reference data set and larger compared to FluxCom (not shown), implying that model values lie within the uncertainty of the reference data. The corresponding $\varepsilon_{\mathrm{iav}}$ values range from 0.25 to 2.42 (unitless), where the higher values are obtained when evaluating CLASSIC against FluxCom (Fig. 5). For soil heat flux, modeled interannual variability is very similar to the corresponding reference values (not shown), leading to low values of $\varepsilon_{\mathrm{iav}}$ (0.28 to 0.36). Values of σ and R are reasonable for sensible heat flux (σ: 0.74 to 1.26, R: 0.78 to 0.89) and latent heat flux (σ: 0.94 to 1.08, R: 0.84 to 0.97), and poor for soil heat flux (σ: 0.15 to 0.17, R: 0.04 to 0.21). Consistently, the corresponding S_{overall} values are higher for latent and sensible heat flux (0.64 to 0.73) and lower for soil heat flux (0.53) (Fig. 9). From all three heat fluxes, sensible heat flux is most sensitive to the choice of forcing, with $\Delta S_{\mathrm{overall}}$ equal to 0.04 when driving the model with GSWP3W5E5.

3.5 Soil moisture

Global mean values in soil moisture (MRSLL) of the top 10 cm are very close to the corresponding reference values from ESA (1979–2017), with biases ranging from $-1\,\%$ ($-0.28\,\mathrm{kg\,m^{-2}}$, CRUNCEP) to $+5\,\%$ (about $+1\,\mathrm{kg\,m^{-2}}$, CRUJRAv2 and GSWP3W5E5) (Fig. 5). Zonal mean values are well reproduced with a tendency towards a positive bias north of 60° N ($+10\,\mathrm{kg\,m^{-2}}$) (Fig. 11). This bias is caused by clusters of grid cells located in different parts of the Arctic (Fig. A8b). The positive bias is present in all ensemble members and occurs from June to September (Fig. A12). Other statistical metrics confirm that the performance does not differ largely among ensemble members (Fig. 9). Values of $\varepsilon_{\mathrm{rmse}}$ and time lags of seasonal peaks (θ) are higher compared to most other variables ($\varepsilon_{\mathrm{rmse}} \leq 1.5$, $\theta \approx 1.5$ months). The spatial variability is larger than in the reference data ($\sigma \approx 1.44$, unitless) and correlation coefficients are relatively low ($R \approx 0.62$). The corresponding overall score is 0.52 and is not very sensitive to the choice of forcing ($\Delta S_{\mathrm{overall}} = 0.01$) (Fig. 9).

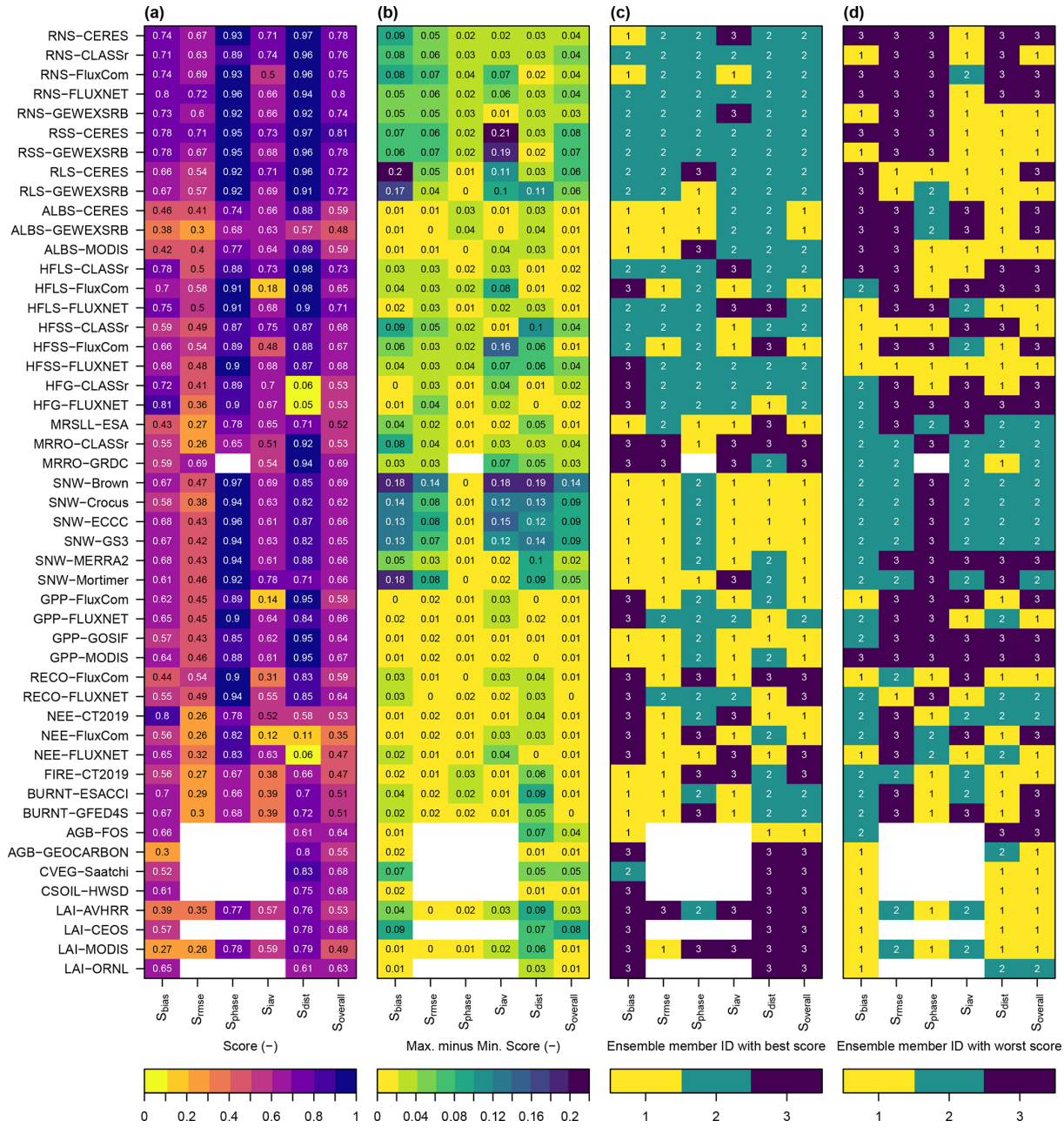

Figure 9. **(a)** Mean ensemble score, **(b)** maximum score difference among ensemble members, and ensemble member with the **(c)** highest and **(d)** lowest score. Ensemble member IDs refer to model runs conducted with data from (1) CRUJRAv2, (2) GSWP3W5E5, and (3) CRUNCEP.

3.6 Runoff and streamflow

Global mean runoff (MRRO) biases range from +1 % ($< 0.005\,\mathrm{kg\,m^{-2}\,d^{-1}}$, CRUNCEP) to +34 % ($0.22\,\mathrm{kg\,m^{-2}\,d^{-1}}$, GSWP3W5E5) when evaluated against CLASSr (2003–2009) (Fig. 5). This difference is consistent with the higher precipitation values found for the GSWP3W5E5 forcing. CLASSIC reproduces the zonal mean pattern of gridded reference runoff from CLASSr well, with mainly positive biases when forcing CLASSIC with GSWP3W5E5 (Fig. 11). Smaller, including negative, zonal mean biases occur when forcing CLASSIC with CRUNCEP. Consistently, the sign of the bias varies for many grid cells depending on the forcing (Fig. A13a). Furthermore, the sign of the bias varies for most latitudes and months (Fig. A14), suggesting that modeled runoff is within the uncertainty range of the meteorological forcing. CLASSIC reproduces the spatial patterns of runoff reasonably well, with σ values between 1.04 and 1.14 and R values between 0.87 and 0.88 (Fig. 5). The timing of the seasonal peaks, however, differs

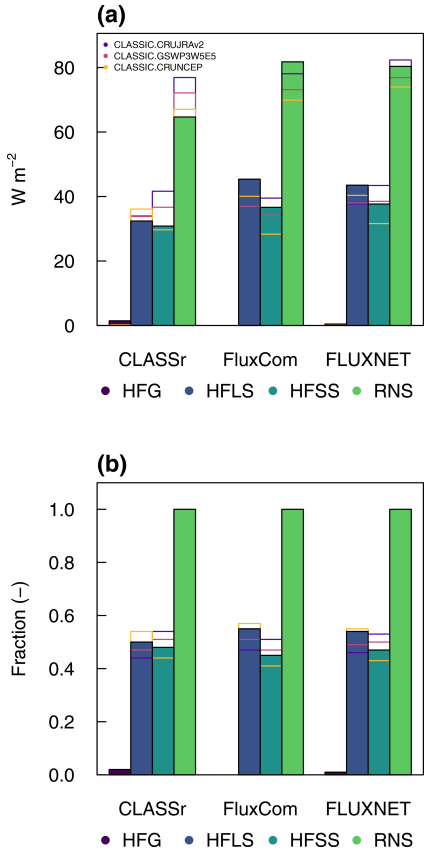

Figure 10. (a) Annual mean values of soil heat flux (HFG), latent heat flux (HFLS), sensible heat flux (HFSS), and net surface radiation (RNS) for the three CLASSIC ensemble members and for the reference data (CLASSr, FluxCom, and FLUXNET). Panel **(b)** shows the corresponding fractions with respect to net surface radiation.

by more than 2 months, and $\varepsilon_{\text{rmse}}$ values exceed 1 when using CLASSr as reference data. The overall score averaged over the three ensemble members is 0.53, with a better score when driving CLASSIC with CRUNCEP and worse when using GSWP3W5E5 as a forcing data set (Fig. 9).

Summing up the annual runoff from all grid cells located within each of the world's 50 largest river basins and comparing those values against annual mean streamflow at the corresponding river mouths yields biases that range between $-14\,\%$ ($-0.11\,\text{kg}\,\text{m}^{-2}\,\text{d}^{-1}$, CRUNCEP, GRDC) to $+20\,\%$ ($+0.16\,\text{kg}\,\text{m}^{-2}\,\text{d}^{-1}$, GSWP3W5E5, GRDC). For most basins, the sign of the bias remains constant regardless of the forcing (Fig. 1e for CRUJRAv2 as an example). CLASSIC reproduces the spatial patterns of streamflow reasonably well, with σ values close to 1 and R values ranging between 0.87 and 0.92 (Fig. 5). The overall score averaged over the three ensemble members is 0.69. Performance is better when driving CLASSIC with CRUNCEP and worse when using GSWP3W5E5 as a forcing data set (Fig. 9).

3.7 Snow water equivalent

Global mean SNW biases range from $-14\,\%$ ($-0.49\,\text{cm}$, CRUNCEP-driven run evaluated against MERRA-2) to $+70\,\%$ ($+1.88\,\text{cm}$, GSWP3W5E5-driven run evaluated against Crocus) for the 1981–2017 period (Fig. 5). Evaluating CLASSIC against in situ measurements compiled by Mortimer et al. (2020) shows that the mean bias can be negative ($-10\,\%$, $-0.88\,\text{cm}$, CRUNCEP) or positive ($+23\,\%$, $+2.07\,\text{cm}$, GSWP3W5E5), depending on the forcing (Figs. 5 and 1f). The higher snow water equivalent values in GSWP3W5E5-driven results are consistent with the higher precipitation values found for this forcing. For most evaluations, CLASSIC reproduces zonal mean values well when driven with CRUJRAv2 or CRUNCEP (Fig. 11). The sign of the bias ranges from positive to negative for most grid cells (Fig. A13b). Positive biases tend to be larger in eastern Canada, northern Scandinavia, and northern Siberia, and during NH spring (Fig. A15b).

Other statistical metrics confirm that CLASSIC's ability to reproduce snow water equivalent tends to be weaker when driven with GSWP3W5E5 ($\varepsilon_{\text{rmse}}$: 0.55 to 0.91, θ: 0.38 to 0.77, ε_{iav}: 0.19 to 0.61, σ: 0.68 to 1.68, and R: 0.54 to 0.87) (Fig. 5). The corresponding S_{overall} values range from 0.62 to 0.69, with lower scores when driving CLASSIC with GSWP3W5E5 and higher scores when using CRUJRAv2 (Fig. 9). Site-level-based score values are similar to grid-level score values, with an overall score of 0.66.

3.8 Gross primary productivity

Globally accumulated annual mean values of GPP computed from different reference data are of similar magnitude, with $115\,\text{Pg}\,\text{C}\,\text{yr}^{-1}$ for FluxCom (1980–2013), $121\,\text{Pg}\,\text{C}\,\text{yr}^{-1}$ for MODIS (2000–2016), and $130\,\text{Pg}\,\text{C}\,\text{yr}^{-1}$ for GOSIF (2000–2017) (Table 3). The corresponding model biases are reasonably small, with negative values when comparing CLASSIC against GOSIF ($-2\,\%$ to $-10\,\%$), and positive values when evaluating against MODIS ($<+1\,\%$ to $+10\,\%$) or FluxCom ($+4\,\%$ to $+14\,\%$). An assessment with eddy covariance data shows that biases range from $-6\,\%$ to $-10\,\%$ (Fig. 5). CLASSIC captures zonal mean GPP patterns well with no systematic bias pattern among evaluations (Fig. 12). Annual mean biases range from negative to positive for most grid cells, depending on the forcing and reference data (Fig. A16a). Hovmöller diagrams show that CLASSIC captures the seasonal GPP cycle well across all latitudes, with no systematic biases (Fig. A17). A more regional assessment, however, shows that the seasonal GPP cycle is out of phase in parts of the humid tropics of South America and Africa (not shown). Values in $\varepsilon_{\text{rmse}}$ are less than 1 for all evaluations, and time lags of seasonal peaks are about 1 month on average (Fig. 5). CLASSIC reproduces the interannual variability of GPP well when compared against FLUXNET, GOSIF, and MODIS ($\varepsilon_{\text{iav}} < 0.6$). Higher ε_{iav} values are found when

Table 3. Globally summed mean values and corresponding biases.

Variable	Ref. ID	Model ID	Ref.	Model	Bias	Bias (%)	Unit	Period
GPP	FluxCom	CLASSIC.CRUJRAv2	115.41	131.11	15.70	13.60	$\mathrm{Pg\,C\,yr^{-1}}$	1980–2013
GPP	FluxCom	CLASSIC.GSWP3W5E5	115.31	124.23	8.92	7.74	$\mathrm{Pg\,C\,yr^{-1}}$	1980–2013
GPP	FluxCom	CLASSIC.CRUNCEP	115.39	119.87	4.48	3.88	$\mathrm{Pg\,C\,yr^{-1}}$	1980–2013
GPP	GOSIF	CLASSIC.CRUJRAv2	130.01	127.84	−2.17	−1.67	$\mathrm{Pg\,C\,yr^{-1}}$	2000–2017
GPP	GOSIF	CLASSIC.GSWP3W5E5	129.61	120.02	−9.59	−7.40	$\mathrm{Pg\,C\,yr^{-1}}$	2000–2016
GPP	GOSIF	CLASSIC.CRUNCEP	129.73	116.60	−13.13	−10.12	$\mathrm{Pg\,C\,yr^{-1}}$	2000–2016
GPP	MODIS	CLASSIC.CRUJRAv2	121.47	133.07	11.60	9.55	$\mathrm{Pg\,C\,yr^{-1}}$	2000–2016
GPP	MODIS	CLASSIC.GSWP3W5E5	121.33	125.29	3.96	3.26	$\mathrm{Pg\,C\,yr^{-1}}$	2000–2016
GPP	MODIS	CLASSIC.CRUNCEP	121.46	121.74	0.28	0.23	$\mathrm{Pg\,C\,yr^{-1}}$	2000–2016
NEE	CT2019	CLASSIC.CRUJRAv2	−4.15	−4.19	−0.04	−0.96	$\mathrm{Pg\,C\,yr^{-1}}$	2000–2017
NEE	CT2019	CLASSIC.GSWP3W5E5	−4.18	−4.36	−0.18	−4.31	$\mathrm{Pg\,C\,yr^{-1}}$	2000–2016
NEE	CT2019	CLASSIC.CRUNCEP	−4.19	−4.60	−0.41	−9.79	$\mathrm{Pg\,C\,yr^{-1}}$	2000–2016
NEE	FluxCom	CLASSIC.CRUJRAv2	-24.88	-3.54	21.34	85.77	$\mathrm{Pg\,C\,yr^{-1}}$	1980–2013
NEE	FluxCom	CLASSIC.GSWP3W5E5	−24.85	−3.87	20.98	84.43	$\mathrm{Pg\,C\,yr^{-1}}$	1980–2013
NEE	FluxCom	CLASSIC.CRUNCEP	−24.88	−4.13	20.75	83.40	$\mathrm{Pg\,C\,yr^{-1}}$	1980–2013
FIRE	CT2019	CLASSIC.CRUJRAv2	2.88	1.82	−1.06	−36.81	$\mathrm{Pg\,C\,yr^{-1}}$	2000–2017
FIRE	CT2019	CLASSIC.GSWP3W5E5	2.90	2.70	−0.20	−6.90	$\mathrm{Pg\,C\,yr^{-1}}$	2000–2016
FIRE	CT2019	CLASSIC.CRUNCEP	2.90	2.16	−0.74	−25.52	$\mathrm{Pg\,C\,yr^{-1}}$	2000–2016
BURNT	ESACCI	CLASSIC.CRUJRAv2	4.61	3.31	−1.30	−28.20	$10^6\ \mathrm{km^2\,yr^{-1}}$	2001–2017
BURNT	ESACCI	CLASSIC.GSWP3W5E5	4.62	5.27	0.65	14.07	$10^6\ \mathrm{km^2\,yr^{-1}}$	2001–2016
BURNT	ESACCI	CLASSIC.CRUNCEP	4.62	4.05	−0.57	−12.34	$10^6\ \mathrm{km^2\,yr^{-1}}$	2001–2016
BURNT	GFED4S	CLASSIC.CRUJRAv2	4.56	3.32	−1.24	−27.19	$10^6\ \mathrm{km^2\,yr^{-1}}$	2001–2015
BURNT	GFED4S	CLASSIC.GSWP3W5E5	4.56	5.29	0.73	16.01	$10^6\ \mathrm{km^2\,yr^{-1}}$	2001–2015
BURNT	GFED4S	CLASSIC.CRUNCEP	4.56	4.06	−0.50	−10.96	$10^6\ \mathrm{km^2\,yr^{-1}}$	2001–2015
AGB	GEOCARBON	CLASSIC.CRUJRAv2	214.53	442.25	227.72	106.15	Pg C	1980–2017
AGB	GEOCARBON	CLASSIC.GSWP3W5E5	214.50	384.31	169.81	79.17	Pg C	1980–2017
AGB	GEOCARBON	CLASSIC.CRUNCEP	214.53	364.94	150.41	70.11	Pg C	1980–2017
CSOIL	HWSD	CLASSIC.CRUJRAv2	1199.32	1108.79	−90.53	−7.55	Pg C	1980–2017
CSOIL	HWSD	CLASSIC.GSWP3W5E5	1198.49	1023.48	−175.01	−14.60	Pg C	1980–2017
CSOIL	HWSD	CLASSIC.CRUNCEP	1198.26	955.57	−242.69	−20.25	Pg C	1980–2017

evaluating the model against FluxCom. However, it must be noted that interannual variability patterns from FluxCom are not reliable (Jung et al., 2020). Spatial patterns of GPP are well captured, with values of σ close to 1 and correlation coefficients around 0.91. The corresponding overall scores range from 0.58 (FluxCom) to 0.67 (MODIS), with higher scores when forcing CLASSIC with CRUJRAv2 (Fig. 9).

3.9 Ecosystem respiration

Annual mean biases of RECO are reasonably small across FLUXNET sites, with mean biases ranging from +5 % to +12 %, depending on the forcing (Fig. 5). Looking at the spatial distribution of FLUXNET sites suggests that the positive biases are more evident in the NH extratropics rather than in the tropics (Fig. 2b). Larger biases are present when comparing CLASSIC with FluxCom; however, those values are not reliable as FluxCom underestimates ecosystem respiration in the tropics (Jung et al., 2020). Other statistical metrics derived from FLUXNET data suggest that CLASSIC reproduces ecosystem respiration reasonably well ($\varepsilon_{\mathrm{rmse}} \approx 0.75$, $\theta < 1$, $\varepsilon_{\mathrm{iav}} \approx 0.66$, $\sigma \approx 1.0$, $R \approx 0.69$) (Fig. 5). The corresponding overall score is 0.64 (Fig. 9). This score value is insensitive to the choice of forcing.

3.10 Net ecosystem exchange

Globally accumulated annual mean NEE computed from CT2019 is $-4.2\,\mathrm{Pg\,C\,yr^{-1}}$ for the period 2000–2016 (Table 3). The corresponding value from FluxCom is $-24.9\,\mathrm{Pg\,C\,yr^{-1}}$ for the period 1980–2013, which is unrealistic and due to an underestimation of ecosystem respiration in the tropics (Jung et al., 2020). CLASSIC's estimate of annual mean NEE ranges from -4.2 to $-4.6\,\mathrm{Pg\,C\,yr^{-1}}$ (2000–2016), which corresponds to a bias of about −1 % to −10 % when compared against CT2019, depending on the meteorological forcing. However, much larger biases apply when evaluating CLASSIC against FLUXNET (+88 %; Figs. 2c and 5).

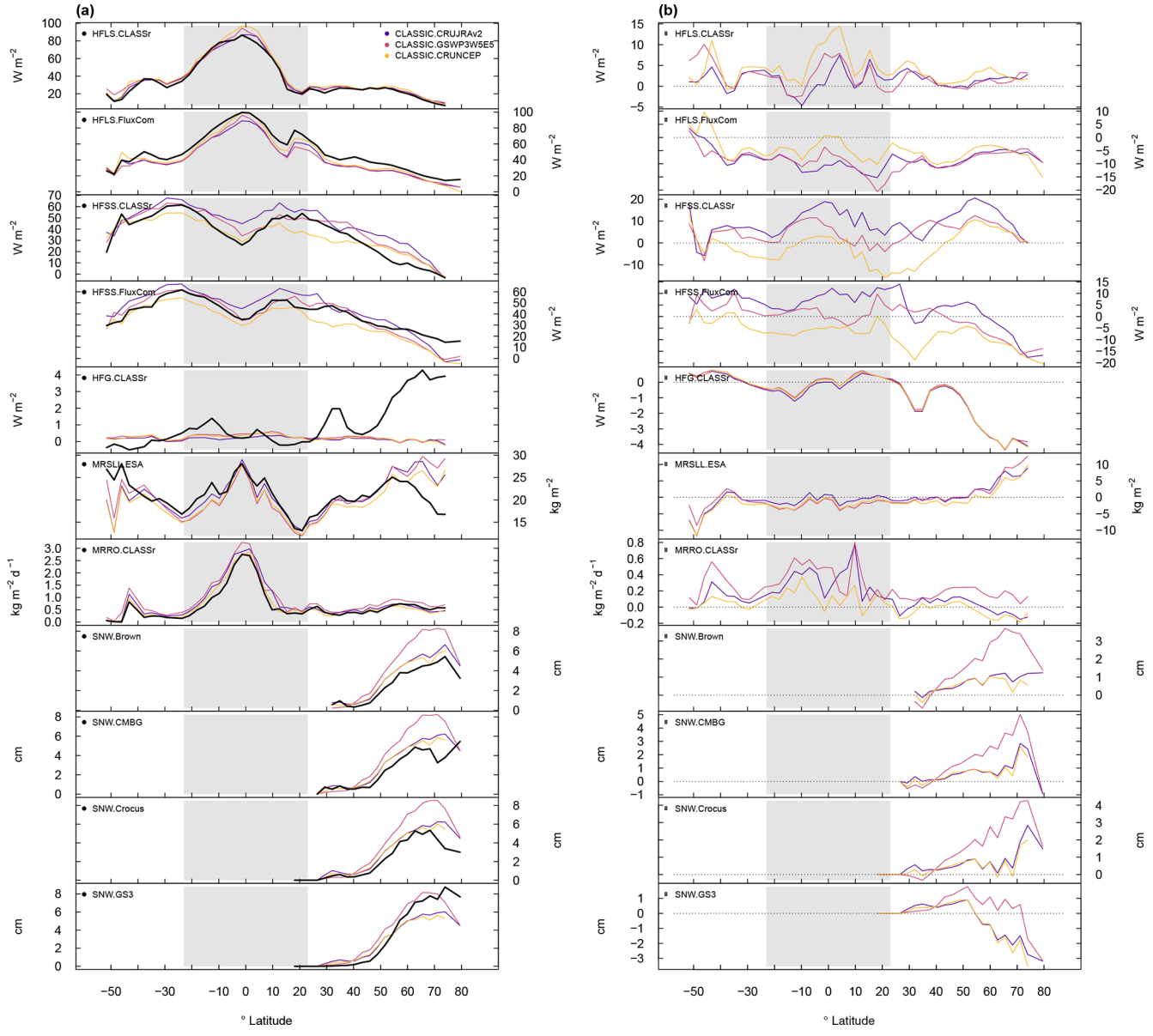

Figure 11. **(a)** Zonal mean values and **(b)** corresponding biases for heat fluxes and variables related to the water cycle.

Zonal mean NEE patterns resemble values from CT2019 to a certain degree (Fig. 12). The strength of the carbon sink is modeled to be stronger in the tropics ($-0.15\,\mathrm{g\,C\,m^{-2}\,d^{-1}}$) and weaker in the extratropics ($+0.10\,\mathrm{g\,C\,m^{-2}\,d^{-1}}$) compared to CT2019. FLUXNET data confirm the tendency towards a positive bias, i.e. weaker sink, in the NH extratropics (Fig. 2c). In the tropics, however, the sign of the bias is mixed, and the number of sites is too small to confirm whether or not CLASSIC overestimates the strength of the carbon sink at low latitudes.

Hovmöller diagrams suggest that CLASSIC reproduces the general pattern of the seasonal NEE cycle in the extratropics, where the strength of the carbon sink is strongest during the warm period (Fig. A18). Other statistical metrics suggest a considerable mismatch between model and reference NEE ($\varepsilon_{\mathrm{rmse}} > 1$, $\theta > 1$, and $R < 0.2$). The corresponding overall scores are 0.47 for FLUXNET and 0.53 for CT2019 (Fig. 9).

3.11 Area burned and resulting emissions

Wildfires are estimated to burn an area of about $4.6\times 10^6\,\mathrm{km^2\,yr^{-1}}$ (ESACCI from 2001–2017 and GFED4S from 2001–2015), causing emissions of about $2.9\,\mathrm{Pg\,C\,yr^{-1}}$ (CT2019, 2000–2017) (Table 3, where BURNT refers to area burned and FIRE refers to the associated emissions). CLASSIC reproduces the spatial extent of wildfires well con-

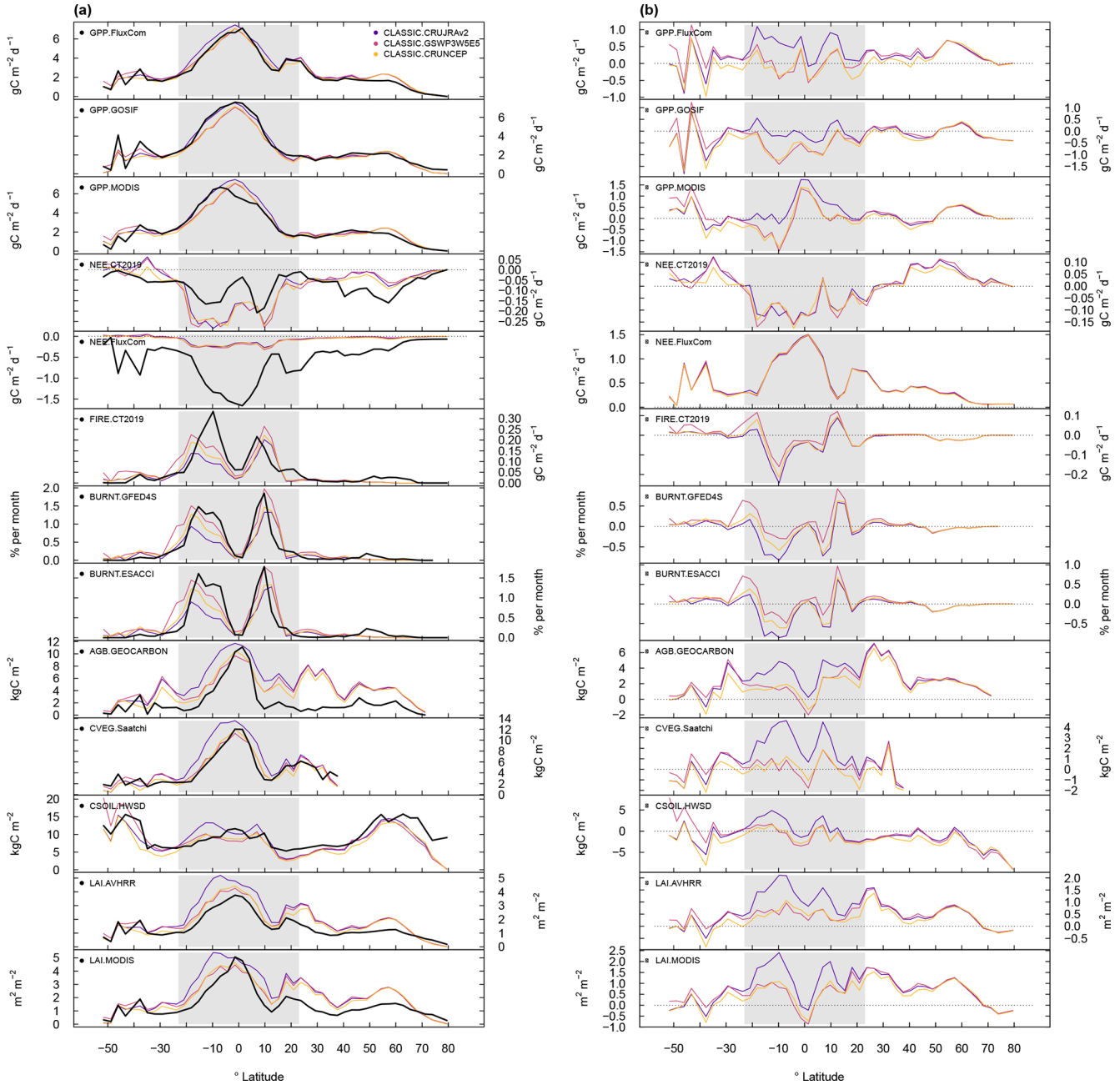

Figure 12. **(a)** Zonal mean values and **(b)** corresponding biases for variables related to the carbon cycle.

sidering the uncertainty of the meteorological forcing, with biases ranging from −28 % when forcing CLASSIC with CRUJRAv2, to +14 % for GSWP3W5E5. The higher fractional area burned values in GSWP3W5E5 are not related to precipitation, as GSWP3W5E5 is the wettest of the three forcing data sets across all latitudes and seasons. Instead, GSWP3W5E5 has larger values in near-surface wind speeds compared to the other forcing data sets, which increase the probability of wildfires. Model values for emissions from wildfires tend to be lower than observed, with biases ranging from −37 % (CRUJRAv2) to −7 % (GSWP3W5E5).

CLASSIC reproduces the zonal mean pattern of both variables well, with values that are highest at 10° S and 10° N (Fig. 12). The sign of the zonal mean bias is consistent between both variables. Biases tend to be negative in the humid parts of the tropics and positive in the drier parts of the tropics (Figs. 12 and A19). CLASSIC reproduces the seasonality of both variables well, with lower values when the solar zenith angle in the tropics is small (Figs. A20 and A21),

which is indicative of the position of the intertropical convergence zone. While biases are most evident in the tropics, the model also underestimates fractional area burned and associated emissions in the boreal forests. Other statistical metrics suggest that considerable differences between model and reference values remain ($\varepsilon_{\text{rmse}} = 1.0$, $\theta > 2$ months, $\varepsilon_{\text{iav}} \approx 1$, and $R \leq 0.51$) (Fig. 5). The resulting overall scores for fractional area burned and emissions from fires are 0.51 and 0.47, respectively (Fig. 9). These score values are not very sensitive to the choice of forcing ($\Delta S_{\text{overall}} \leq 0.01$).

3.12 Vegetation biomass

Reference values from GEOCARBON estimate globally accumulated AGB with 214.5 Pg C. The corresponding values from CLASSIC range from 364.9 Pg C (+70 %) to 442.3 Pg C (+106 %) (Table 3). The bias is largest between 20° N and 40° N, in particular in the eastern US and in East Asia (Figs. 12 and A22a). Comparing CLASSIC against vegetation carbon (i.e. above- and below-ground living biomass; CVEG) for the tropics provided by Saatchi et al. (2011) suggests smaller biases (+2 % for CRUNCEP to +33 % for CRUJRAv2) (Figs. 5, 12, and A22b). Spatial correlations are weak, with $R \leq 0.62$ for above-ground biomass and $R \leq 0.71$ for vegetation carbon (Fig. 5). The corresponding overall scores are 0.55 ($\Delta S_{\text{overall}} = 0.01$) for above-ground biomass and 0.68 ($\Delta S_{\text{overall}} = 0.05$) for vegetation carbon in the tropics (Fig. 9).

To investigate differences between CLASSIC and GEOCARBON in greater detail, we evaluate model and reference above-ground biomass values against forest inventory data (FOS). The comparison shows a larger negative bias for GEOCARBON (−50 %) and a smaller negative bias for CLASSIC (−3 % for CRUJRAv2 to −21 % for CRUNCEP) (Figs. 5 and 2d). This finding suggests that GEOCARBON may be subject to considerable uncertainties, which may contribute to the large bias when evaluating CLASSIC against GEOCARBON. More forest inventory data are required to reach a conclusive result, as the current spatial coverage of inventory plots provided by FOS is modest. Evaluating CLASSIC against FOS yields an overall score of 0.64 (Fig. 9).

3.13 Soil carbon

The global stock of soil carbon (CSOIL) is estimated with 1198 Pg C (HWSD). The corresponding values from CLASSIC range from about 956 Pg C (−20 %; CRUNCEP) to 1109 Pg C (−8 %; CRUJRAv2) (Table 3). CLASSIC reproduces the zonal mean patterns of soil carbon well, with a tendency for negative biases in the Arctic ($-10\,\text{kg C m}^{-2}$) (Fig. 12). Some of this bias is related to a small number of grid cells at high latitudes with particularly large biases ($-40\,\text{kg C m}^{-2}$) (Fig. A23a). The higher values in the reference data may be related to the presence of permafrost and peatland, which are currently not represented in CLASSIC. The spatial variability of soil carbon is well reproduced ($\sigma \approx 0.92$), but the spatial correlation is rather weak ($R \approx 0.5$) (Fig. 5). The corresponding overall score is 0.68 ($\Delta S_{\text{overall}} = 0.01$) (Fig. 9).

3.14 Leaf area index

Evaluating CLASSIC with gridded LAI data from AVHRR (1982–2010) and MODIS (2000–2009) suggests large positive global mean biases that range from +31 % ($+0.42\,\text{m}^{-2}\,\text{m}^{-2}$) to +59 % ($+0.88\,\text{m}^{-2}\,\text{m}^{-2}$), depending on the forcing and reference data (Fig. 5). While CLASSIC reproduces the general pattern of zonal mean LAI reasonably well, the model has a positive bias across most latitudes (Fig. 12). However, biases have a local minimum at the Equator, especially when using MODIS as a reference. All six evaluations tend towards positive biases in forested regions of the NH extratropics and the drier parts of the tropics (Fig. A23b). In the more humid tropics of the Amazon and Congo basin, biases range from negative to positive values, depending on the forcing and reference data. Hovmöller diagrams show that the positive bias in the NH extratropics occurs mainly during the cold season, suggesting a possible relation with snow, leaf shedding, and/or the larger solar zenith angle (Fig. A24). For most of the remaining months and latitudes, LAI biases range from positive to negative biases, suggesting that modeled values are reasonably accurate when considering the uncertainty of the meteorological forcing and reference data.

Other statistical metrics suggest further differences between model and reference LAI. Values of $\varepsilon_{\text{rmse}}$ are higher than for most other variables (0.99 to 1.31), and time lags in seasonal peaks range between 1 and 2 months (Fig. 5). The spatial variability of LAI is considerably larger for model data than gridded reference data (σ ranges from 1.27 to 1.66), and the correlation coefficients are reasonable but not impressive (0.7 to 0.8). The corresponding overall scores are 0.49 ($\Delta S_{\text{overall}} = 0.01$) and 0.53 ($\Delta S_{\text{overall}} = 0.03$) when evaluating CLASSIC against gridded LAI reference data from MODIS and AVHRR, respectively (Fig. 9).

To investigate the apparent LAI bias further, we evaluate model and reference LAI values with in situ measurements of maximum LAI provided by ORNL. The comparisons show larger negative biases for MODIS (−37 %) and AVHRR (−33 %), and smaller negative biases for CLASSIC (−19 % to −24 %) (Figs. 5 and 2f for CLASSIC only). This suggests that (i) it is unlikely that CLASSIC overestimates maximum LAI, and (ii) the large positive LAI bias discussed above may in part be due to an underestimation of LAI by the gridded reference data set. Furthermore, evaluating CLASSIC against in situ LAI observations from CEOS shows mean biases that range from +1 % to +22 % (Figs. 5 and 2e). The corresponding values for MODIS and AVHRR are −6 % and −4 %, respectively (not shown). Given these findings, we

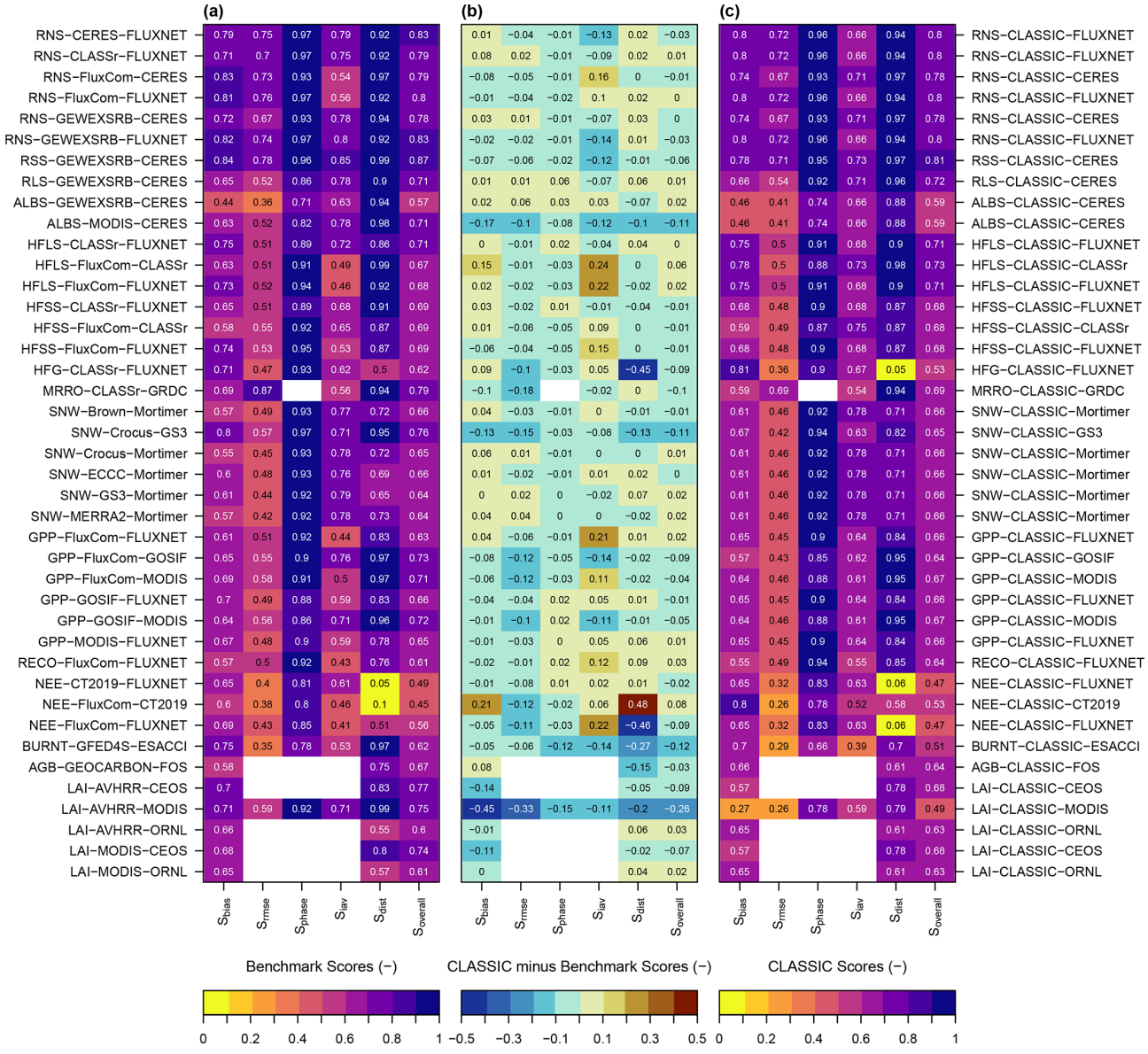

Figure 13. **(a)** Benchmark scores computed from reference data only, **(c)** mean model ensemble skill scores, and **(b)** corresponding score difference.

conclude that the large positive LAI found for CLASSIC when evaluated against AVHRR and MODIS must be interpreted with caution. One possible reason for lower LAI in satellite-based reference data at high latitudes is the problems associated with LAI retrievals during winter months when the sun is low and the surface is snow covered.

3.15 Benchmark scores derived from reference data only

Benchmark scores are produced by comparing two independent reference data sets (Fig. 3). This exercise fulfills three purposes. First, benchmark scores quantify observational uncertainty, where lower values imply larger uncertainties. Second, we can evaluate the quality of gridded products using in situ measurements. Third, comparing model scores against benchmark scores shows how well CLASSIC performs relative to the performance of a reference data set. The comparison of two reference data sets is limited to the time period for which both data sets overlap.

Our results demonstrate that most benchmarks are considerably low, suggesting that observational uncertainties are large. For most variables, model skill scores are very similar to benchmark skill scores, implying that CLASSIC performs similarly to a reference data set. In some cases, CLASSIC even outperforms gridded reference data sets when evaluating gridded reference data and model output against in situ

measurements. The remaining section discusses these findings in more detail.

Contrary to our expectations, most benchmark scores are far below unity (Fig. 13). Only two variables show S_{overall} values equal to or greater than 0.8 (net surface radiation and net SW radiation). The largest uncertainties are found for NEE, with $S_{\text{overall}} = 0.49$ when comparing CT2019 against FLUXNET. Even lower score values are found when evaluating FluxCom NEE against FLUXNET due to the known limitations of the FluxCom data set discussed in Sect. 2.4.2.

Comparing gridded reference data against in situ measurements shows subtle differences in how well gridded products perform for the same variable. For net surface radiation, S_{overall} values range from 0.79 for CLASSr to 0.83 for CERES when evaluated against FLUXNET. For latent heat flux, CLASSr ($S_{\text{overall}} = 0.71$) performs slightly better than FluxCom ($S_{\text{overall}} = 0.68$). For sensible heat flux, CLASSr and FluxCom perform equally well ($S_{\text{overall}} = 0.69$). For snow water equivalent, the blended multi-reference mean data (ECCC; $S_{\text{overall}} = 0.66$) perform very similarly to the individual reference data (S_{overall}: 0.64 to 0.66) when evaluated against in situ measurements provided by Mortimer et al. (2020). For GPP, the S_{overall} values are highest for GOSIF (0.66), followed by MODIS (0.65) and FluxCom (0.63) when evaluated against FLUXNET. In the case of NEE, S_{overall} values are higher for FluxCom (0.56) than for CT2019 (0.49). However, it must be noted that the majority of FLUXNET sites are located in the NH extratropics, implying that corresponding scores reflect the performance in primarily non-tropical regions. For LAI, the S_{overall} values are 0.61 for MODIS and 0.60 for AVHRR when evaluated against in situ measurements from ORNL and 0.74 (MODIS) and 0.77 (AVHRR) when compared against data from CEOS.

In most cases, model skill score values are very similar to the corresponding benchmark score values. For instance, evaluating CERES net surface radiation against FLUXNET yields $S_{\text{overall}} = 0.83$. The corresponding mean score value from all three CLASSIC ensemble members is 0.80. For a number of cases, the CLASSIC ensemble performs just as well as independent reference data sets. For instance, S_{overall} values for LW radiation are both 0.8 when evaluating either GEWEXSRB against CERES or the CLASSIC ensemble against CERES. In three cases (fractional area burned, LAI, and snow water equivalent), model scores and benchmark scores differ by more than 0.1, which is discussed next.

The S_{overall} benchmark score for fractional area burned is 0.62 (GFED4S versus ESACCI) and the corresponding model skill score is 0.51 (CLASSIC ensemble versus ESACCI). The score difference is driven by deficiencies in reproducing spatial patterns ($\Delta S_{\text{dist}} = -0.27$). For LAI, the S_{overall} benchmark score is 0.75 (AVHRR versus MODIS) and the S_{overall} model score is 0.49 (CLASSIC ensemble versus MODIS). This difference is primarily driven by lower model skill scores in S_{bias} and S_{rmse}. However, in Sect. 3.14, we conclude that LAI biases derived from gridded reference data must be interpreted with caution. Comparing LAI benchmark scores derived from in situ measurements (AVHRR versus ORNL, $S_{\text{overall}} = 0.60$) shows much greater similarity with the corresponding model skill scores (CLASSIC ensemble versus ORNL, $S_{\text{overall}} = 0.63$). Similar results apply when conducting the same analysis with LAI from MODIS rather than AVHRR.

For snow water equivalent, the S_{overall} benchmark skill score exceeds the corresponding model skill score by 0.11 when evaluating Crocus and the CLASSIC ensemble against GS3. However, using in situ measurements yields benchmark and model skill scores that are more similar. For instance, evaluating the snow water equivalent reference ensemble ECCC and the CLASSIC ensemble against data from Mortimer et al. (2020) yields identical S_{overall} values (0.66).

For some variables, CLASSIC even outperforms gridded reference data when evaluated against in situ observations. This applies to net surface radiation from CLASSr, latent heat flux from FluxCom, snow water equivalent from Crocus, GS3, and MERRA-2, GPP from FluxCom and MODIS, and LAI from MODIS and AVHRR.

To rank model performance by variable, we compute the average model-benchmark score difference across all evaluations for each variable. Following this approach, model performance increases as follows, where the average model-benchmark score differences are provided in parenthesis: fractional area burned (-0.12), runoff (-0.1), soil heat flux (-0.09), LAI (-0.07), net SW radiation (-0.06), NEE (-0.05), above-ground biomass (-0.03), GPP (-0.03), surface albedo (-0.02), snow water equivalent (-0.01), net surface radiation (-0.01), sensible heat flux (-0.01), net LW radiation (0.01), latent heat flux (0.03), and ecosystem respiration (0.03). This ranking excludes the FluxCom NEE evaluation against CT2019 due to the unrealistic FluxCom NEE values in the tropics. To conclude, comparing model skill scores against benchmark scores derived from reference data suggests that model performance is reasonably good when considering the uncertainties associated with the reference data.

4 Discussion

We begin our discussion by comparing our findings with the skill scores presented for CLASS-CTEM by Friedlingstein et al. (2019). While our results are largely consistent, we can now provide more insights that facilitate the interpretation of those scores. Our findings show that the poor performance of LAI is mainly due to a large positive bias that occurs in the NH extratropics during the cold season. An evaluation against in situ LAI measurements shows that biases can be positive, negative, or close to zero, depending on the respective forcing and in situ reference data set. Furthermore, CLASSIC performs very similar to MODIS and AVHRR when comparing all three data sets against in

situ LAI measurements. Also, evaluating MODIS against AVHRR yields an overall score of 0.75, which implies a considerable mismatch between both reference data sets. Given the uncertainties in remotely sensed LAI retrievals discussed in Sect. 2.4.2, it is very possible that the mismatch between modeled and gridded reference LAI is, at least to some extent, due to deficiencies in remotely sensed LAI.

The NEE scores presented in Friedlingstein et al. (2019) are based on in situ observations and empirically upscaled eddy covariance measurements by Jung et al. (2009). The latter presents an earlier version of the FluxCom data set employed in this study. Both data sets have a much stronger carbon sink in the tropics compared to results from inversion models (Jung et al., 2020). Potential reasons for this mismatch include (i) a sampling bias towards ecosystems with a large carbon sink, (ii) missing predictor variables related to disturbance and site history, and (iii) biases of eddy covariance NEE measurements, possibly due to nighttime advection of carbon dioxide under tall tropical forest canopies (Jung et al., 2020). Furthermore, the interannual variability of FluxCom data is considerably lower compared to results from inversion methods. Evaluating FluxCom and CLASSIC against the inversion model results from CT2019 shows that the model outperforms FluxCom. Furthermore, the NEE skill scores of CT2019 and CLASSIC are similar when assessing both data sets against FLUXNET (0.49 and 0.47, respectively). We conclude that the low NEE score presented in Friedlingstein et al. (2019) is mainly a sign of observational uncertainties rather than model deficiencies.

The overall score (S_{overall}) values documented in this study are similar to the ones reported by Melton et al. (2020). The values are 0.66 (0.69) for GPP, 0.71 (0.66) for HFLS, 0.68 (0.66) for HFSS, 0.47 (0.44) for NEE, 0.64 (0.65) for RECO, and 0.8 (0.82) for RNS, where values in parenthesis correspond to results from Melton et al. (2020). The fact that both approaches yield similar results suggests that a comparison of global model output and eddy covariance data at the grid cell level provides meaningful results. This finding implies that the low scores for NEE documented here are not due to the fact that we used global rather than site-specific model inputs. However, it must be noted that interpreting the differences between both studies is challenging, given that our evaluation is based on 204 sites, while Melton et al. (2020) used 31 sites. Furthermore, our approach averages observations across sites that are located within the same model grid cell and cover the same time period.

Our benchmarking approach is subject to limitations, which are discussed next. Evaluating CLASSIC against eddy covariance measurements has numerous shortcomings. First, eddy covariance measurements from FLUXNET lack energy balance closure, where sensible and latent heat fluxes are about 20 % lower compared to the available energy (Wilson et al., 2002). Second, FLUXNET data are subject to gaps, which often occur when conditions are unsuitable for making measurements. Third, eddy covariance measurements have a much smaller footprint size (Schmid, 1994) compared to our grid cell size of $2.8125° \times 2.8125°$. The spatial mismatch implies that the measurements are not necessarily representative for the area covered by the grid cell. Considerable differences between measured and modeled values are therefore expected. Finally, the periods of evaluation differ among reference data sets. For instance, LAI data from MODIS and AVHRR cover the periods 2000–2017 and 1982–2010, respectively. Three variables assessed in this study were evaluated against one reference data set only, namely soil moisture, CO_2 emissions from fires, and soil carbon. It remains therefore unclear to what extent model performance for those three variables is affected by observational uncertainty. Future evaluations will therefore consider additional reference data sets, including soil carbon from the World Soil Information Service (Batjes et al., 2017), as well as additional in situ measurements for surface radiation provided by the Baseline Surface Radiation Network (Driemel et al., 2018). Future developments of AMBER could introduce a new score that quantifies how well a model reproduces the sensitivity of ecosystems to climate extremes, which are shown to affect terrestrial ecosystems in profound ways (Reichstein et al., 2013).

5 Conclusions

This study assesses the energy, water, and carbon cycle simulated by CLASSIC, using a wide range of in situ measurements and globally gridded reference data. We account for uncertainties in the meteorological forcing by evaluating an ensemble of three model runs that are produced with different forcing data sets. Uncertainties in the reference data are accounted for by comparing model scores against benchmark scores, which are based on reference data only. The method has been implemented in a new R package referred to as AMBER, which is publicly available at the Comprehensive R Archive Network (https://CRAN.R-project.org/package=amber, last access: 28 April 2021).

The main deficiencies that should be addressed in future model development are the (i) positive albedo bias and resulting SW radiation bias in parts of the NH extratropics and Tibetan Plateau, (ii) out-of-phase seasonal GPP cycle in the humid tropics of South America and Africa, (iii) lacking spatial correlation of annual mean NEE measured by FLUXNET sites, (iv) underestimation of fractional area burned and corresponding emissions in the boreal forests, (v) negative soil organic carbon bias in high latitudes, and (vi) time lag in seasonal LAI maxima in the NH extratropics. Recent model development has started addressing some of those issues already, including the improvement of LAI seasonality through the incorporation of non-structural carbohydrates, which will form part of the next model version release (Asaadi et al., 2018). Further research is required to separate the impact of observational uncertainties on biases in LAI and above-

ground biomass. For LAI, we propose to add additional observation-based reference data sets and reduce the spatial coverage to high-quality grid cells that have not been gap filled. For biomass, we propose to extend our current forest inventory database to achieve greater spatial coverage.

The complexity of LSMs and increasing availability of global Earth observations demand comprehensive methods of model evaluation. Such methods must account for uncertainties associated with model inputs and observation-based reference data to yield robust results. Future research could explore the question of why meteorological forcing data sets used in this study differ and how this uncertainty could be reduced. A centralized and updated reference database that includes spatially explicit information of uncertainties would also be highly valuable.

Our results show that model performance is very sensitive to the choice of meteorological forcing and reference data, confirming recent findings presented in Lawrence et al. (2019) and Bonan et al. (2019). For instance, for about half of all variables assessed in this study, the sign of the bias changes depending on what forcing and reference data are used. Comparing model and benchmark scores confirms that CLASSIC performance is strongly affected by observational uncertainty. Taking this uncertainty into account shows that CLASSIC performs very well across a wide range of variables. Our results will serve as a baseline for guiding and monitoring future CLASSIC development.

Appendix A

The computation of GPP is based on the parameterizations by Farquhar et al. (1980), Collatz et al. (1991), and Collatz et al. (1992), with some minor adjustments documented in Arora and Boer (2003). The gross leaf photosynthesis rate (G_o) is limited by the availability of photosynthetically active radiation (PAR) (J_e), RuBisCO (J_c), and the transport capacity (J_s) (Melton and Arora, 2016).

To compute GPP, G_o is first downregulated in response to nutrient availability and then multiplied with the fraction of photosynthetically active radiation f_{PAR}:

$$f_{\mathrm{PAR}} = \frac{1}{k_{\mathrm{n}}}\left(1 - \exp(-k_{\mathrm{n}}\Lambda)\right), \tag{A1}$$

where k_{n} is the extinction coefficient that describes the nitrogen and time-mean PAR profile along the depth of the canopy and Λ is the leaf area index (Arora et al., 2009).

Λ is computed as the product of specific leaf area (SLA) and leaf biomass (C_{L}), where

$$\mathrm{SLA} = 25.0\Upsilon_{\mathrm{L}}^{-0.5}, \tag{A2}$$

and Υ_{L} is a PFT-specific leaf life span measured in years.

Ecosystem respiration is the sum of autotrophic (R_{a}) and heterotrophic respiration (R_{h}). Autotrophic respiration is the sum of maintenance (R_{m}) and growth respiration (R_{g}). Maintenance respiration is divided into respiration associated with the maintenance of leaves (R_{mL}), stems (R_{mS}), and roots (R_{mR}). Leaf maintenance respiration is scaled from the leaf to the canopy level using f_{PAR}, and growth respiration (R_{g}) is estimated as a fraction of GPP. Heterotrophic respiration is the sum of respiration from litter and soil carbon and is affected by soil temperature and soil moisture.

The reflectivity of plant surfaces varies considerably within the shortwave spectrum, with higher absorptivity in the visible range due to photosynthesis and lower absorptivity in the near-infrared (IR) range. CLASSIC therefore computes surface albedo for the visible and near-IR spectrums separately. Both approaches are identical, except for the use of different constants that affect the background albedos for soil, vegetation, and snow, as well as canopy transmissivities. The surface albedo for the visible spectrum is the sum of the visible albedos of (i) bare ground ($\alpha_{\mathrm{vs,g}}$), (ii) open snow ($\alpha_{\mathrm{vs,sn}}$), (iii) vegetation over bare ground ($\alpha_{\mathrm{vs,cn}}$), and (iv) vegetation over snow ($\alpha_{\mathrm{vs,cs}}$), multiplied with the corresponding subarea fractional coverage (f_{g}, f_{gs}, f_{c}, and f_{cs}, respectively; Fig. A25):

$$\alpha_{\mathrm{vs}} = f_{\mathrm{c}}\alpha_{\mathrm{vs,cn}} + f_{\mathrm{g}}\alpha_{\mathrm{vs,g}} + f_{\mathrm{cs}}\alpha_{\mathrm{vs,cs}} + f_{\mathrm{gs}}\alpha_{\mathrm{vs,sn}}. \tag{A3}$$

The visible albedo of bare ground ($\alpha_{\mathrm{vs,g}}$) ranges from the visible albedo of wet soil ($\alpha_{\mathrm{g,w,v}}$) to the visible albedo of dry soil ($\alpha_{\mathrm{g,d,v}}$), depending on the soil moisture content in the first soil layer ($\theta_{\mathrm{l,1}}$). The visible albedo of open snow ($\alpha_{\mathrm{vs,sn}}$) depends on the all-wave albedo of the snowpack (α_{sno}). The visible albedo of vegetation over bare ground ($\alpha_{\mathrm{vs,cn}}$) is the product of the total vegetation albedo of a PFT over bare soil ($\alpha_{\mathrm{sn,n}}$) and its corresponding fractional coverage (f_{can}). The former depends on the canopy albedo ($\alpha_{\mathrm{vs,cx}}$), the visible albedo of ground under vegetation canopy ($\alpha_{\mathrm{vs,gc}}$), the sky view factor (χ), and the visible transmissivity of the vegetation (τ_{vs}):

$$\alpha_{\mathrm{vs,n}} = (1-\chi)\alpha_{\mathrm{vs,cx}} + \chi\tau_{\mathrm{vs}}\alpha_{\mathrm{vs,gc}}. \tag{A4}$$

The visible albedo of ground under vegetation canopy ($\alpha_{\mathrm{vs,gc}}$) is set equal to the visible albedo of bare ground ($\alpha_{\mathrm{vs,g}}$) computed above. The canopy albedo $\alpha_{\mathrm{vs,cx}}$ is a function of the fractional coverage of canopy by frozen water ($f_{\mathrm{snow,c}}$) and the average background visible albedo of snow-covered vegetation ($\alpha_{\mathrm{vs,wc}}$):

$$\alpha_{\mathrm{vs,cx}} = f_{\mathrm{snow,c}}\alpha_{\mathrm{vs,wc}} + (1.0 - f_{\mathrm{snow,c}})\alpha_{\mathrm{vs,c}}. \tag{A5}$$

The visible transmissivity of the vegetation (τ_{vs}) follows Beer's law of radiation transfer:

$$\tau_{\mathrm{vs}} = \exp(-\kappa\Lambda_{\mathrm{pai}}), \tag{A6}$$

where κ is the canopy extinction coefficient, which depends on the cloud fraction, the solar zenith angle, and PFT-specific constants. The sky view factor (χ) expresses the degree of closure of a canopy, and is a function of a PFT-specific constant (c) and the plant area index (Λ_{pai}), which is the sum of Λ and stem area index:

$$\chi = \exp(c\Lambda_{\mathrm{pai}}). \tag{A7}$$

The same approach described for computing $\alpha_{\mathrm{vs,cn}}$ applies to the visible albedo of vegetation over snow ($\alpha_{\mathrm{vs,cs}}$), except that the terms $f_{\mathrm{can}}\alpha_{\mathrm{vs,n}}$ and $\alpha_{\mathrm{vs,gc}}$ are replaced with the corresponding terms for canopy over snow, i.e. $f_{\mathrm{cans}}\alpha_{\mathrm{vs,s}}$ and $\alpha_{\mathrm{vs,sc}}$.

The visible and near-IR albedos and canopy transmissivities obtained above are now used to compute the net SW radiation for the canopy ($K_{*,\mathrm{c}}$) and underlying ground (or snow) ($K_{*,\mathrm{g}}$):

$$\begin{aligned} K_{*,\mathrm{g}} &= \hat{\tau}_{\mathrm{vs}}K^{\downarrow}_{\mathrm{vs}}\left(1-\alpha_{\mathrm{vs,g}}\right) + \hat{\tau}_{\mathrm{nir}}K^{\downarrow}_{\mathrm{nir}}\left(1-\alpha_{\mathrm{nir,g}}\right) \text{ and} \\ K_{*,\mathrm{c}} &= K^{\downarrow}_{\mathrm{vs}}\left(1-\alpha_{\mathrm{vs,c}}\right) + K^{\downarrow}_{\mathrm{nir}}\left(1-\alpha_{\mathrm{nir,c}}\right) - K_{*,\mathrm{g}}, \end{aligned} \tag{A8}$$

where $K^{\downarrow}$ is the downwelling SW radiation, carets denote values that are averaged over multiple PFTs, and the suffix "nir" indicates that the respective variable applies to the near-IR spectrum (Fig. A26). The net SW radiation is then the sum of $K_{*,\mathrm{g}}$, $K_{*,\mathrm{c}}$, and the SW radiation transmitted into snow, if present (Q_{trans}).

The net LW radiation for the canopy ($L_{*,\mathrm{g}}$) and underlying ground (or snow) ($L_{*,\mathrm{c}}$) follow the Stefan–Boltzmann law:

$$\begin{aligned} L_{*,\mathrm{g}} &= \left(1-\hat{\chi}\right)\sigma\overline{T}_{\mathrm{c}}^{4} + \hat{\chi}L^{\downarrow} - \sigma T(0)^{4} \text{ and} \\ L_{*,\mathrm{c}} &= \left(1-\hat{\chi}\right)\left(L^{\downarrow} + \sigma T(0)^{4} - 2\sigma\overline{T}_{\mathrm{c}}^{4}\right), \end{aligned} \tag{A9}$$

where $L^{\downarrow}$ is the downwelling LW radiation, σ is the Stefan–Boltzmann constant, $\overline{T}_c$ is the effective canopy temperature, and $T(0)$ is the temperature of the underlying ground (Fig. A26).

The sensible heat flux from the ground ($Q_{H,g}$) and canopy $Q_{H,c}$ are given as

$$Q_{H,g} = \rho_a c_p (T(0)_{pot} - T_{a,c})/r_{a,g} \text{ and}$$
$$Q_{H,c} = \rho_a c_p (\overline{T}_c - T_{a,c})/r_b, \quad \text{(A10)}$$

where ρ_a is the air density, c_p is the specific heat capacity, $T(0)_{pot}$ is the potential air temperature at the ground, $T_{a,c}$ is the canopy air temperature, $r_{a,g}$ is the surface resistance, and r_b is the leaf boundary resistance. Air density is a function of air temperature and pressure as defined in the ideal gas law. The resistances $r_{a,g}$ and r_b are defined as

$$\frac{1}{r_{a,g}} = 1.9 \times 10^{-3} (T(0)_v - T_{ac,v})^{1/3} \text{ and}$$
$$\frac{1}{r_b} = v_{ac}^{0.5} \sum f_i \gamma_i \Lambda_i^{0.5} / 0.75[1 - \exp(-0.75\Lambda_i^{0.5})], \quad \text{(A11)}$$

where $T(0)_v$ is the virtual potential temperature of the surface, $T_{ac,v}$ is the virtual temperature of the air in the canopy space, v_{ac} is the wind speed in the canopy air space, f_i is the fractional coverage of each vegetation type i over the subarea in question, and γ_i is a vegetation-dependent parameter that incorporates the effects of leaf dimension and sheltering.

The latent heat flux from the ground ($Q_{E,g}$) and canopy ($Q_{E,c}$) are

$$Q_{E,g} = L_v \rho_a (q(0) - q_{a,c})/r_{a,g} \text{ and}$$
$$Q_{E,c} = L_v \rho_a (q_c - q_{a,c})/(r_b + r_c), \quad \text{(A12)}$$

where L_v is the latent heat of vaporization, $q(0)$ is the specific humidity at surface, $q_{a,c}$ is the specific humidity of air within vegetation canopy space, q_c is the saturated specific humidity at canopy temperature, and r_c is the stomatal resistance to transpiration. In the absence of water stress, the unstressed r_c is a function of $K^{\downarrow}$, Λ, and κ. If soil moisture is limiting transpiration, r_c depends on air temperature, vapor pressure deficit, and a soil moisture suction coefficient.

The soil heat flux ($G(0)$) is estimated from the vertical soil temperature gradient and the thermal conductivity of the first three soil layers:

$$G(0) = -\lambda_t \frac{\partial T}{\partial z}. \quad \text{(A13)}$$

The same approach is applied in the presence of snow, where λ_t is replaced with the thermal conductivity of snow (λ_s) and T becomes the snow temperature (T_s). The latter depends on the snow heat capacity, which is a function of snow water content.

The surface and canopy energy balance are then expressed as

$$K_{*,g} + L_{*,g} - Q_{H,g} - Q_{E,g} - G(0) = 0 \text{ and}$$
$$K_{*,c} + L_{*,c} - Q_{H,c} + Q_{H,g} - Q_{E,c} - \Delta Q_{S,c} = 0, \quad \text{(A14)}$$

where $\Delta Q_{S,c}$ represents the change of energy storage in the canopy. In practice, the terms on the left-hand side do not necessarily add up to zero. CLASSIC attempts to approximate a zero net balance by slightly adjusting $T(0)$ until the absolute residual of the surface energy balance is less than $5.0\,W\,m^{-2}$, or until the absolute value of the iteration step most recently used is less than 0.01 K.

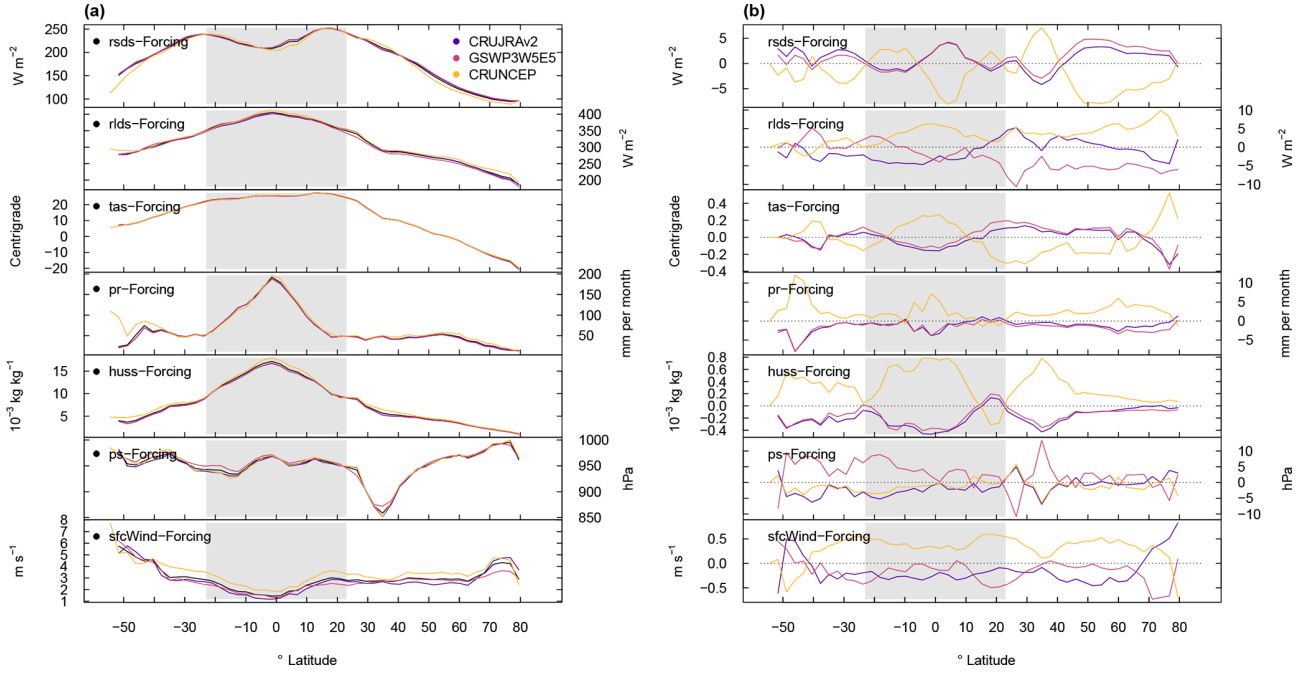

Figure A1. **(a)** Zonal mean values of meteorological forcing data and **(b)** corresponding difference between the individual forcing and the mean of the three forcings, where rsds and rlds are the surface downwelling SW and LW radiation, respectively, tas is the near-surface air temperature, pr is the precipitation, huss is the near-surface specific humidity, ps is the surface pressure, and sfcWind is the near-surface wind speed.

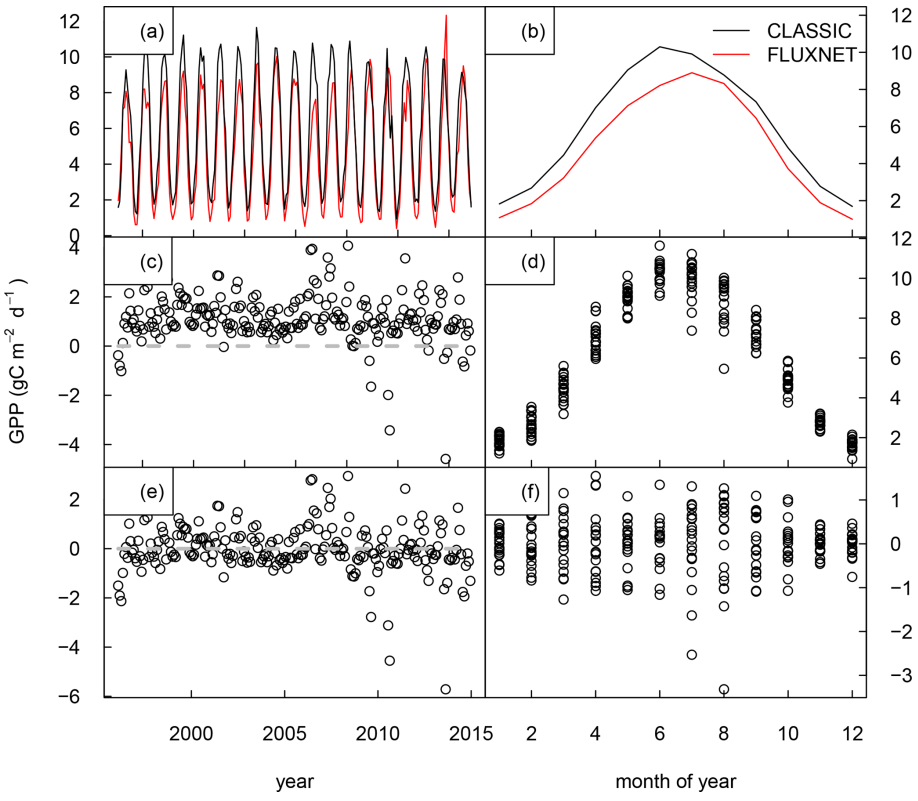

Figure A2. The left-hand column shows monthly values of **(a)** modeled (black) and observed (red) gross primary productivity, **(c)** residuals, and **(e)** residuals of the anomalies. The right-hand column gives the corresponding model values for the **(b)** climatological mean seasonal cycle, **(d)** monthly variability, and **(f)** monthly variability of anomalies. The data correspond to CLASSIC model output for the location of an eddy covariance measurement site in the Netherlands (NL-Loo).

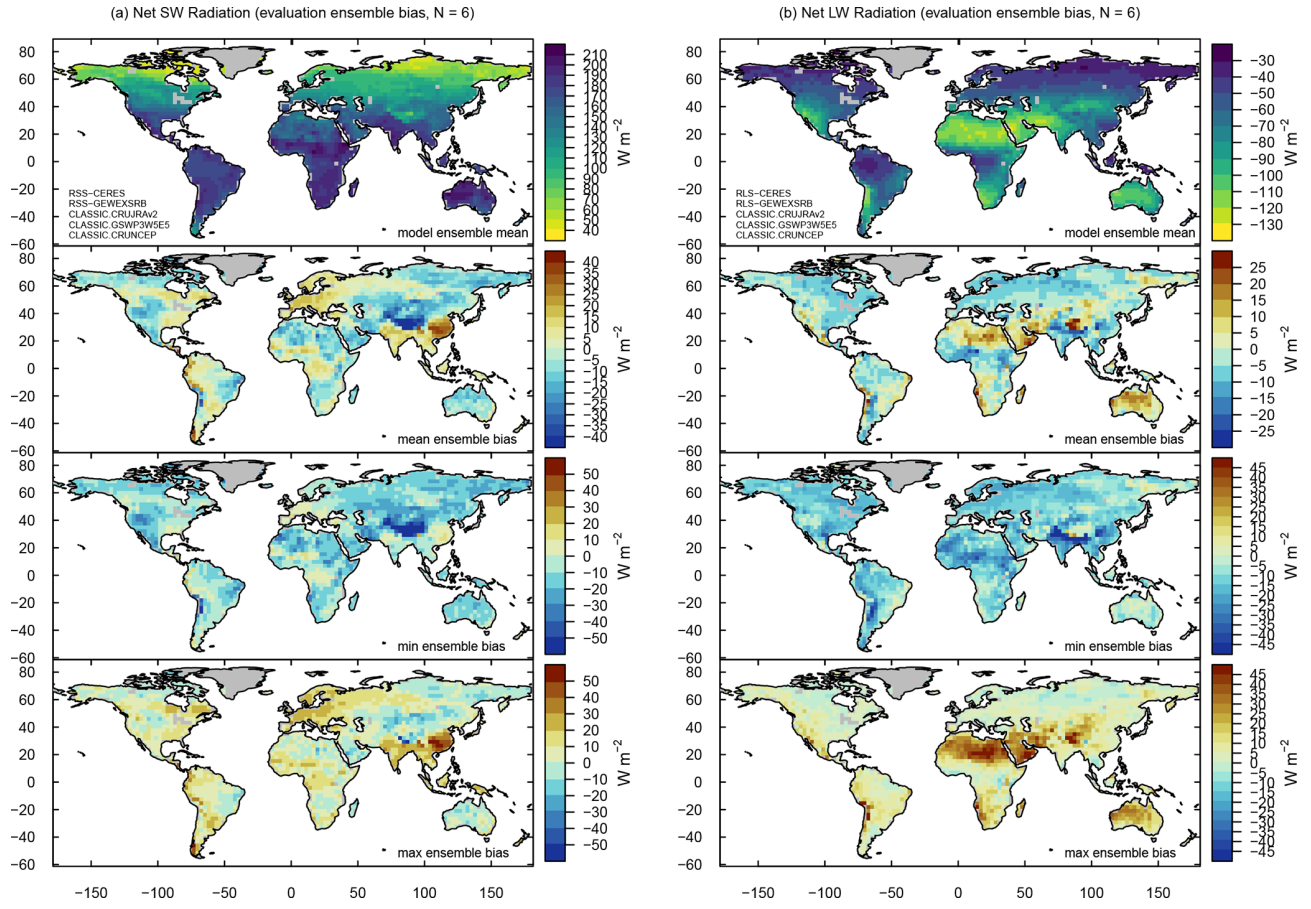

Figure A3. Same as Fig. 7 but for **(a)** net SW radiation and **(b)** net LW radiation.

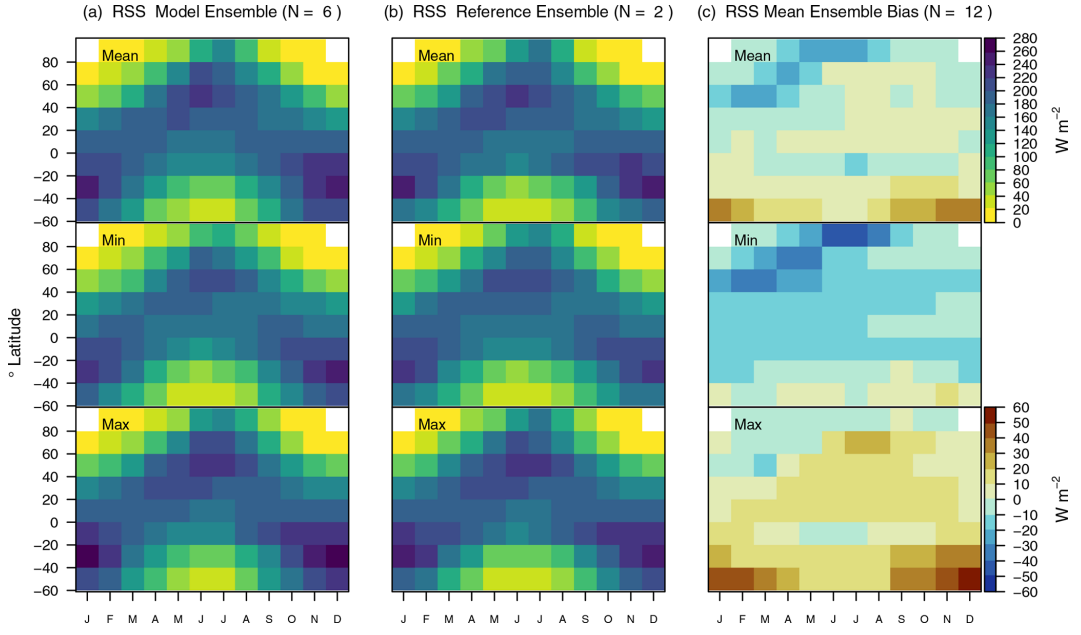

Figure A4. Same as in Fig. 8 but for net SW radiation.

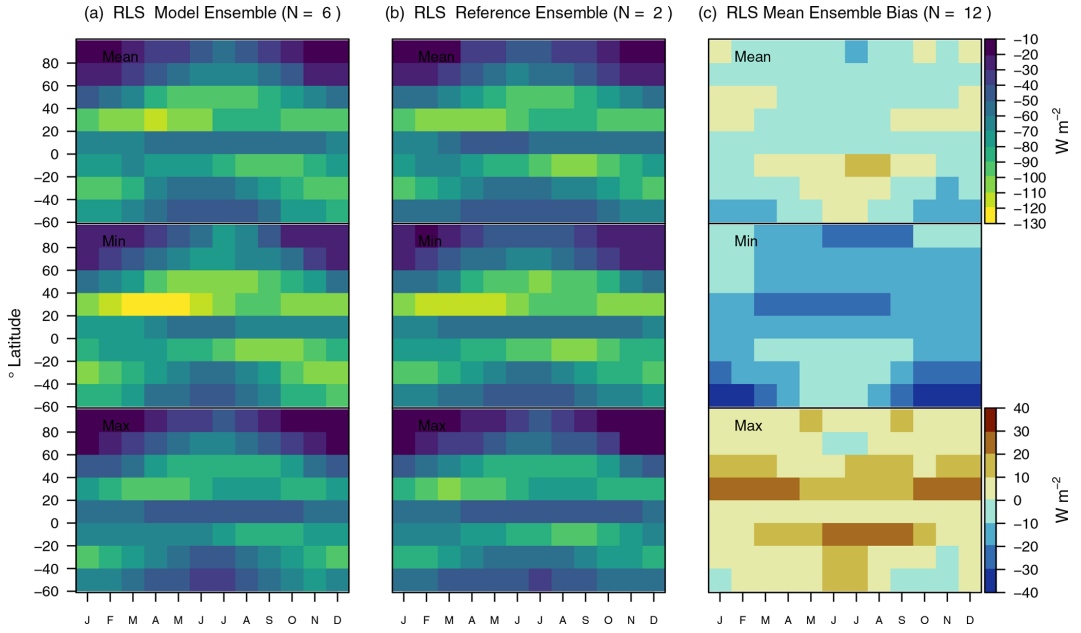

Figure A5. Same as in Fig. 8 but for net LW radiation.

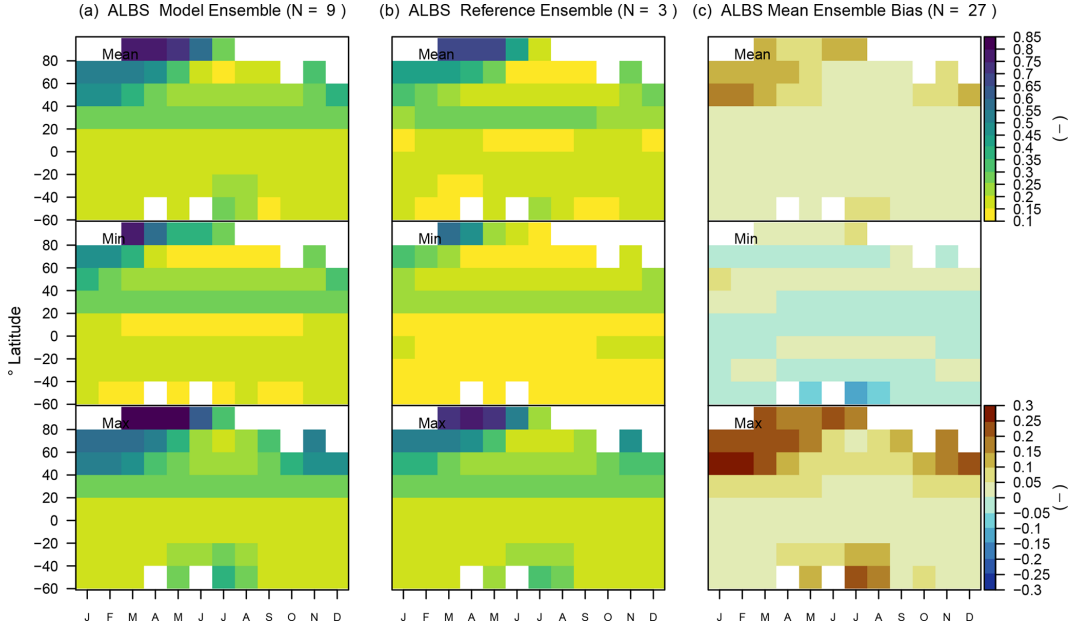

Figure A6. Same as in Fig. 8 but for surface albedo.

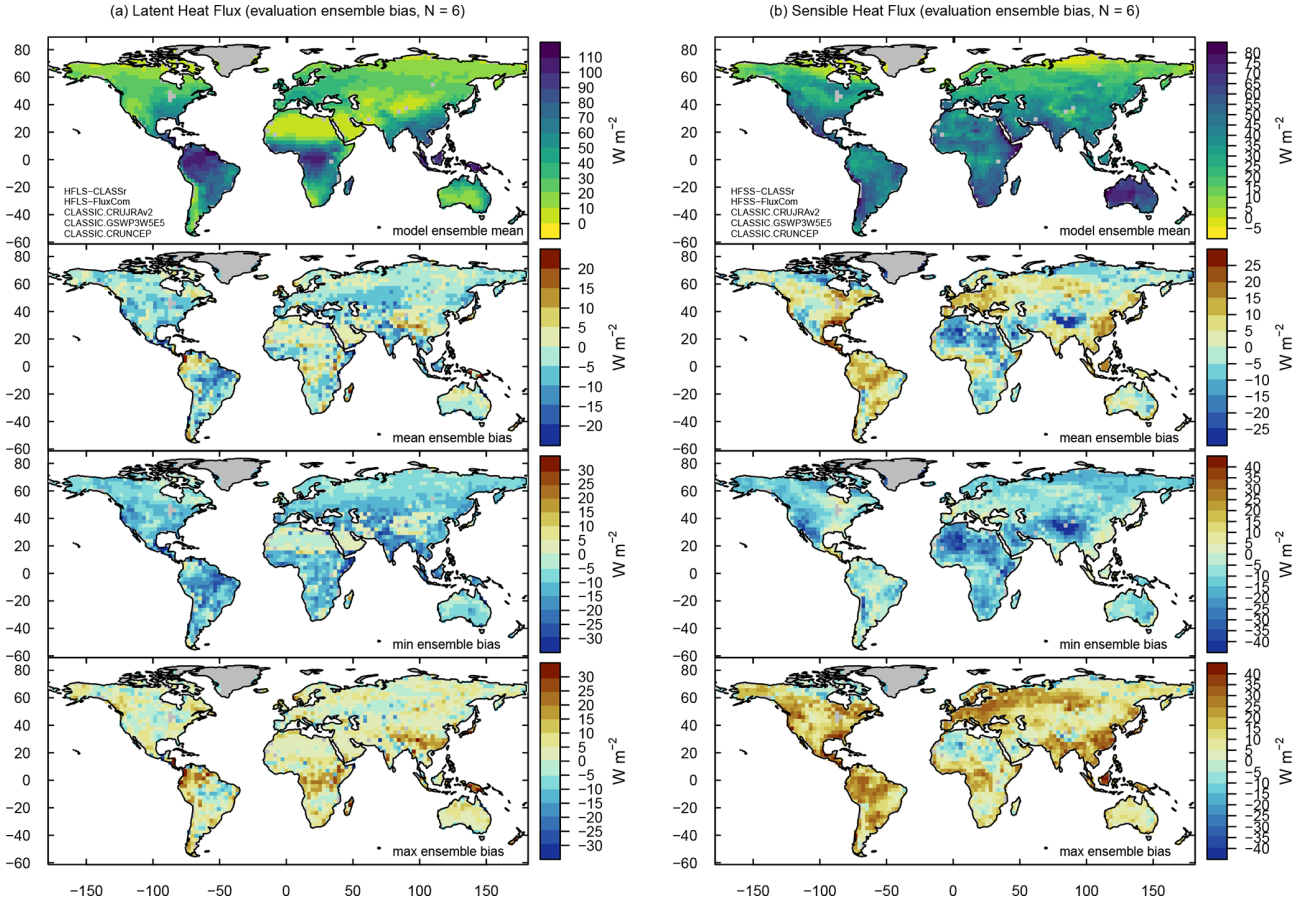

Figure A7. Same as Fig. 7 but for **(a)** latent heat flux and **(b)** sensible heat flux.

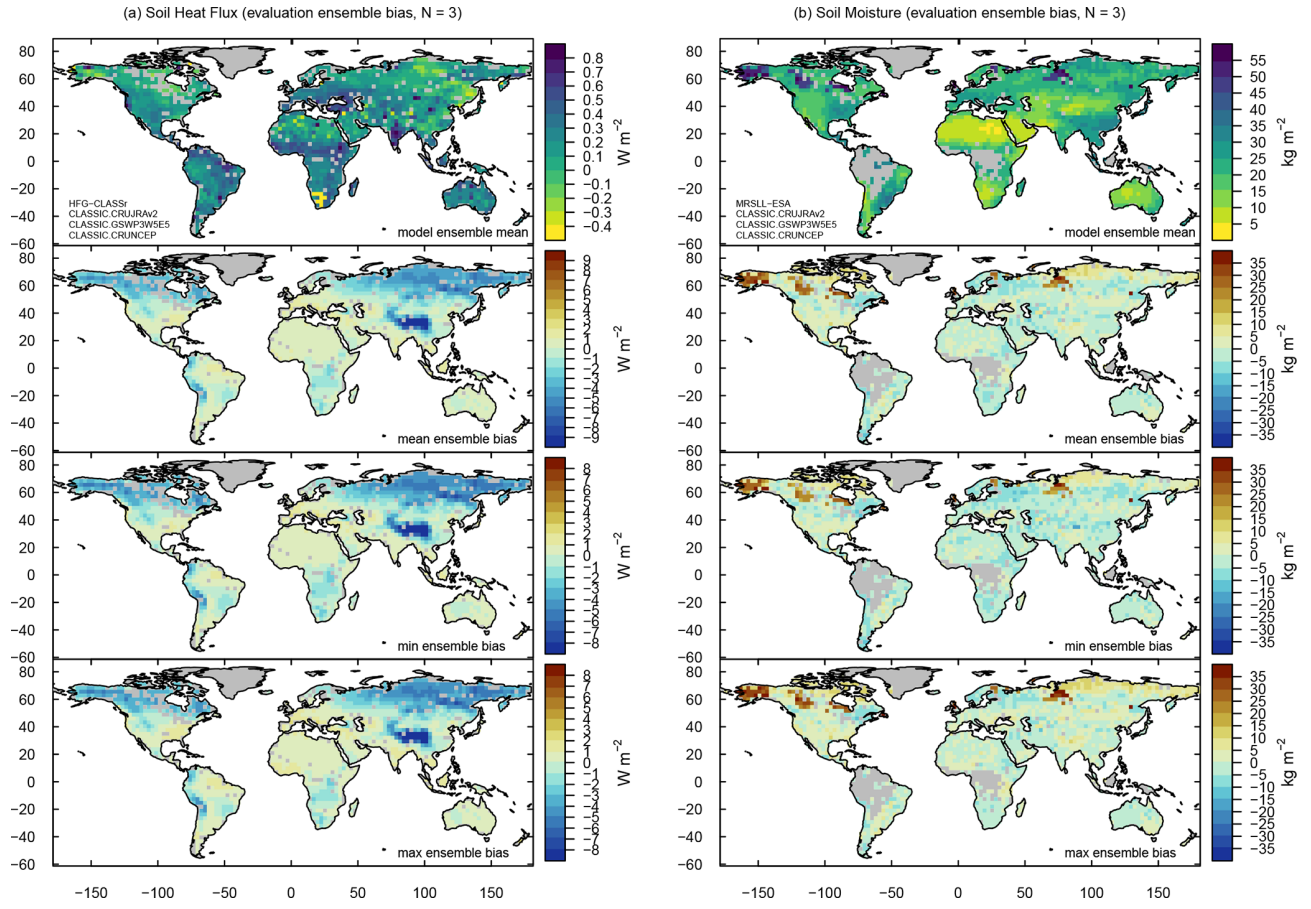

Figure A8. Same as Fig. 7 but for **(a)** soil heat flux and **(b)** soil moisture in the top 10 cm.

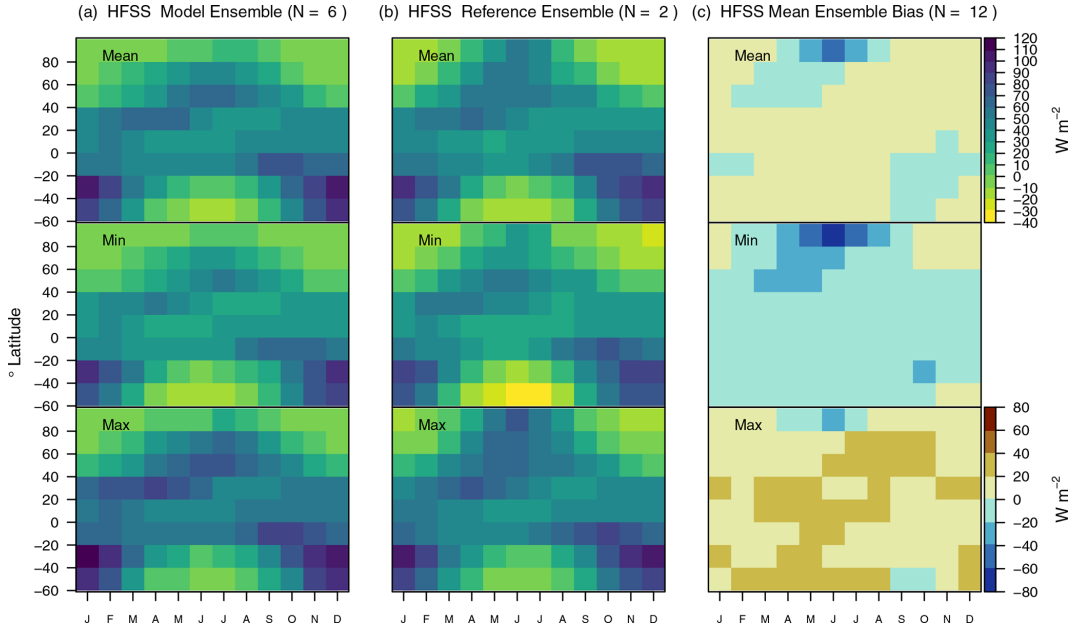

Figure A9. Same as in Fig. 8 but for sensible heat flux.

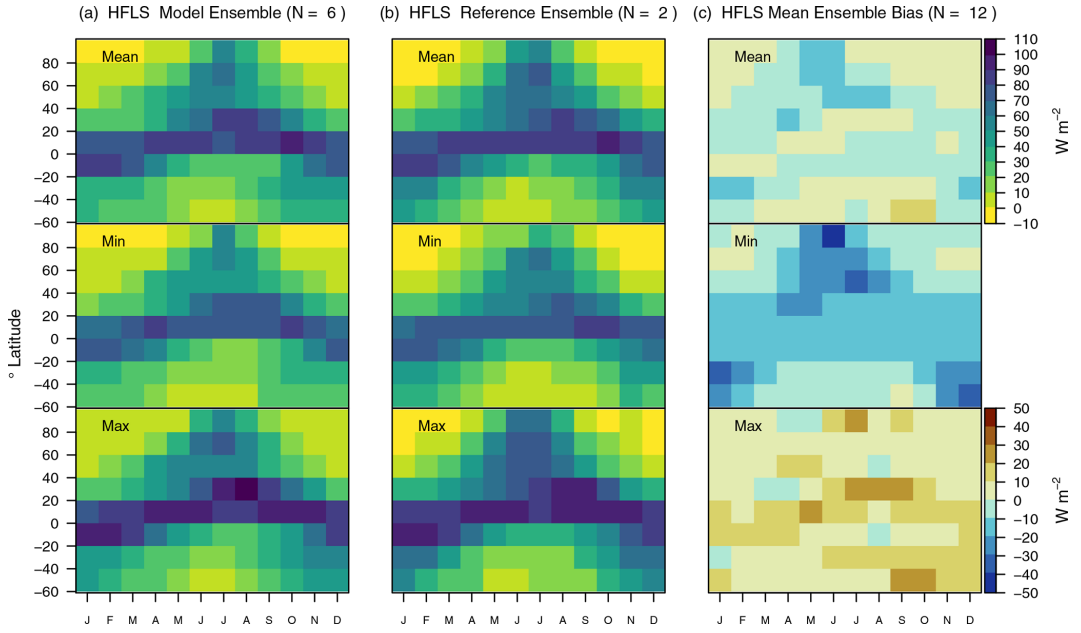

Figure A10. Same as in Fig. 8 but for latent heat flux.

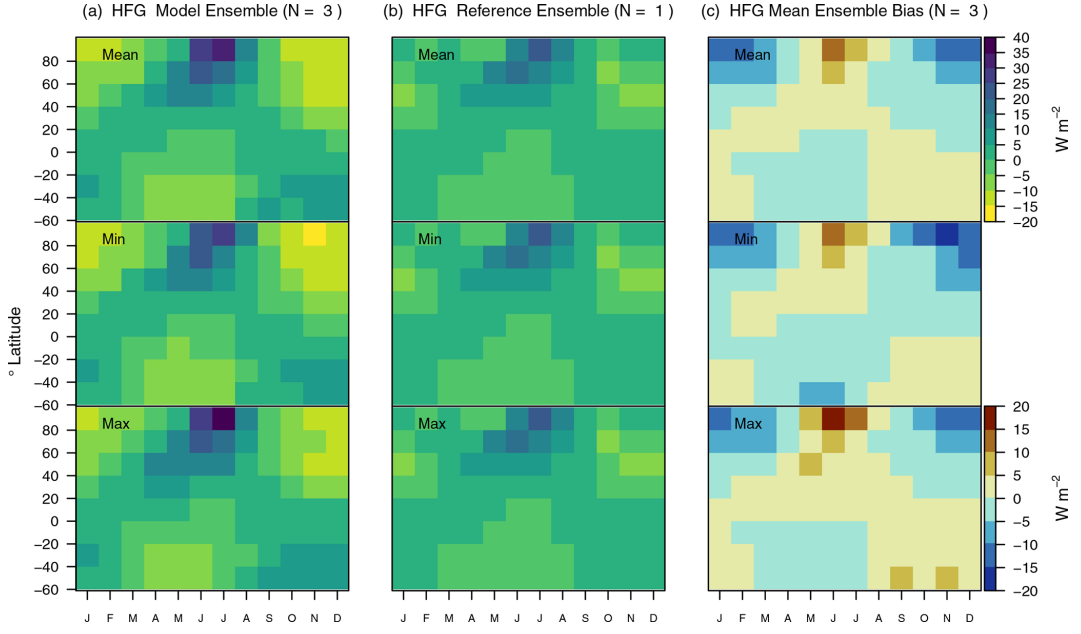

Figure A11. Same as in Fig. 8 but for soil heat flux.

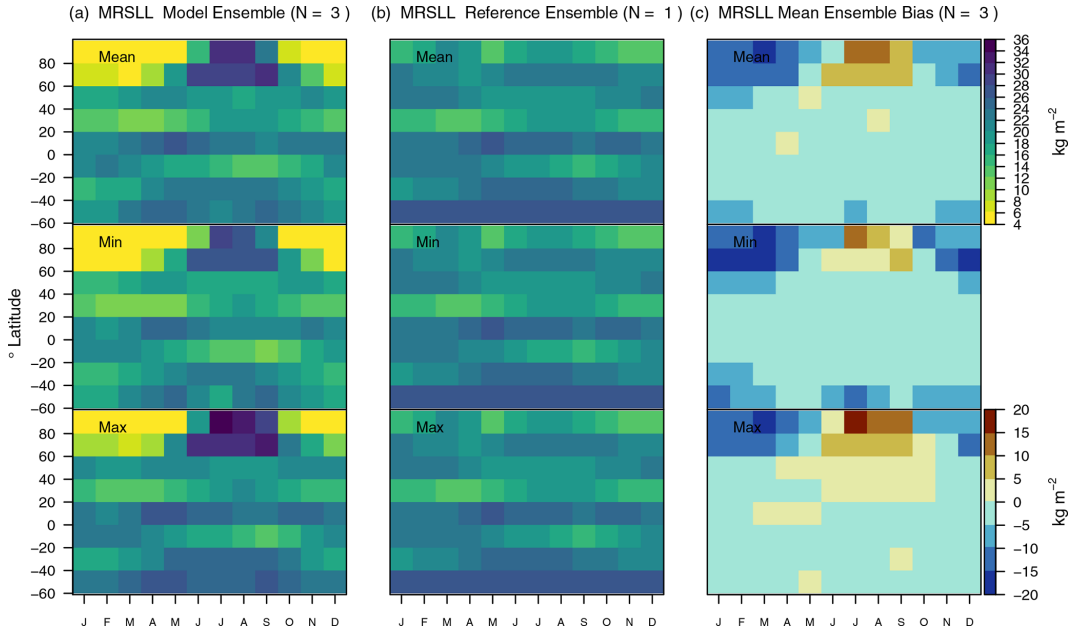

Figure A12. Same as in Fig. 8 but for soil moisture.

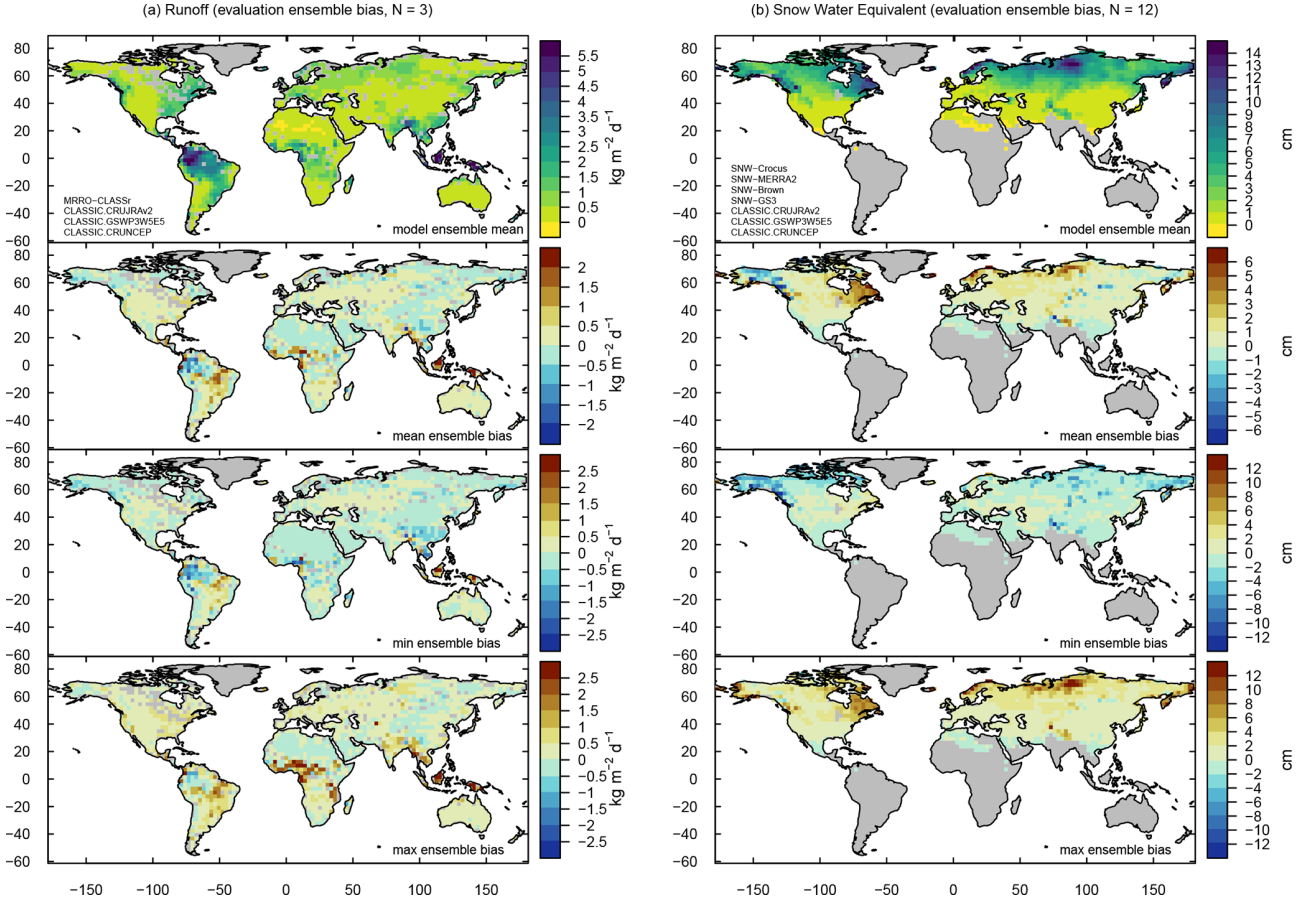

Figure A13. Same as Fig. 7 but for **(a)** runoff and **(b)** snow water equivalent.

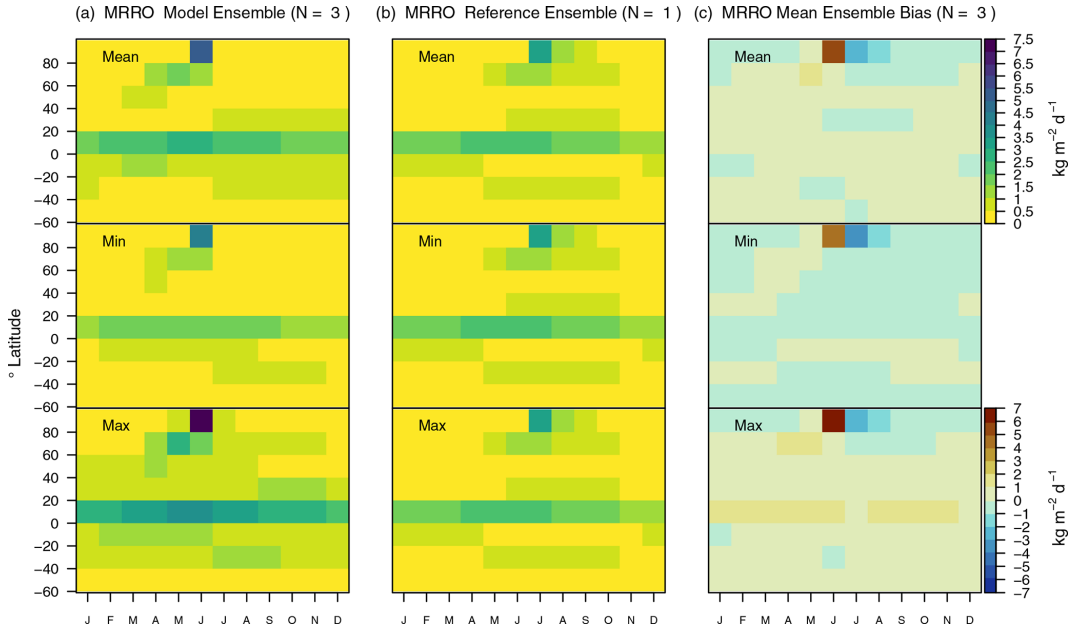

Figure A14. Same as in Fig. 8 but for runoff.

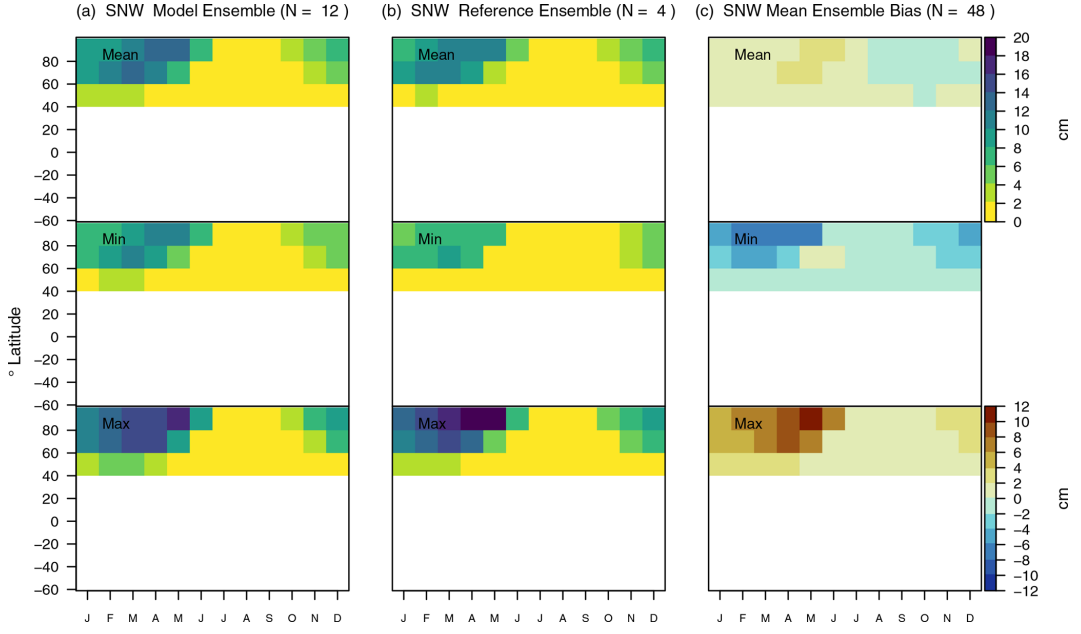

Figure A15. Same as in Fig. 8 but for snow water equivalent.

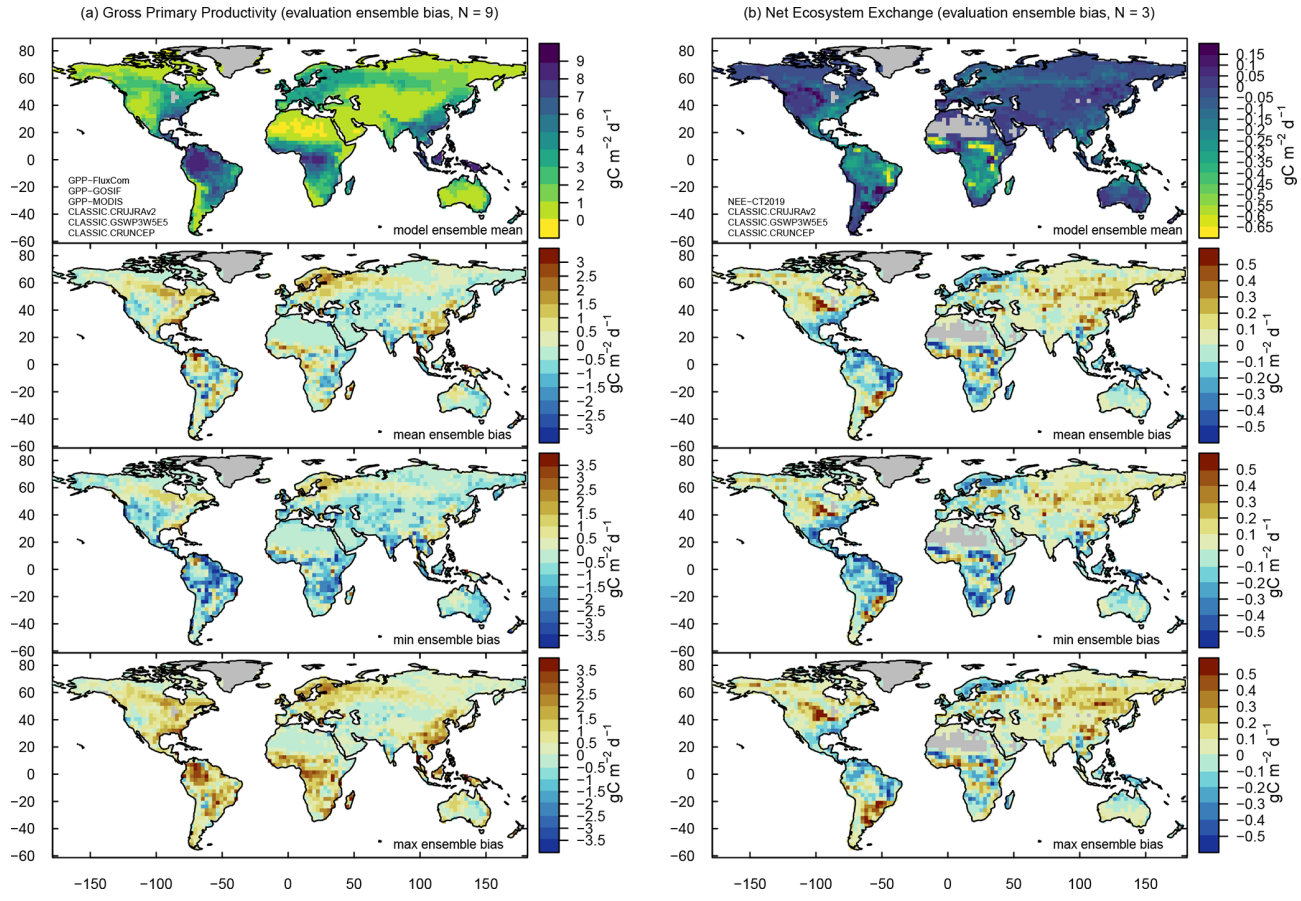

Figure A16. Same as Fig. 7 but for **(a)** gross primary productivity and **(b)** net ecosystem exchange.

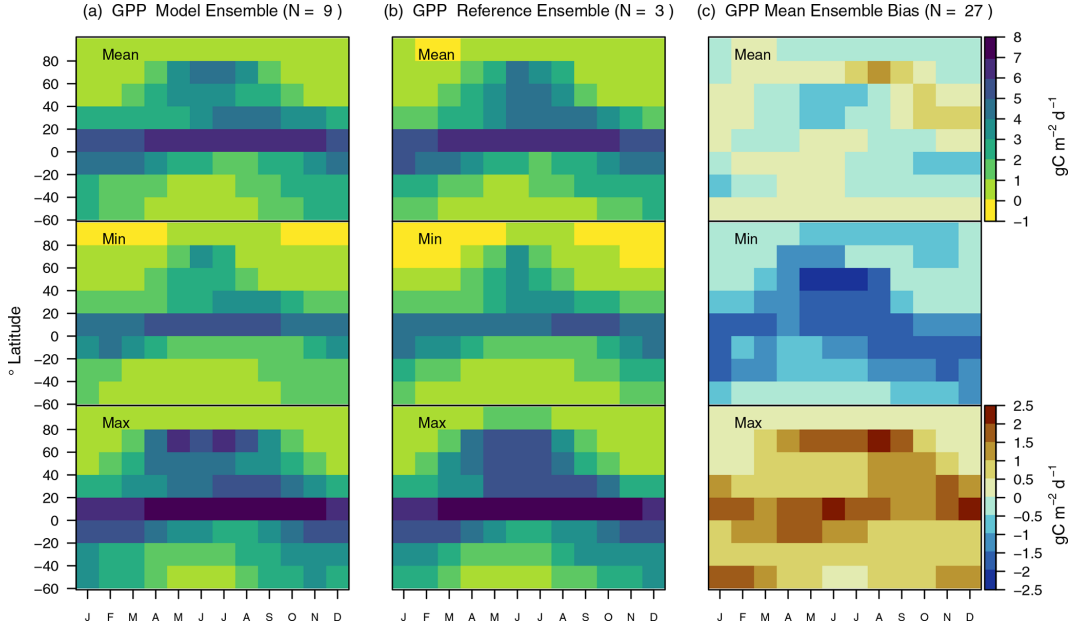

Figure A17. Same as in Fig. 8 but for gross primary productivity.

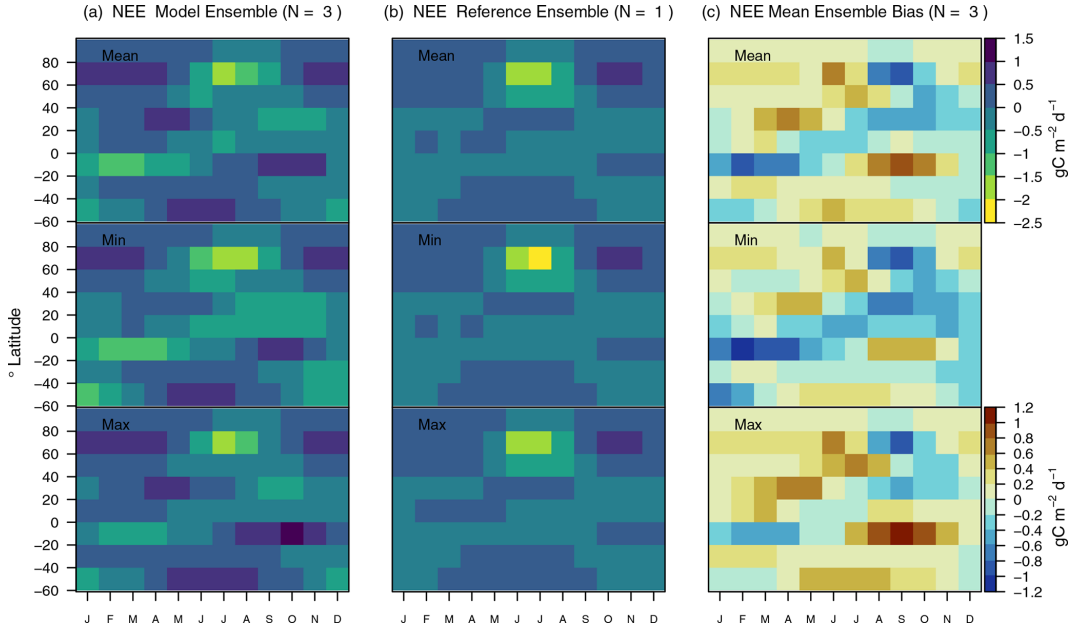

Figure A18. Same as in Fig. 8 but for net ecosystem exchange.

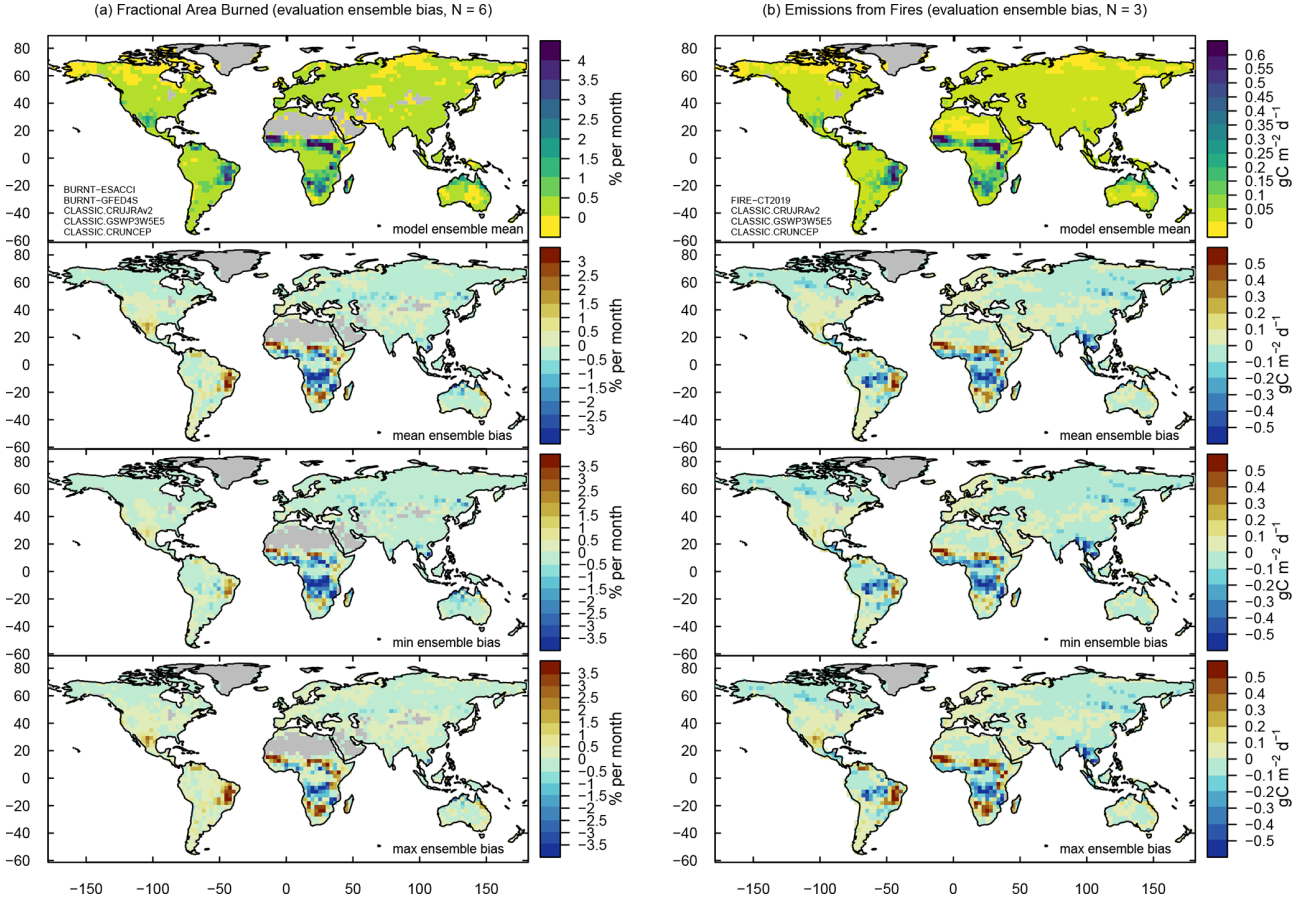

Figure A19. Same as Fig. 7 but for **(a)** fractional area burned and **(b)** emissions from wildfires.

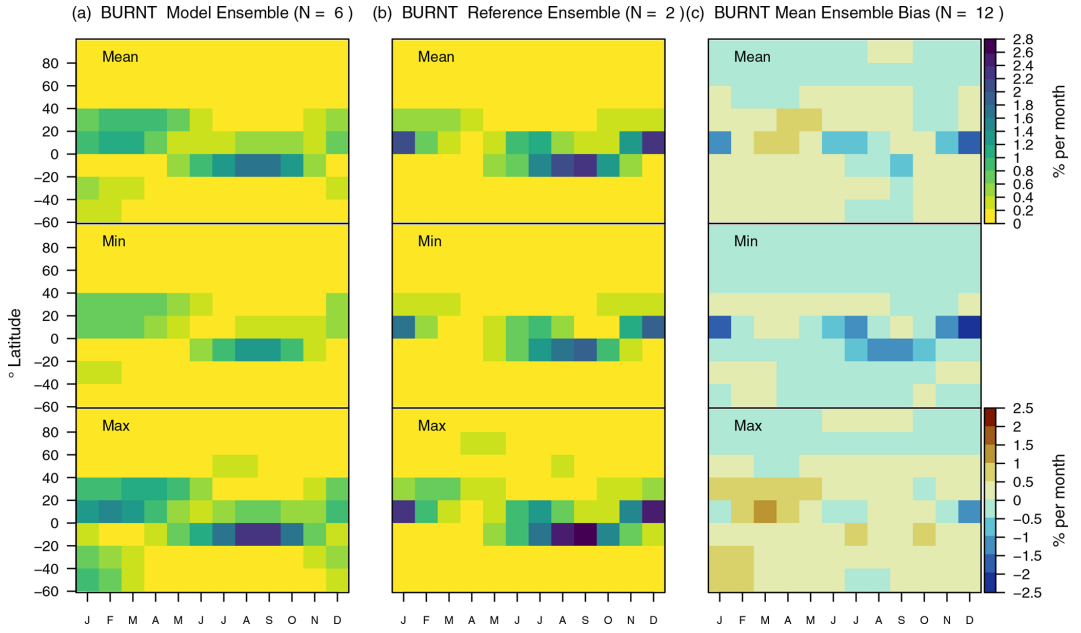

Figure A20. Same as in Fig. 8 but for fractional area burned.

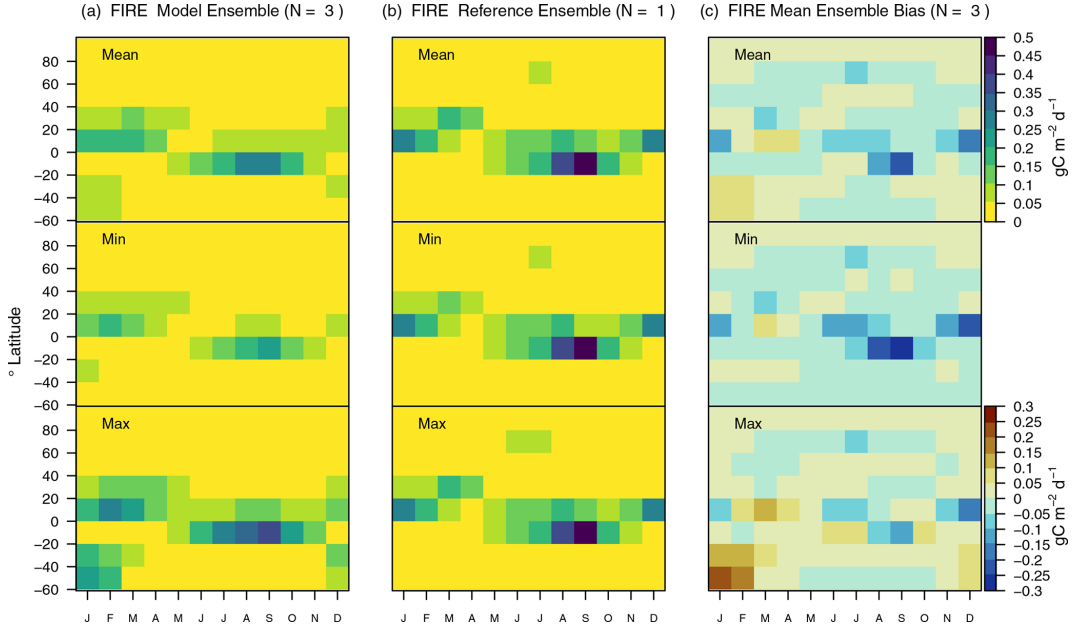

Figure A21. Same as in Fig. 8 but for emissions from wildfires.

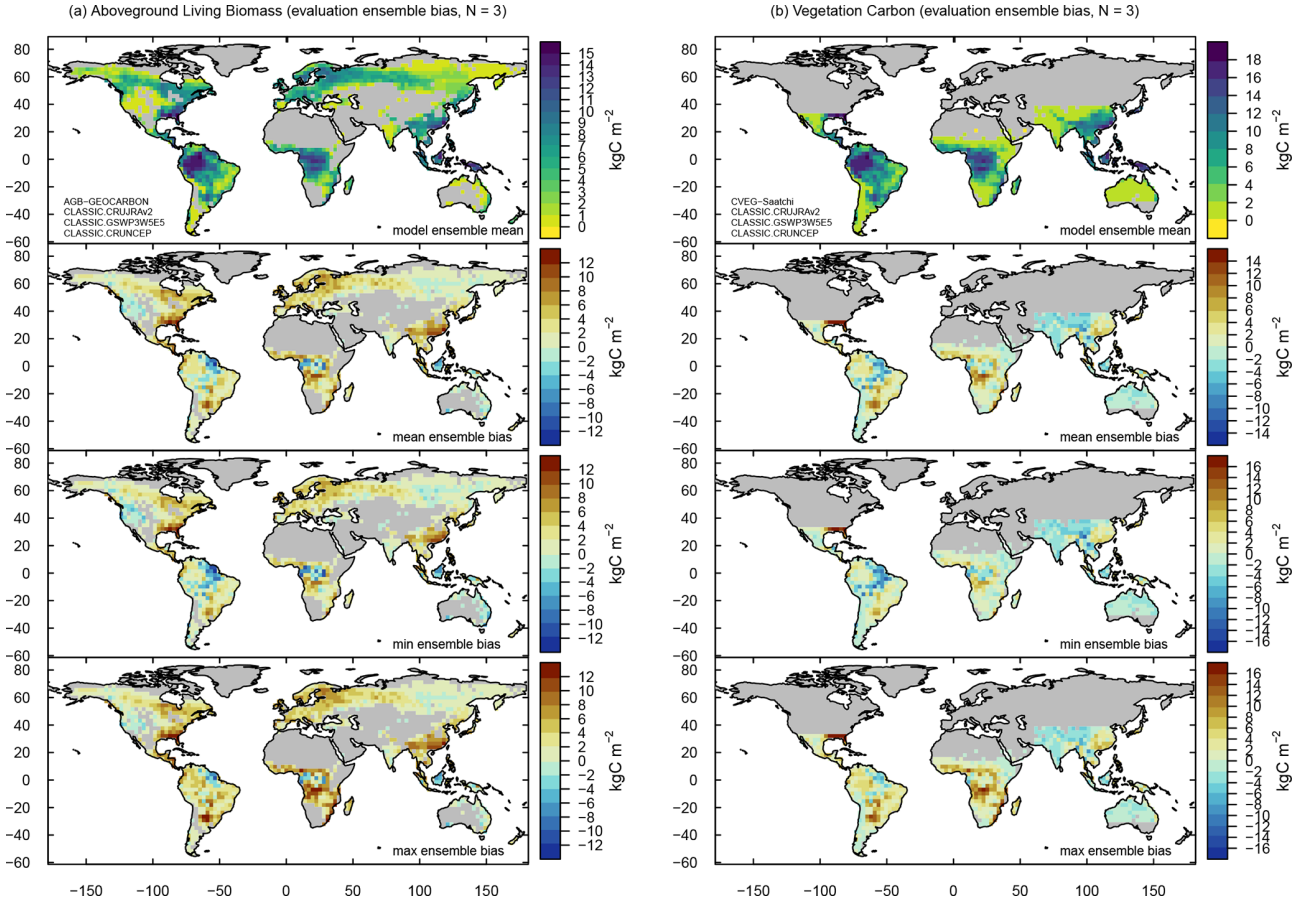

Figure A22. Same as Fig. 7 but for **(a)** above-ground living biomass and **(b)** vegetation carbon in the tropics.

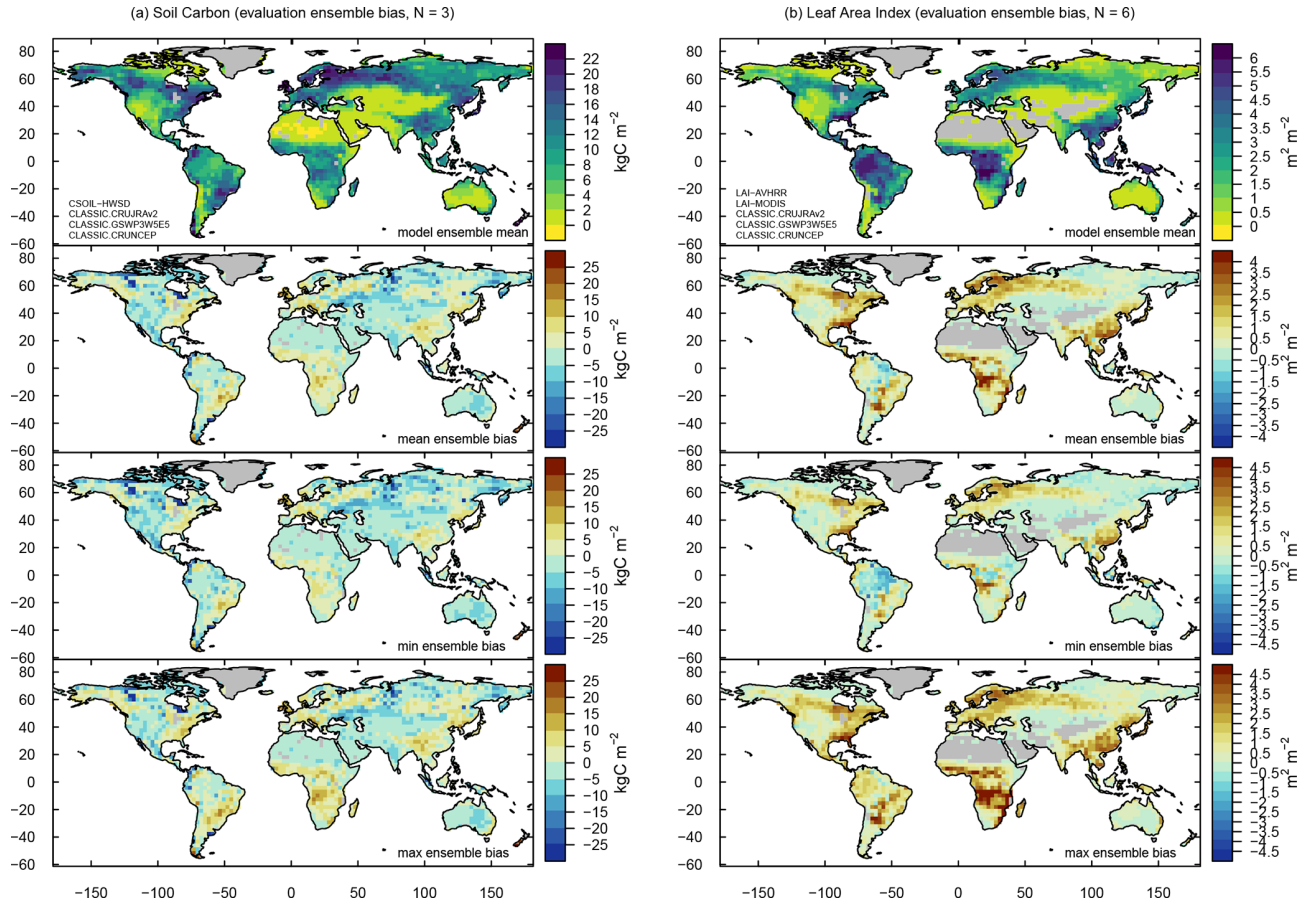

Figure A23. Same as Fig. 7 but for **(a)** soil organic carbon mass and **(b)** leaf area index.

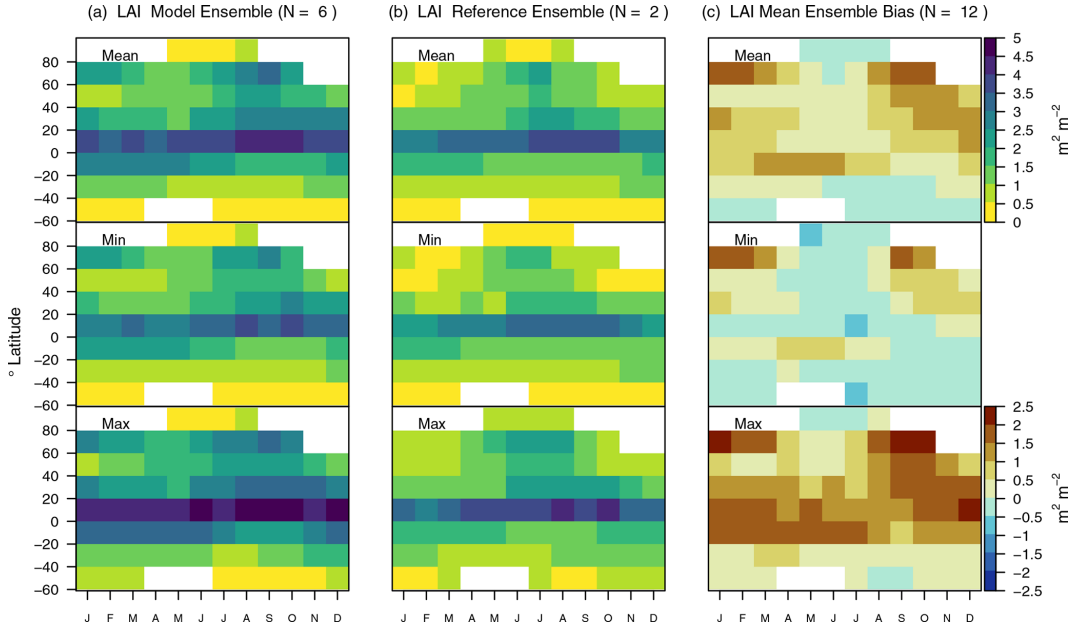

Figure A24. Same as in Fig. 8 but for leaf area index.

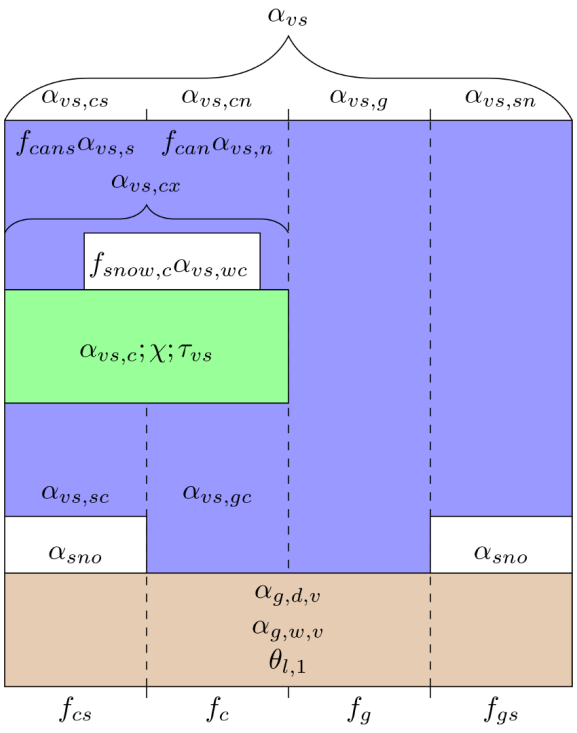

Figure A25. Contributions from the bare ground (brown), canopy (green), and snow (white) to the total visible albedo of the land surface. All variables are defined in the Appendix.

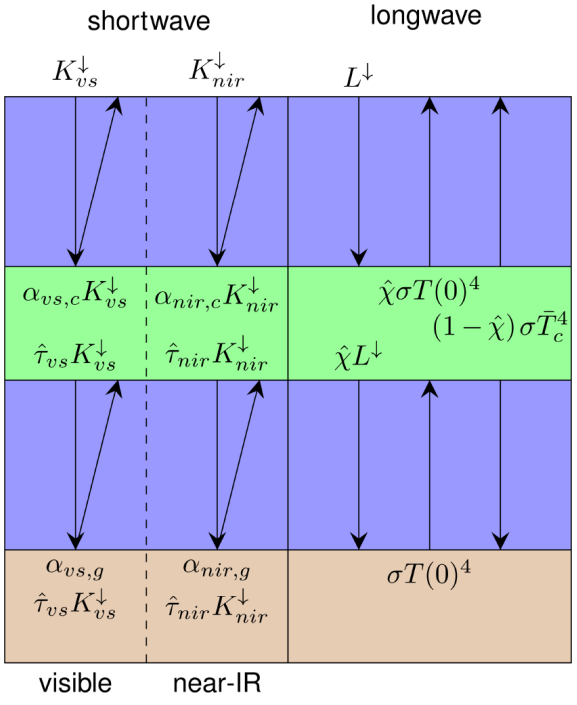

Figure A26. Longwave and shortwave radiation fluxes between the ground (brown), canopy (green), and overlying atmosphere. The meaning of each arrow is given at its tail. All variables are defined in the Appendix.

Code and data availability. The data, scripts, code, dependencies, and instructions required for reproducing the results presented in this paper can be downloaded from https://doi.org/10.5281/zenodo.4596546 (Seiler et al., 2020). AMBER (version 1.0.3) and its documentation are also available through the Comprehensive R Archive Network (CRAN; https://CRAN.R-project.org/package=amber, last access: 28 April 2021, Seiler, 2020). The full set of Figures produced by AMBER for this study can be accessed at https://cseiler.shinyapps.io/ShinyAmber (last access: 28 April 2021, Seiler, 2021). The CLASSIC code and its software container are available at https://doi.org/10.5281/zenodo.3522407 (Melton et al., 2019).

Author contributions. CS developed AMBER, conducted the analysis, and wrote the first draft of the manuscript. JM conducted the CLASSIC experiments assessed in this study and contributed to the methods that address uncertainties in the meteorological forcing and reference data. VA created CTEM. LW processed albedo and leaf area index MODIS data. All authors contributed to the final version of the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors wish to thank all groups that provided public access to the reference data listed in Table 1. The eddy covariance data that are shared by the FLUXNET community include the AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC networks. The FLUXNET eddy covariance data processing and harmonization were 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. Colleen Mortimer compiled in situ measurements of snow water equivalent. We are grateful for the constructive comments from two anonymous reviewers.

Review statement. This paper was edited by Gerd A. Folberth and reviewed by two anonymous referees.

References

Arora, V. K. and Boer, G. J.: A Representation of Variable Root Distribution in Dynamic Vegetation Models, Earth Interact., 7, 1–19, 2003.

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

Arora, V. K. and Boer, G. J.: A parameterization of leaf phenology for the terrestrial ecosystem component of climate models, Glob. Chang. Biol., 11, 39–59, https://doi.org/10.1111/j.1365-2486.2004.00890.x, 2005b.

Arora, V. K. and Boer, G. J.: Simulating Competition and Coexistence between Plant Functional Types in a Dynamic Vegetation Model, Earth Interact., 10, 1–30, 2006.

Arora, V. K. and Boer, G. J.: Uncertainties in the 20th century carbon budget associated with land use change, Glob. Chang. Biol., 16, 3327–3348, 2010.

Arora, V. K. and Melton, J. R.: Reduction in global area burned and wildfire emissions since 1930s enhances carbon uptake by land, Nat. Commun., 9, 1326, https://doi.org/10.1038/s41467-018-03838-0, 2018.

Arora, V. K., Boer, G. J., Christian, J. R., Curry, C. L., Denman, K. L., Zahariev, K., Flato, G. M., Scinocca, J. F., Merryfield, W. J., and Lee, W. G.: The Effect of Terrestrial Photosynthesis Down Regulation on the Twentieth-Century Carbon Budget Simulated with the CCCma Earth System Model, J. Clim., 22, 6066–6088, 2009.

Asaadi, A. and Arora, V. K.: Implementation of nitrogen cycle in the CLASSIC land model, Biogeosciences, 18, 669–706, https://doi.org/10.5194/bg-18-669-2021, 2021.

Asaadi, A., Arora, V. K., Melton, J. R., and Bartlett, P.: An improved parameterization of leaf area index (LAI) seasonality in the Canadian Land Surface Scheme (CLASS) and Canadian Terrestrial Ecosystem Model (CTEM) modelling framework, Biogeosciences, 15, 6885–6907, https://doi.org/10.5194/bg-15-6885-2018, 2018.

Avitabile, V., Herold, M., Heuvelink, G. B. M., Lewis, S. L., Phillips, O. L., Asner, G. P., Armston, J., Ashton, P. S., Banin, L., Bayol, N., Berry, N. J., Boeckx, P., de Jong, B. H. J., DeVries, B., Girardin, C. A. J., Kearsley, E., Lindsell, J. A., Lopez-Gonzalez, G., Lucas, R., Malhi, Y., Morel, A., Mitchard, E. T. A., Nagy, L., Qie, L., Quinones, M. J., Ryan, C. M., Ferry, S. J. W., Sunderland, T., Laurin, G. V., Gatti, R. C., Valentini, R., Verbeeck, H., Wijaya, A., and Willcock, S.: An integrated pan-tropical biomass map using multiple reference datasets, Glob. Change Biol., 22, 1406–1420, https://doi.org/10.1111/gcb.13139, 2016.

Baccini, A., Goetz, S. J., Walker, W. S., Laporte, N. T., Sun, M., Sulla-Menashe, D., Hackler, J., Beck, P. S. A., Dubayah, R., Friedl, M. A., Samanta, S., and Houghton, R. A.: Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps, Nat. Clim. Chang., 2, 182–185, 2012.

Baret, F., Morissette, J. T., Fernandes, R. A., Champeaux, J. L., Myneni, R. B., Chen, J., Plummer, S., Weiss, M., Bacour, C., Garrigues, S., and Nickeson, J. E.: Evaluation of the representativeness of networks of sites for the global validation and intercomparison of land biophysical products: proposition of the CEOS-BELMANIP, IEEE T. Geosci. Remote Sens., 44, 1794–1803, 2006.

Bartholome, E. and Belward, A. S.: GLC2000: a new approach to global land cover mapping from Earth observation data, Int. J. Remote Sens., 26, 1959–1977, 2005.

Batjes, N. H., Ribeiro, E., van Oostrum, A., Leenaars, J., Hengl, T., and Mendes de Jesus, J.: WoSIS: providing standardised soil profile data for the world, Earth Syst. Sci. Data, 9, 1–14, https://doi.org/10.5194/essd-9-1-2017, 2017.

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. J.,

Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J. A., Stevens, L., and Vuichard, N.: The Plumbing of Land Surface Models: Benchmarking Model Performance, J. Hydrometeorol., 16, 1425–1442, 2015.

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

Brown, R. D., Brasnett, B., and Robinson, D.: Gridded North American monthly snow depth and snow water equivalent for GCM evaluation, Atmos.-Ocean, 41, 1–14, 2003.

Brun, E., Vionnet, V., Boone, A., Decharme, B., Peings, Y., Valette, R., Karbou, F., and Morin, S.: Simulation of Northern Eurasian Local Snow Depth, Mass, and Density Using a Detailed Snowpack Model and Meteorological Reanalyses, J. Hydrometeorol., 14, 203–219, 2013.

Chuvieco, E., Lizundia-Loiola, J., Pettinari, M. L., Ramo, R., Padilla, M., Tansey, K., Mouillot, F., Laurent, P., Storm, T., Heil, A., and Plummer, S.: Generation and analysis of a new global burned area product based on MODIS 250 m reflectance bands and thermal anomalies, Earth Syst. Sci. Data, 10, 2015–2031, https://doi.org/10.5194/essd-10-2015-2018, 2018.

Claverie, M., Matthews, J. L., Vermote, E. F., and Justice, C. O.: A 30+ Year AVHRR LAI and FAPAR Climate Data Record: Algorithm Description and Validation, Remote Sensing, 8, 263, https://doi.org/10.3390/rs8030263, 2016.

Collatz, G. J., Ball, J. T., Grivet, C., and Berry, J. A.: Physiological and environmental regulation of stomatal conductance, photosynthesis and transpiration: a model that includes a laminar boundary layer, Agr. Forest. Meteorol., 54, 107–136, 1991.

Collatz, G. J., Ribas-Carbo, M., and Berry, J. A.: Coupled Photosynthesis-Stomatal Conductance Model for Leaves of C4 Plants, Funct. Plant Biol., 19, 519–538, 1992.

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., and Randerson, J. T.: The International Land Model Benchmarking (ILAMB) system: design, theory, and implementation, J. Adv. Model. Earth Sy., 10, 2731–2754, 2018.

Dai, A. and Trenberth, K. E.: Estimates of Freshwater Discharge from Continents: Latitudinal and Seasonal Variations, J. Hydrometeorol., 3, 660–687, 2002.

Dickinson, R. E.: Modeling evapotranspiration for three-dimensional global climate models, Climate processes and climate sensitivity, 29, 58–72, 1984.

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

Driemel, A., Augustine, J., Behrens, K., Colle, S., Cox, C., Cuevas-Agulló, E., Denn, F. M., Duprat, T., Fukuda, M., Grobe, H., Haeffelin, M., Hodges, G., Hyett, N., Ijima, O., Kallis, A., Knap, W., Kustov, V., Long, C. N., Longenecker, D., Lupi, A., Maturilli, M., Mimouni, M., Ntsangwane, L., Ogihara, H., Olano, X., Olefs, M., Omori, M., Passamani, L., Pereira, E. B., Schmithüsen, H., Schumacher, S., Sieger, R., Tamlyn, J., Vogt, R., Vuilleumier, L., Xia, X., Ohmura, A., and König-Langlo, G.: Baseline Surface Radiation Network (BSRN): structure and data description (1992–2017), Earth Syst. Sci. Data, 10, 1491–1501, https://doi.org/10.5194/essd-10-1491-2018, 2018.

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

Farquhar, G. D., von Caemmerer, S., and Berry, J. A.: A biochemical model of photosynthetic CO2 assimilation in leaves of C_3 species, Planta, 149, 78–90, 1980.

Foley, J. A., Prentice, I. C., Ramankutty, N., Levis, S., Pollard, D., Sitch, S., and Haxeltine, A.: An integrated biosphere model of land surface processes, terrestrial carbon balance, and vegetation dynamics, Global Biogeochem. Cy., 10, 603–628, 1996.

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Hauck, J., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Bakker, D. C. E., Canadell, J. G., Ciais, P., Jackson, R. B., Anthoni, P., Barbero, L., Bastos, A., Bastrikov, V., Becker, M., Bopp, L., Buitenhuis, E., Chandra, N., Chevallier, F., Chini, L. P., Currie, K. I., Feely, R. A., Gehlen, M., Gilfillan, D., Gkritzalis, T., Goll, D. S., Gruber, N., Gutekunst, S., Harris, I., Haverd, V., Houghton, R. A., Hurtt, G., Ilyina, T., Jain, A. K., Joetzjer, E., Kaplan, J. O., Kato, E., Klein Goldewijk, K., Korsbakken, J. I., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lenton, A., Lienert, S., Lombardozzi, D., Marland, G., McGuire, P. C., Melton, J. R., Metzl, N., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S.-I., Neill, C., Omar, A. M., Ono, T., Peregon, A., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Séférian, R., Schwinger, J., Smith, N., Tans, P. P., Tian, H., Tilbrook, B., Tubiello, F. N., van der Werf, G. R., Wiltshire, A. J., and Zaehle, S.: Global Carbon Budget 2019, Earth Syst. Sci. Data, 11, 1783–1838, https://doi.org/10.5194/essd-11-1783-2019, 2019.

Garrigues, S., Lacaze, R., Baret, F., Morisette, J. T., Weiss, M., Nickeson, J. E., Fernandes, R., Plummer, S., Shabanov, N. V., Myneni, R. B., Knyazikhin, Y., and Yang, W.: Validation and intercomparison of global Leaf Area Index products derived from remote sensing data, J. Geophys. Res.-Biogeo., 113, G02028, https://doi.org/10.1029/2007JG000635, 2008.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), J. Clim., 30, 5419–5454, 2017.

Giglio, L., Loboda, T., Roy, D. P., Quayle, B., and Justice, C. O.: An active-fire based burned area mapping algorithm for the MODIS sensor, Remote Sens. Environ., 113, 408–420, 2009.

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

Goll, D. S., Brovkin, V., Parida, B. R., Reick, C. H., Kattge, J., Reich, P. B., van Bodegom, P. M., and Niinemets, Ü.: Nutrient limitation reduces land carbon uptake in simulations with a model

of combined carbon, nitrogen and phosphorus cycling, Biogeosciences, 9, 3547–3569, https://doi.org/10.5194/bg-9-3547-2012, 2012.

Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H.: Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset, Int. J. Climatol., 34, 623–642, 2014.

Henderson-Sellers, A., Yang, Z.-L., and Dickinson, R. E.: The Project for Intercomparison of Land-surface Parameterization Schemes, B. Am. Meteorol. Soc., 74, 1335–1350, 1993.

Hobeichi, S., Abramowitz, G., and Evans, J.: Conserving Land-Atmosphere Synthesis Suite (CLASS), J. Clim., 33, 1821–1844, 2019.

Iio, A. and Ito, A.: A Global Database of Field-observed Leaf Area Index in Woody Plant Species, 1932–2011, Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA, https://doi.org/10.3334/ORNLDAAC/1231, 2014.

ISIMIP: The Inter-Sectoral Impact Model Intercomparison Project, available at: https://www.isimip.org/gettingstarted/input-data-bias-correction/details/80/, last access: 29 July 2020.

Jacobson, A. R., Schuldt, K. N., Miller, J. B., Oda, T., Tans, P., Arlyn Andrews, Mund, J., Ott, L., Collatz, G. J., Aalto, T., Afshar, S., Aikin, K., Aoki, S., Apadula, F., Baier, B., Bergamaschi, P., Beyersdorf, A., Biraud, S. C., Bollenbacher, A., Bowling, D., Brailsford, G., Abshire, J. B., Chen, G., Huilin Chen, Lukasz Chmura, Sites Climadat, Colomb, A., Conil, S., Cox, A., Cristofanelli, P., Cuevas, E., Curcoll, R., Sloop, C. D., Davis, K., Wekker, S. D., Delmotte, M., DiGangi, J. P., Dlugokencky, E., Ehleringer, J., Elkins, J. W., Emmenegger, L., Fischer, M. L., Forster, G., Frumau, A., Galkowski, M., Gatti, L. V., Gloor, E., Griffis, T., Hammer, S., Haszpra, L., Hatakka, J., Heliasz, M., Hensen, A., Hermanssen, O., Hintsa, E., Holst, J., Jaffe, D., Karion, A., Kawa, S. R., Keeling, R., Keronen, P., Kolari, P., Kominkova, K., Kort, E., Krummel, P., Kubistin, D., Labuschagne, C., Langenfelds, R., Laurent, O., Laurila, T., Lauvaux, T., Law, B., Lee, J., Lehner, I., Leuenberger, M., Levin, I., Levula, J., Lin, J., Lindauer, M., Loh, Z., Lopez, M., Myhre, C. L., Machida, T., Mammarella, I., Manca, G., Manning, A., Manning, A., Marek, M. V., Marklund, P., Martin, M. Y., Matsueda, H., McKain, K., Meijer, H., Meinhardt, F., Miles, N., Miller, C. E., Mölder, M., Montzka, S., Moore, F., Josep-Anton Morgui, Morimoto, S., Munger, B., Jaroslaw Necki, Newman, S., Nichol, S., Niwa, Y., O'Doherty, S., Mikaell Ottosson-Löfvenius, Paplawsky, B., Peischl, J., Peltola, O., Jean-Marc Pichon, Piper, S., Plass-Dölmer, C., Ramonet, M., Reyes-Sanchez, E., Richardson, S., Riris, H., Ryerson, T., Saito, K., Sargent, M., Sawa, Y., Say, D., Scheeren, B., Schmidt, M., Schmidt, A., Schumacher, M., Shepson, P., Shook, M., Stanley, K., Steinbacher, M., Stephens, B., Sweeney, C., Thoning, K., Torn, M., Turnbull, J., Tørseth, K., Bulk, P. V. D., Laan-Luijkx, I. T. V. D., Dinther, D. V., Vermeulen, A., Viner, B., Vitkova, G., Walker, S., Weyrauch, D., Wofsy, S., Worthy, D., Dickon Young, and Miroslaw Zimnoch: CarbonTracker CT2019, Global Monitoring Laboratory, https://doi.org/10.25925/39M3-6069, 2020.

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

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., and Reichstein, M.: The FLUXCOM ensemble of global land-atmosphere energy fluxes, Sci. Data, 6, 74, https://doi.org/10.1038/s41597-019-0076-8, 2019.

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S., Bodesheim, P., Carvalhais, N., Chevallier, F., Gans, F., Goll, D. S., Haverd, V., Köhler, P., Ichii, K., Jain, A. K., Liu, J., Lombardozzi, D., Nabel, J. E. M. S., Nelson, J. A., O'Sullivan, M., Pallandt, M., Papale, D., Peters, W., Pongratz, J., Rödenbeck, C., Sitch, S., Tramontana, G., Walker, A., Weber, U., and Reichstein, M.: Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the FLUXCOM approach, Biogeosciences, 17, 1343–1365, https://doi.org/10.5194/bg-17-1343-2020, 2020.

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

Kato, S., Loeb, N. G., Rose, F. G., Doelling, D. R., Rutan, D. A., Caldwell, T. E., Yu, L., and Weller, R. A.: Surface Irradiances Consistent with CERES-Derived Top-of-Atmosphere Shortwave and Longwave Irradiances, J. Clim., 26, 2719–2740, 2013.

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

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, J. Meteorol. Soc. Jpn., 93, 5–48, 2015.

Koster, R. D., Sud, Y. C., Guo, Z., Dirmeyer, P. A., Bonan, G., Oleson, K. W., Chan, E., Verseghy, D., Cox, P., Davies, H., Kowalczyk, E., Gordon, C. T., Kanae, S., Lawrence, D., Liu, P., Mocko, D., Lu, C.-H., Mitchell, K., Malyshev, S., McAvaney, B., Oki, T., Yamada, T., Pitman, A., Taylor, C. M., Vasic, R., and Xue, Y.: GLACE: The Global Land–Atmosphere Coupling Experiment. Part I: Overview, J. Hydrometeorol., 7, 590–610, 2006.

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., Kampenhout, L., Kennedy, D., Kluzek, E., Lawrence, P. J., Li, F., Li, H., Lombardozzi, D., Riley, W. J., Sacks, W. J., Shi, M., Vertenstein, M., Wieder, W. R., Xu, C., Ali, A. A., Badger, A. M., Bisht, G., Broeke, M., Brunke, M. A., Burns, S. P., Buzan, J., Clark, M., Craig, A., Dahlin, K., Drewniak, B., Fisher, J. B., Flanner, M., Fox, A. M., Gentine, P., Hoffman, F., Keppel-Aleks, G., Knox, R., Kumar, S., Lenaerts, J., Leung, L. R., Lipscomb, W. H., Lu, Y., Pandey, A., Pelletier, J. D., Perket, J., Randerson, J. T., Ricciuto, D. M., Sanderson, B. M., Slater, A., Subin, Z. M., Tang, J., Thomas, R. Q., Val Martin, M., and Zeng, X.: The community land model version 5: Description of new features, benchmarking, and impact of forcing uncertainty, J. Adv. Model. Earth Syst., 11, 4245–4287, 2019.

Li, X. and Xiao, J.: Mapping Photosynthesis Solely from Solar-Induced Chlorophyll Fluorescence: A Global, Fine-Resolution Dataset of Gross Primary Production Derived from OCO-2, Remote Sensing, 11, 2563, https://doi.org/10.3390/rs11212563, 2019.

Liu, Y. Y., Parinussa, R. M., Dorigo, W. A., De Jeu, R. A. M., Wagner, W., van Dijk, A. I. J. M., McCabe, M. F., and Evans, J. P.: Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals, Hydrol. Earth Syst. Sci., 15, 425–436, https://doi.org/10.5194/hess-15-425-2011, 2011.

Manabe, S.: Climate and the ocean circulation: I. The atmospheric circulation and the hydrology of the earth's surface, Mon. Weather Rev., 97, 739–774, 1969.

Manabe, S. and Bryan, K.: Climate calculations with a combined ocean-atmosphere model, J. Atmos. Sci., 26, 786–789, 1969.

Masarie, K. A., Peters, W., Jacobson, A. R., and Tans, P. P.: ObsPack: a framework for the preparation, delivery, and attribution of atmospheric greenhouse gas measurements, Earth Syst. Sci. Data, 6, 375–384, https://doi.org/10.5194/essd-6-375-2014, 2014.

Melton, J. R. and Arora, V. K.: Competition between plant functional types in the Canadian Terrestrial Ecosystem Model (CTEM) v. 2.0, Geosci. Model Dev., 9, 323–361, https://doi.org/10.5194/gmd-9-323-2016, 2016.

Melton, J. R., Shrestha, R. K., and Arora, V. K.: The influence of soils on heterotrophic respiration exerts a strong control on net ecosystem productivity in seasonally dry Amazonian forests, Biogeosciences, 12, 1151–1168, https://doi.org/10.5194/bg-12-1151-2015, 2015.

Melton, J. R., Arora, V., Wisernig-Cojoc, E., Seiler, C., Fortier, M., Chan, E., and Teckentrup, L.: The Canadian Land Surface Scheme including Biogeochemical Cycles (Version 1.0), Zenodo, https://doi.org/10.5281/zenodo.3522407, 2019.

Melton, J. R., Arora, V. K., Wisernig-Cojoc, E., Seiler, C., Fortier, M., Chan, E., and Teckentrup, L.: 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, Geosci. Model Dev., 13, 2825–2850, https://doi.org/10.5194/gmd-13-2825-2020, 2020.

Mortimer, C., Mudryk, L., Derksen, C., Luojus, K., Brown, R., Kelly, R., and Tedesco, M.: Evaluation of long-term Northern Hemisphere snow water equivalent products, The Cryosphere, 14, 1579–1594, https://doi.org/10.5194/tc-14-1579-2020, 2020.

Mudryk, L.: Historical gridded snow water equivalent and snow cover fraction over Canada from remote sensing and land surface models, available at: http://climate-scenarios.canada.ca/?page=blended-snow-data, last access: 7 May 2020.

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

Myneni, R. B., Hoffman, S., Knyazikhin, Y., Privette, J. L., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y., Smith, G. R., Lotsch, A., Friedl, M., Morisette, J. T., Votava, P., Nemani, R. R., and Running, S. W.: Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data, Remote Sens. Environ., 83, 214–231, 2002.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, J. Geophys. Res.-Atmos., 116, D12109, https://doi.org/10.1029/2010JD015139, 2011.

Pastorello, G., Papale, D., Chu, H., Trotta, C., Agarwal, D., Canfora, E., Baldocchi, D., and Torn, M.: A new data set to keep a sharper eye on land-air exchanges, Eos T. Am. Geophys. Union, 98, https://doi.org/10.1029/2017EO071597, 2017.

Pinker, R. T. and Laszlo, I.: Modeling Surface Solar Irradiance for Satellite Applications on a Global Scale, J. Appl. Meteorol., 31, 194–211, 1992.

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

Reichstein, M., Bahn, M., Ciais, P., Frank, D., Mahecha, M. D., Seneviratne, S. I., Zscheischler, J., Beer, C., Buchmann, N., Frank, D. C., Papale, D., Rammig, A., Smith, P., Thonicke, K., van der Velde, M., Vicca, S., Walz, A., and Wattenbach, M.: Climate extremes and the carbon cycle, Nature, 500, 287–295, 2013.

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. A., Salas, W., Zutta, B. R., Buermann, W., Lewis, S. L., Hagen, S., Petrova, S., White, L., Silman, M., and Morel, A.: Benchmark map of forest carbon stocks in tropical regions across three continents, P. Natl. Acad. Sci., 108, 9899–9904, 2011.

Santoro, M., Beaudoin, A., Beer, C., Cartus, O., Fransson, J. E. S., Hall, R. J., Pathe, C., Schmullius, C., Schepaschenko, D., Shvidenko, A., Thurner, M., and Wegmüller, U.: Forest growing stock volume of the northern hemisphere: Spatially explicit estimates for 2010 derived from Envisat ASAR, Remote Sens. Environ., 168, 316–334, 2015.

Schepaschenko, D., Chave, J., Phillips, O. L., Lewis, S. L., Davies, S. J., Réjou-Méchain, M., Sist, P., Scipal, K., Perger, C., Herault, B., Labrière, N., Hofhansl, F., Affum-Baffoe, K., Aleinikov, A., Alonso, A., Amani, C., Araujo-Murakami, A., Armston, J., Arroyo, L., Ascarrunz, N., Azevedo, C., Baker, T., Bałazy, R., Bedeau, C., Berry, N., Bilous, A. M., Bilous, S. Y., Bissiengou, P., Blanc, L., Bobkova, K. S., Braslavskaya, T., Brienen, R., Burslem, D. F. R. P., Condit, R., Cuni-Sanchez, A., Danilina, D., Del Castillo Torres, D., Derroire, G., Descroix, L., Sotta, E. D., d'Oliveira, M. V. N., Dresel, C., Erwin, T., Evdokimenko, M. D., Falck, J., Feldpausch, T. R., Foli, E. G., Foster, R., Fritz, S., Garcia-Abril, A. D., Gornov, A., Gornova, M., Gothard-Bassébé, E., Gourlet-Fleury, S., Guedes, M., Hamer, K. C., Susanty, F. H., Higuchi, N., Coronado, E. N. H., Hubau, W., Hubbell, S., Ilstedt, U., Ivanov, V. V., Kanashiro, M., Karlsson, A., Karminov, V. N., Killeen, T., Koffi, J.-C. K., Konovalova, M., Kraxner, F., Krejza, J., Krisnawati, H., Krivobokov, L. V., Kuznetsov, M. A., Lakyda, I., Lakyda, P. I., Licona, J. C., Lucas, R. M., Lukina, N., Lussetti, D., Malhi, Y., Manzanera, J. A., Marimon, B., Junior, B. H. M., Martinez, R. V., Martynenko, O. V., Matsala, M., Matyashuk, R. K., Mazzei, L., Memiaghe, H., Mendoza, C., Mendoza, A. M., Moroziuk, O. V., Mukhortova, L., Musa, S., Nazimova, D. I., Okuda, T., Oliveira, L. C., Ontikov, P. V., Osipov, A. F., Pietsch, S., Playfair, M., Poulsen, J., Radchenko, V. G., Rodney, K., Rozak, A. H., Ruschel, A., Rutishauser, E., See, L., Shchepashchenko, M., Shevchenko, N., Shvidenko, A., Silveira, M., Singh, J., Sonké, B., Souza, C., Stereńczak, K., Stonozhenko, L., Sullivan, M. J. P., Szatniewska, J., Taedoumg, H., Ter Steege, H., Tikhonova, E., Toledo, M., Trefilova, O. V., Valbuena, R., Gamarra, L. V., Vasiliev, S., Vedrova, E. F., Ver-

hovets, S. V., Vidal, E., Vladimirova, N. A., Vleminckx, J., Vos, V. A., Vozmitel, F. K., Wanek, W., West, T. A. P., Woell, H., Woods, J. T., Wortel, V., Yamada, T., Nur Hajar, Z. S., and Zo-Bi, I. C.: The Forest Observation System, building a global reference dataset for remote sensing of forest biomass, Sci. Data, 6, 198, https://doi.org/10.1038/s41597-019-0196-1, 2019.

Schmid, H. P.: Source areas for scalars and scalar fluxes, Bound.-Lay. Meteorol., 67, 293–318, 1994.

Seiler, C.: AMBER: Automated Model Benchmarking R Package, r package version 1.0.3, available at: https://CRAN.R-project.org/package=amber (last access: 28 April 2021), 2020.

Seiler, C.: AMBER Results for CLASSIC v1.0, available at: https://cseiler.shinyapps.io/ShinyAmber, last access: 29 April 2021.

Seiler, C., Melton, J. R., Arora, V. K., and Wang, L.: 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 (Version 1.1), Zenodo, https://doi.org/10.5281/zenodo.4596546, 2020.

Sellers, P. J., Randall, D. A., Collatz, G. J., Berry, J. A., Field, C. B., Dazlich, D. A., Zhang, C., Collelo, G. D., and Bounoua, L.: A Revised Land Surface Parameterization (SiB2) for Atmospheric GCMS. Part I: Model Formulation, J. Clim., 9, 676–705, 1996.

Stackhouse, Jr, P. W., Gupta, S. K., Cox, S. J., Zhang, T., Mikovitz, J. C., and Hinkelman, L. M.: The NASA/GEWEX surface radiation budget release 3.0: 24.5-year dataset, Gewex news, 21, 10–12, 2011.

Strahler, A. H., Muller, J. P., and MODIS Science Team Members: MODIS BRDF/albedo product: algorithm theoretical basis document version 5.0, MODIS, available at: https://modis.gsfc.nasa.gov/data/atbd/atbd_mod09.pdf (last access: 12 April 2021), 1999.

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim, L., von Salzen, K., Yang, D., and Winter, B.: The Canadian Earth System Model version 5 (CanESM5.0.3), Geosci. Model Dev., 12, 4823–4873, https://doi.org/10.5194/gmd-12-4823-2019, 2019.

Takala, M., Luojus, K., Pulliainen, J., Derksen, C., Lemmetyinen, J., Kärnä, J.-P., Koskinen, J., and Bojkov, B.: Estimating northern hemisphere snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements, Remote Sens. Environ., 115, 3517–3529, 2011.

Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E. A. G., and Allison, S. D.: Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations, Biogeosciences, 10, 1717–1736, https://doi.org/10.5194/bg-10-1717-2013, 2013.

Viovy, N.: CRUNCEP Version 7 – Atmospheric Forcing Data for the Community Land Model, Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, https://doi.org/10.5065/PZ8F-F017, 2018.

Wang, Z., Schaaf, C. B., Sun, Q., Shuai, Y., and Román, M. O.: Capturing rapid land surface dynamics with Collection V006 MODIS BRDF/NBAR/Albedo (MCD43) products, Remote Sens. Environ., 207, 50–64, 2018.

Wieder, W.: Regridded Harmonized World Soil Database v1.2, ORNL DAAC, Oak Ridge, Tennessee, USA, https://doi.org/10.3334/ORNLDAAC/1247, 2014.

Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., Bernhofer, C., Ceulemans, R., Dolman, H., Field, C., Grelle, A., Ibrom, A., Law, B. E., Kowalski, A., Meyers, T., Moncrieff, J., Monson, R., Oechel, W., Tenhunen, J., Valentini, R., and Verma, S.: Energy balance closure at FLUXNET sites, Agr. Forest. Meteorol., 113, 223–243, 2002.

Zhang, T., Stackhouse, P. W., Cox, S. J., Mikovitz, J. C., and Long, C. N.: Clear-sky shortwave downward flux at the Earth's surface: Ground-based data vs. satellite-based data, J. Quant. Spectrosc. Ra., 224, 247–260, 2019.

Zhang, Y., Xiao, X., Wu, X., Zhou, S., Zhang, G., Qin, Y., and Dong, J.: A global moderate resolution dataset of gross primary production of vegetation for 2000–2016, Sci. Data, 4, 170165, https://doi.org/10.1038/sdata.2017.165, 2017.