

A multi-year gridded data ensemble of surface biogenic carbon fluxes for North America: evaluation and analysis of results

Yu Zhou^{1*}, Christopher A. Williams^{1*}, Thomas Lauvaux², Kenneth J. Davis², Sha Feng², Ian Baker³, Scott Denning³, Yaxing Wei⁴

¹ Graduate School of Geography, Clark University, Worcester, MA 01610, USA

² Department of Meteorology and Atmospheric Science, Pennsylvania State University, University Park, PA 16802, USA

³ Department of Atmospheric Science, Colorado State University, 1371 Campus Delivery, Fort Collins, CO 80523, USA

⁴ Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA

*Corresponding authors: Yu Zhou (yuzhou2@clarku.edu) and Christopher A. Williams (cwilliams@clarku.edu)

Key Points: [140 characters]

- We present a gridded ensemble of biogenic carbon fluxes by perturbing three key model parameters.
- Mean of the new ensemble outperforms other model products in comparison to flux tower data.
- Seasonal net carbon exchange shows consistency with other model products **in high- and mid-latitude ecoregions and biomes** and at continental scales.

Abstract [250 words]

Accurate and fine-scale estimates of biogenic carbon fluxes are critical for measuring and monitoring the biosphere's responses and feedbacks to the climate system. Currently available data products from flux towers and model-intercomparison projects struggle to adequately represent spatiotemporal dynamics of surface biogenic carbon fluxes, and to quantify their uncertainties, which also are crucial to atmospheric inversion systems. To address these gaps, we introduce a new perturbed-parameter model ensemble with the CASA model to estimate surface biogenic carbon fluxes at monthly and 3-hourly scales for North America at ~500 m and 5 km resolutions. We first use the Extended Fourier Amplitude Sensitivity Testing to choose the three most sensitive parameters to be perturbed, maximum light-use-efficiency (E_{max}), optimal temperature of photosynthesis (T_{opt}), and temperature response of respiration (Q_{10}). The initial range for each parameter is broadly sampled for the L1 ensemble, but then we pruned E_{max} with site-level primary productivity to derive an L2 ensemble with narrower uncertainty ranges. Ensembles are strongly correlated with site-level results at both monthly and 3-hourly scales and encompass the pooled AmeriFlux observations. **The monthly L2 ensemble mean achieve 85% of the observed variability. Although a less carbon net uptake is found in the cold seasons, the L2 ensemble outperforms diverse data products with the highest Taylor skill scores at diurnal to annual scales.** This ensemble presents a seasonal consistency for most biome types and in high- and mid-latitudes with other modeled data, but inconsistencies are found in tropical ecoregions and for annual totals over North America.

Keywords: parameters, sensitivity analysis, model-data comparison, biogeochemical modeling

1. Introduction

Accurate quantification of terrestrial carbon sequestration and exchange with the atmosphere over regions and continents is crucial for understanding ecosystem function, assessing biosphere-atmosphere interactions, and diagnosing terrestrial ecosystem contributions to the global coupled carbon-climate system (Schimel et al. 2001; Denman et al. 2007; Beer et al. 2010). Long-term assessment over regions is needed to characterize the magnitude, location, and trend of changes in terrestrial carbon sources/sinks, to evaluate impacts of climate dynamics, disturbance and management activities, to improve understanding of the terrestrial carbon cycle locally, and to reduce global carbon budget errors (Penman et al. 2003; IPCC 2014; West et al. 2018). More operationally, surface biogenic carbon fluxes have become a key component of monitoring, reporting, and verification (MRV) systems that seek to assess greenhouse gas (GHG) emissions and removals (Keenan and Williams 2018). Top-down approaches to MRV, that assimilate atmospheric GHG observations to adjust surface fluxes, require accurate and spatially explicit surface biogenic carbon flux priors at fine spatial (kilometric) and temporal (hourly) scales. Because atmospheric inversions adjust existing biogenic flux estimates through an optimization procedure, inverse estimates highly depend on the reliability, i.e. the uncertainties, of prior fluxes (Chevallier et al. 2006; Lauvaux et al. 2012). Such priors characterize how the biosphere responds to weather systems, seasonal climate dynamics, and climate variability and extremes. In the ideal scenario, robust flux uncertainties allow observed variations in atmospheric concentrations to adjust carbon sources and sinks (Tarantola 2005). This study's goal is to provide such a dataset for North America. It is performed with a diagnostic carbon cycle model driven by meteorological and satellite-based vegetation phenology datasets and parameterized with eddy covariance flux tower data.

Several datasets are presently available to aid in this goal. Eddy covariance flux towers measure carbon exchanges between terrestrial ecosystem and the atmosphere at (half-)hourly temporal scale, but cannot provide direct quantification of continental-scale fluxes because of the sparseness of the network (339 sites in the AmeriFlux), limitations in the spatial scale and representativeness (footprints of roughly 1 km²), relatively short and intermittent temporal sampling (mostly less than ten years), and other technical issues (e.g., low turbulence conditions such as night time/winter) (Baldocchi 2003; Aubinet 2008). Satellite remote sensing data offer spatially continuous characterizations of vegetation cover (Friedl et al. 2002), leaf area (Yang et al. 2006; Zhu et al. 2013), absorption of photosynthetically active radiation (Asrar et al. 1992; Pinker and Laszlo 1992), and solar-induced chlorophyll fluorescence (Meroni et al. 2009; Li and Xiao 2019), but do not directly observe ecosystem productivity and offer limited insights into ecosystem respiration. Integrating these data streams into a carbon cycle model can fill these gaps and provide a comprehensive description of ecosystem carbon fluxes and stocks.

Biogeochemical models are able to simulate surface carbon fluxes from site level to global scales as they are driven by climatic forcing (e.g., precipitation, temperature) and biophysical factors (e.g., vegetation indices, fraction of absorbed photosynthetically active radiation, leaf area index, soil type), and while linking aboveground and belowground processes by exchanging carbon, water, nitrogen among vegetation, soils and the atmosphere (e.g., Potter et al. 1993; Thornton et al. 2002; Krinner et al. 2005; Zeng et al. 2005). However, model intercomparison studies show wide divergence and poor skill compared to carbon flux observations, deriving from variation in model parameters, structures, and initializations (Schwalm et al. 2010; Huntzinger et al. 2012; Friedlingstein et al. 2014). Moreover, currently available products are mostly characterized with relatively low spatial resolutions which present a mismatch with the scale needed for local or regional assessment of carbon fluxes and stock.

The emerging need of monitoring surface biogenic carbon fluxes at relatively high spatial and temporal resolution is not only an important part of regional to global assessment of the carbon cycle, but also a necessary component of the regional atmospheric inversion systems (Schuh et al. 2010; Lauvaux et al. 2012). Inversion frameworks aiming to quantify surface-atmosphere carbon fluxes combine inventories of fossil fuel carbon emission and biogenic flux estimates with atmospheric measurements (e.g., Enting et al. 1995; Gurney et al. 2002; Baker et al. 2006). The biogenic fluxes simulated by biogeochemical models are often used to provide prior flux information, one of the most critical components of inversion systems. An inversion can be seen as an optimization of surface prior fluxes and hence depends primarily on the robustness of their estimates, including uncertainties. However, prior flux uncertainties often lack (a) realistic spatial and temporal structures (Lauvaux et al. 2009), and (b) a reliable estimate of the error variances (Chevallier et al. 2006). Spatial error correlations can lead to biased inverse flux estimates, and potentially provide misleading information. Currently, prior biogenic flux errors are based on expert knowledge (Enting 2002; Berchet et al. 2014), statistical criteria (Chevallier et al. 2006), or small-size ensembles (Kountouris et al. 2015). Ensemble-based approaches have been commonly used to describe uncertainties of model estimates (e.g. Schwalm et al. 2015), particularly in perturbed parameter ensembles typical of climate model analyses and deriving from its tradition in weather forecasting (Epstein 1969; Lynch 2008; Leonardo Di G et al. 2014). For surface biogenic fluxes, only a few studies have provided uncertainty estimates, typically at continental scales (Zaehle et al. 2005; Huntzinger et al. 2017). This is true in spite of several recent multi-model synthesis and intercomparison studies including the Multi-scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP,

ends at 2010; Huntzinger et al. 2013; Wei et al. 2014; Huntzinger et al. 2018), Trends in Net Land-Atmosphere Carbon Exchange (TRENDY; Sitch et al. 2008; Canadell et al. 2011) and Coupled Model Intercomparison Project (CMIP5; Taylor et al. 2012). These founding experiments were either too coarsely resolved in time or space, or excluded the characterization of error structures, or both.

In this paper, we introduce an ensemble of surface biogenic carbon fluxes for North America with relatively high spatial resolutions (~500 m and 5 km) available from 2003 to 2017 generated by perturbing model parameters. We evaluate the robustness of the ensemble with estimates of carbon fluxes derived from site-level measurements and several modeled products. We also evaluate its representativeness across various temporal scales for future local and regional carbon assessment and atmospheric inversion studies.

2. Method and Data

2.1 CASA model and temporal downscaling

The modeling approach is based on the CASA biogeochemical model (Potter et al. 1993; Randerson et al. 1996). In CASA, net primary productivity (NPP) is calculated with a light use efficiency model driven by the absorbed fraction of photosynthetically active radiation ($fPAR$) and scaled by maximum light use efficiency (E_{max}), temperature stress (T_{NPP}), and moisture stresses (W_{NPP}), all evaluated at a spatial location (x, y) and time (t) (Eq. 1). The default value of E_{max} is 0.55 for all biome types. W_{NPP} is derived based on a ratio of estimated evapotranspiration to potential evapotranspiration, and allowed to varying from 0.5 in arid conditions to 1 when fully wet. T_{NPP} is defined as $T_1 \times T_{2low} \times T_{2high}$. T_1 reflects the empirical observation that plants in very cold habitats typically have low maximum growth rate (Eq. 2). T_2 reflects the concept that the efficiency of light utilization should be depressed when plants grow at a temperature displaced from their locally acclimated temperature optimum (Eqs. 3 and 4). T_2 has an asymmetric bell shape that falls off more quickly at high than at low temperatures. The default T_{opt} is defined as the air temperature in the month when the NDVI or LAI reaches its maximum for the year.

$$NPP(x,y,t) = E_{max}(x,y) \cdot fPAR(x,y,t) \cdot PAR(x,y,t) \cdot T_{NPP}(x,y,t) \cdot W_{NPP}(x,y,t) \quad (1)$$

$$T_1 = 0.8 + (0.02 \times T_{opt}) - 0.0005 \times T_{opt}^2 \quad (2)$$

$$T_{2low} = (1 + e^{0.2 \times (T_{opt} - 10 - T)})^{-1} \quad (3)$$

$$T_{2high} = (1 + e^{0.3 \times (-10 - T_{opt} - T)})^{-1} \quad (4)$$

On a monthly time step, NPP is allocated to leaves, roots and wood, with a default fractional allocation ratio of 1:1:1. Carbon allocation are assumed to the same to aboveground and belowground woods. Each of these pools has a turnover time that specifies the rate at which carbon moves to litter pools (surface fine litter, soil fine litter, coarse woody debris). Carbon in the litterfall pool is either transferred to the microbial and soil organic matter pools or decomposed during the process (Fig. S1). The amount of carbon emitted to the atmosphere from microbial decomposition of soil and litter, i.e., heterotrophic respiration (R_h), depends on the rate and efficiency of heterotrophic consumption (Eq. 5), which varies between pools in the model and also depends on biome- and pool-specific chemistry and site-specific setting, such as soil moisture and temperature.

$$Rh(x,y,t) = \sum_{i=1}^p C_i(x,y,t) \cdot K_i(x,y) \cdot W_{resp}(x,y,t) \cdot T_{resp}(x,y,t) \cdot D_\varepsilon(x,y) \quad (5)$$

where p is the number of pools, C_i is the carbon content of pool i , K_i is the pool-specific decay rate constant, W_{resp} and T_{resp} are the effect of soil moisture and temperature on decomposition, and D_ε is microbial carbon decomposition efficiency. The effect of temperature on soil carbon fluxes (T_{resp}) is treated uniformly as an exponential (Q_{10}) response as shown in Eq. 6:

$$T_{resp}(x,y,t) = Q_{10}^{\frac{T(x,y,t) - 30}{10}} \quad (6)$$

where Q_{10} is the multiplicative increase in soil biological activity for a 10 °C rise in soil temperature and $T(x,y,t)$ is monthly averaged air temperature. The default value of Q_{10} is 1.4.

Before the simulation, an initialization of carbon pools is achieved by a 1000-year spin-up run with repeating climatology to reach an equilibrium state for each carbon pool. The CASA inputs include climatic drivers (air temperature, total precipitation, and solar radiation), vegetation type, fractional tree and herbaceous vegetation covers, biophysical properties ($fPAR$), and soil properties. The data used as input to the model are described in Section 2.2.2 and Table 1. Net ecosystem productivity (NEP) is computed as the difference between NPP and Rh. Assuming a carbon use efficiency of 0.5, gross primary productivity (GPP) is $2 \times NPP$. Correspondingly, total ecosystem respiration (RECO) is the sum of NPP and Rh, and net ecosystem exchange (NEE) is equal to RECO – GPP, and also equal to Rh – NPP.

To temporally downscale the model's monthly carbon flux estimates to 3-hourly, we used the North American Regional Reanalysis (NARR) 3-hourly (UTC) air temperature (T_{air}) and downward shortwave radiation ($DWSW$). We distributed monthly estimates to the 3-hourly temporal scale with a simple assumption of linear dependence on light for GPP (Eq. 7) and within-month, non-linear dependence on temperature for RECO (Eq. 8) (Olsen and Randerson 2004; Fisher et al. 2016a).

$$GPP(t) = GPP_{mo} \times \frac{DWSW(t)}{\sum_{i=1}^{i=8 \times days} DWSW(i)} \quad (t = 1, \dots, 8 \times days) \quad (7)$$

$$RECO(t) = RECO_{mo} \times \frac{T_{resp}(t)}{\sum_{i=1}^{i=8 \times days} T_{resp}(i)} \quad (t = 1, \dots, 8 \times days) \quad (8)$$

where T_{resp} is the temperature scalar defined in Eq. 6.

2.2. Data sources

2.2.1 Model driver data for conterminous United States and North America

The MODIS product team provide ~500m (MCD15A2H) and 1km (MCD15A2) spatial resolution $fPAR$ datasets. We used these two datasets to drive the CASA model at conterminous US and North America domains, respectively. The 1km $fPAR$ dataset ended at 2016, therefore, we used MCD15A2H to simulate carbon fluxes in 2017 for North America. Herbaceous and woody (tree) vegetation covers were obtained from MODIS product (MOD44B) rescaled from its native 250m spatial resolution. The conterminous United States (CONUS) and North America (NA) biogenic fluxes were designed to simulate at ~500m and 5km resolutions, respectively. We used spatial averaging to upscale $fPAR$ and vegetation cover datasets to match the target CONUS and NA grids.

The meteorological driver datasets differ between the two spatial domains because of their differing spatial resolutions and limits to data availability. For the CONUS domain, we used the PRISM dataset for precipitation and air temperature, and NLDAS-2 (North American Land Data Assimilation System) dataset for downward shortwave and longwave radiations, with respective spatial resolutions of 30 arcsec and 0.125 degree. The meteorological drivers for the NA domain were resampled from the NARR (North American Regional Reanalysis) dataset that has a coarser resolution (32 km). An evaluation of precipitation and air temperature of the two datasets with NCDC (National Climatic Data Center) observations is included in the Supplement (Fig. S2 and S3). We used a CONUS-Soil dataset and MsTMIP Soil Map (Liu et al. 2014) to calculate the soil fractions of clay, sand, and silt for CONUS and NA domains, separately. Three-hourly NARR data provided air temperature and downward shortwave radiation for temporal downscaling of monthly carbon fluxes for both CONUS and NA spatial domains. In order to match the target resolutions, a bilinear resampling method was adopted to downscale meteorological datasets (both monthly and three-hourly) and soil maps.

For forests, given their critical importance for carbon exchanges and diverse functional behavior, we used high-resolution forest maps, from the 250m National Forest Type map (Ruefenacht et al. 2008) and 30m NAFD (North American Forest Dynamics) products, to refine the biome types derived from middle-resolution MODIS (MOD12Q1). There were three steps to integrate these datasets. First, we generated binary maps that only presented forest and non-forest pixels from the National Forest Type map and NAFD product. We then upscaled them to the MOD12Q1 resolution (463.31m) based on the majority within the resampling window. We substituted the pixels which indicated as forest in the previous step and assigned the biome type according to the forest type in the National Forest Type map. For the NA domain, the biome type map was upscaled to 5km using the majority resampling method. The resulting biome map for CONUS is shown in Fig. S4.

[Insert Table 1 here]

2.2.2 Site-level carbon fluxes

The eddy covariance technique provides continuous measurements of NEE fluxes at (half-)hourly temporal resolutions between ecosystems and the atmosphere (Baldocchi et al. 2001). NEE is partitioned into GPP and RECO. Flux results were obtained from AmeriFlux and FLUXNET2015 datasets. In the FLUXNET2015 dataset, NEE is filtered with an ensemble of USTAR thresholds calculated with two different methods from Barr et al. (2013) and Papale et al. (2006), and applied following different strategies (threshold variable in time or constant across years). The NEE is partitioned in the two components GPP and RECO using two different methods, a nighttime-based approach to estimate RECO (Reichstein et al. 2005) and a daytime-based approach to estimate both GPP and RECO (Lasslop et al. 2010). In contrast, the AmeriFlux dataset has only a single record for each site, with a QA/QC filtered or gap-filled by the tower team. The sites are listed in Table S1. The site distribution covers a broad range of biome types (Fig. S4).

We use these site-level results in two ways. First, we use AmeriFlux NEE, GPP, and RECO to evaluate model estimates at 3-hourly and monthly scales. To aggregate (half-)hourly records to 3-hourly and monthly scales, gaps were filled by the average of each (half-)hour in a month for each year when less than 30% data was missed. Second, site-level GPP from AmeriFlux and FLUXNET2015 was used to infer each biome-specific E_{max} range in order to

generate the second-level CASA ensemble (Section 2.4). AmeriFlux GPP has the single record provided by the site PI / tower team, while FLUXNET2015 GPP embraces more uncertainties as it adopted a wide range of methods and strategies for estimating GPP. We combined AmeriFlux and FLUXNET2015 GPPs for calculating the E_{max} ranges of different biome types.

2.2.3 Datasets for model evaluation and comparison

To compare the performance of the monthly and 3-hourly CASA modeled carbon fluxes against other models, we used the flux results from three products within the overlapping period from 2006 to 2010 at flux tower sites (Table 2).

(1) CarbonTracker 2017. NOAA's CarbonTracker, version CT2017 (Peters et al. 2007, with updates documented at <http://carbontracker.noaa.gov>), provides global prior and optimized biogenic fluxes (NEE) at one-degree resolution. Prior fluxes are from two CASA runs, including Global Fire Emissions Database (GFED) 4.1s NEE (b4) and Carbon Monitoring System (CMS) NEE (bc), and both of them exclude wildfire emissions and air-sea exchange. Posterior fluxes are optimized by atmospheric observations (Peters et al. 2007). We compared our CASA NEE ensemble to the CT2017 priors and posterior, converting all units from $\text{mol m}^{-2} \text{s}^{-1}$ to $\text{g C m}^{-2} \text{s}^{-1}$.

(2) Multi-scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP) dataset. MsTMIP is a formal multi-scale synthesis involving standardized model simulations to facilitate comparison across models and with observations through an integrated evaluation framework (Huntzinger et al. 2013). We compared CASA modeled NEE to 3-hourly NEE from MsTMIP (Fisher et al. 2016b). This dataset provides global estimation with half-degree resolution. MsTMIP contains 15 modeling results included BIOME_BGC, CLM, CLM4VIC, CLASS_CTEM, DLEM, GTEC, ISAM, LPJ, ORCHIDEE, SIB3, SIBCASA, TEM6, TRIPLEX-GHG, VEGAS, and VISIT (Table S2; Huntzinger et al. 2013), optimal (weighted) and naïve (unweighted) means of the model ensemble (Schwalm et al. 2015). The naïve mean is a single-integrated mean value where each model is weighted equally. The optimal mean is the weighted mean value where each model's weight is the reliability factor in the model output (Schwalm et al. 2015). Here, we used the MsTMIP ensemble from 15 modeling results and optimal and naïve means to evaluate this study's CASA NEE ensemble at both 3-hourly and monthly scales.

(3) Hourly SiB3 dataset (SiB3CSU). In addition to the SIB3 modeling results in the MsTMIP, we also compared the CASA ensemble results with another SiB3 simulation provided on an hourly time step (Baker et al. 2008; Baker et al. 2013). In addition to the NEE results, this dataset also provides hourly GPP and RECO estimates from 2006 to 2015. We aggregated the hourly fluxes to the 3-hourly temporal resolution to evaluate ensembles of CASA NEE, GPP, and RECO.

[Insert Table 2 here]

2.3. Selecting perturbed model parameters for generating flux ensemble

We utilized Extended Fourier Amplitude Sensitivity Testing (EFAST) to determine the sensitivity of carbon fluxes to model parameters. EFAST was applied to analyze the sensitivity of carbon fluxes (NEE, GPP, and RECO) to selected parameters varied in this study (Table 3). We randomly chose 100,000 vegetated pixels to run the sensitivity test. The EFAST is a global and quantitative Sobol' algorithm that can be applied to complex nonlinear and non-monotonic models. By integrating the merits of the FAST and Sobol' algorithms, the EFAST is

characterized by high efficiency and accuracy in addition to the ability to compute the interaction effects among parameters (Saltelli et al. 1999). **Default values were taken from Potter et al. (1993) and Randerson et al. (1996). The range of each parameter was sample within the $\pm 20\%$ of its default value.**

[Insert Table 3 here]

For the parameters listed in Table 3, all the probability distribution functions (PDFs) are assumed to be uniform. The distribution of 12 parameters constitute a K^{12} parameter space. To reduce the computational complexity of searching the whole parameter space, Cukier et al. (1978) proposed to use a multidimensional Fourier decomposition with a sinusoidal function and different frequencies of each parameter:

$$x_i(s) = G_i(\sin w_i s), s \in (-\infty, +\infty) \quad (9)$$

where s is a scalar variable varying over the range $-\infty < s < +\infty$, G_i is the transformation function of searching curve which depends on distribution of parameters, and w_i are a set of frequencies associated with each parameter. Thus, how strongly a parameter's frequency propagates from input, through the model, to the output serves as a measure of model's sensitivity to the parameter. EFAST calculates the first-order and total-order sensitivity indices of each parameter based on the contribution of one parameter to the total variance of Y . The total-order sensitivity index includes higher-order, nonlinear interactions between the parameter of interest and the complementary set of parameters. We selected a final, reduced set of perturbed parameters based on those with the greatest first-order sensitivity indices. The first-order sensitivity index (S_i) of parameter i is calculated as the variance at the parameter's unique high frequency (and harmonics of that frequency) divided by total variance of model estimate (Y) (Eq. 10).

$$S_i = \frac{\text{var}_{x_i}(E(Y | x_i))}{\text{var}(Y)} \quad (10)$$

2.4 Pruning light use efficiency by site-level GPP

In the original version of CASA, all biome types use the same E_{max} in simulating primary productivity. However, many previous studies indicate that E_{max} differs among biome types (e.g., Lobell et al. 2002; Garbulsky et al. 2010), and it is also true for their uncertainties (Madani et al. 2014; Gitelson and Gamon 2015). In order to infer the range of E_{max} for each biome type, we used monthly GPP estimates (GPP_{obs}) within the growing periods at eddy covariance flux sites from AmeriFlux and FLUXNET2015 datasets (listed sites and corresponding years in Table S1). As flux sites are broadly distributed that the growing period varies across space, we defined the growing period as the time period when GPP is higher than the averaged GPP within each year. We firstly calculated maximum potential of GPP (GPP_{max}) when all of the active radiation could be utilized for photosynthesis ($2 \cdot fPAR \cdot PAR_{obs}$). GPP_{max} is only controlled by temperature (T_{NPP}) and moisture (W_{NPP}) scalars, which is shown as the the denominator in Eq. 11. Therefore, E_{max} can be calculated as the ratio of GPP_{obs} and GPP_{max} .

$$E_{max,biome} = \frac{GPP_{obs}}{GPP_{max}} = \frac{GPP_{obs}}{2 \cdot fPAR \cdot PAR_{obs} \cdot T_{NPP} \cdot W_{NPP}} \quad (11)$$

where GPP_{obs} is the monthly GPP estimates estimates sampled from the single record in the AmeriFlux dataset and the ensemble of GPP estimates in the FLUXNET2015 dataset, $fPAR$ is

derived from MOD15A2H for each flux site, and PAR_{obs} is the ground-measured at each site. Temperature and moisture scalars (T_{NPP} and W_{NPP}) were computed using ground-measured precipitation and air temperature. For some sites lacking meteorological observations, we sampled PRISM for precipitation and air temperature and NLDAS-2 for PAR.

For each site, $E_{max,biome}$ was inferred for each month within the defined growing period. Each site has an ensemble of inferred E_{max} with seasonal uncertainty of this site and uncertainties from site-level GPP estimates as they are derived from AmeriFlux and FLUXNET2015 datasets. For sites characterized as the same biome type, the E_{max} range of this biome type is represented by the mean and standard deviation of $E_{max,biome}$ ensembles.

2.5 Evaluating the CASA ensemble

To evaluate the modeled carbon fluxes with site measurements, this study used four metrics including linear correlation coefficient (r , Eq. 12), mean error (ME , Eq. 13), root mean square error ($RMSE$, Eq. 14), and Taylor diagram and skill (S_T , Eq. 15). We used the CONUS results for evaluation in Section 3, and the comparison of spatial maps of CONUS and NA are included in the Supplement. ME quantifies the averaged bias between the model estimate and observation that a positive ME represents overestimation while negative indicates the underestimation.

$$r = \frac{\sum_{i=1}^n [(C_{obs,i} - \overline{C_{obs}}) \times (C_{sim,i} - \overline{C_{sim}})]}{n \times \sigma(C_{obs}) \times \sigma(C_{sim})} \quad (12)$$

$$ME = \frac{1}{n} \sum_{i=1}^n (C_{sim,i} - C_{obs,i}) \quad (13)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (C_{sim,i} - C_{obs,i})^2}{n}} \quad (14)$$

where n is the sample size at each flux site, C_{sim} is simulated carbon fluxes, and C_{obs} is site-level flux result.

The Taylor diagram is a polar coordinate plot, characterizing the correspondence between modeling estimates and observations by exploiting the geometric and algebraic relationships with r , RMSE, and the standard deviation (σ) (Taylor 2001). This diagram compares the distance of a model from the observation, and a short distance indicates high correspondence. **For each modeled result, we calculated r , RMSE, and σ for each site, and then averaged them across sites to create the Taylor diagrams.** We also rank model performance based on Taylor skill score (S_T , Eq. 15) (Taylor 2001).

$$S_T = \frac{2 \times (1 + r)}{\left(\frac{\sigma(C_{obs})}{\sigma(C_{sim})} + \frac{\sigma(C_{sim})}{\sigma(C_{obs})} \right)^2} \quad (15)$$

We calculated S_T for diurnal cycles (3-hourly), monthly and annual variabilities to evaluate model performance across various temporal scales. Diurnal cycles are derived from deseasonalized 3-hourly data by subtracting the mean of a 15-day moving window. **We ranked**

the model performance by S_T in the Table 5, and highlighted the model (Model_H) has highest S_T . We also compared the RMSE of this model is significant from other top-four S_T models (Model_{TF}). A t -test and a F -test were adopted to test if the site-level RMSE means and variances between Model_H and Model_{TF} are significantly different ($p < 0.05$), respectively. If both not, we consider the Model_{TF} has a similarly good performance to the Model_H, which is also highlighted in the Table 5.

To assess regional carbon net uptake, we compared CASA ensemble with other model datasets at biome type level, and ecoregional and continental scales. The biome type map is based on MODIS land cover map for 18 IGBP classes (Fig. S5a), including water, evergreen needleleaf forests (ENF), evergreen broadleaf forests (EBF), deciduous needleleaf forests (DNF), deciduous broadleaf forests (DBF), mixed forests (MF), closed shrublands (CSH), open shrublands (OSH), woody savannas (WSA), savannas (SAV), grasslands (GRA), permanent wetlands, cropland (CRO), urban and built-up, cropland and natural vegetation mosaics (CNM), snow and ice, and barren as well as an unclear class. The ecoregion is derived from Level-1 map which North America is divided into 15 classes (Fig. S5b; <https://www.epa.gov/eco-research/ecoregions>).

3. Results

Figure 1 illustrates the simulation domain and spatial resolution of the CASA ensemble, with comparison to examples from alternate models. The CASA ensemble has a relatively high spatial resolution compared to other, readily available model estimates (Fig. 1). The ensemble's spatial detail derives principally from our use of MODIS $fPAR$ and vegetation cover fraction data. The ensemble spread generally spans 25% to 50% of the ensemble mean. Of the model's shown, the CASA ensemble shows closest correspondence to the CarbonTracker 2017 model priors, which is unsurprising given that those priors are developed with the same basic model and similar data inputs only at coarser spatial resolutions.

We firstly present the key parameter sets used in generating CASA Level-1 and Level-2 (L2) ensembles. We focus on evaluating CASA L2 ensemble in the main text and include CASA L1 results in the Supplement. The evaluations include comparison of CASA L2 ensemble with AmeriFlux dataset in 2003-2017, and the comparison of CASA L2 ensemble with AmeriFlux and gridded datasets in 2006-2010 across time and space.

[Insert Figure 1 here]

3.1 Selected parameter sets for generating CASA L1 and L2 ensemble

GPP is most sensitive to the change of E_{max} because of the simple linear dependence of GPP on light use efficiency in the model's equations. We find a mean first-order sensitivity to E_{max} of 0.71 in 100,000 randomly selected vegetated pixels (Fig. 2a). GPP is less sensitive to T_{opt} , with a mean first-order sensitivity of 0.29. RECO is similarly sensitive to E_{max} and T_{opt} because they determine the amount of carbon input to dead wood, litter, and soil pools for heterotrophic respiration (Fig. 2b). Q_{10} and turnover rates of heterotrophic carbon pools are less important in affecting RECO compared to E_{max} and T_{opt} . As NEE is the difference between RECO and GPP, it is also most sensitive to E_{max} , such that its mean first-order sensitivity is 0.43 (Fig. 2c), followed by T_{opt} (0.33) and Q_{10} (0.23).

[Insert Figure 2 here]

Based on these results, we selected T_{opt} , E_{max} and Q_{10} as the parameters for perturbation in generating the flux ensemble. They are the most important three parameters for controlling productivity, respiration, and net CO₂ exchange. We perturbed the optimum temperature for primary productivity. The CASA model assumes that vegetation is acclimated to experience maximal productivity at an optimal temperature equal to the temperature in the month with peak vegetation greenness, but the true optimum may lie above or below this point due to the uncertainties in the meteorological products. We assumed the optimal temperature derived from reanalysis data might vary within 2 °C, in part based on knowledge that the root mean squared difference of NLDAS-2 air temperature is 2.3 °C (Cosgrove et al. 2003). Therefore, we sampled a range of T_{opt} defined as $[Def - 2, Def + 2]$ (°C), where Def is the default optimal temperature. We perturbed the temperature sensitivity of respiration in the CASA model. Mahecha et al. (2010) reported a global convergence of Q_{10} around 1.4 ± 0.1 for most biome types from 60 FLUXNET sites. In this study we varied Q_{10} from 1.2 to 1.6. We varied the maximal, unstressed light use efficiency of primary productivity (E_{max}). We sampled a range of E_{max} spanning 0.25 to 1.00 (g C MJ⁻¹) considering the variations across different biome types (Turner et al. 2003; Wang et al. 2010). To generate the full carbon flux ensemble (Level 1, L1), we defined 36 perturbed parameter sets by uniformly sampling the space of the three selected parameters (Table S3, E_{max} in [0.25, 0.5, 0.75, 1.00], ΔT_{opt} (difference from default T_{opt}) in [-2, 0, 2], and Q_{10} in [1.2, 1.4, 1.6]).

This broad range of E_{max} was initially sampled in spite of the fact that E_{max} of some bioclimatic settings is unlikely to span the full range, such as low productivity forest groups or arid grasslands that are unlikely to reach an E_{max} of 1 gC/MJ. Pruning the L1 ensemble to the E_{max} range characteristic of each biome type is performed in the development of the second level (L2) CASA ensemble dataset. We selectively remove members of the L1 ensemble that are inconsistent with evaluation data from flux tower observations. The constrained E_{max} ranges and samples used in L2 of each biome type are listed in Table 4. E_{max} ranges of most biome types are within the initial E_{max} range except for cropland (1.01 ± 0.37). Constrained E_{max} samples selected from the original L1 perturbations have a reduced spread for select biome types. In order to represent the full uncertainty range and to have the same number of pruned ensemble members for all the biome types, the L2 E_{max} parameter values were assigned twice when the biome-specific E_{max} standard deviation is small (e.g., woody savanna), or when the E_{max} mean lies between two L1 E_{max} samples (e.g., deciduous broadleaf forest, evergreen needleleaf forest, and open shrubland). In addition, under-estimation of GPP at cropland sites led us to increase the maximum light use efficiency for this biome type to 1.25. Therefore, the number of L2 ensemble members is 27 by combining the initial samples of T_{opt} and Q_{10} (Table S3).

[Insert Table 4 here]

3.2 Evaluating the monthly CASA L2 ensemble at AmeriFlux sites

The ensemble mean L1 carbon fluxes are well correlated with flux tower estimates with a high density along the 1:1 line in scatterplots (Fig. S6, $r_{GPP} = 0.83$; $r_{RECO} = 0.77$; $r_{NEE} = 0.62$). Underestimation of high GPP (approximately > 400 g C m⁻² mo⁻¹) resulted mainly from three cropland sites, including US-Ne1, US-Ne2, US-Ne3 (Table S1). Similarly, underestimation of a large carbon sink (Fig. S6, $NEE < -200$ g C m⁻² mo⁻¹) was detected principally at cropland sites. The model-data agreement for GPP was improved in the L2 ensemble when using the pruned, biome-specific E_{max} ranges. Comparing the difference between means of L1 and L2 ensembles,

the estimates of GPP and NEE are both improved, with higher correlation coefficients (Fig. S6, $r_{GPP} = 0.88$; $r_{NEE} = 0.66$) and lower *RMSEs* (decreasing by 4.71 and 1.53 g C m⁻² mo⁻¹ for GPP and NEE, respectively).

Overall, the CASA L1 and L2 ensemble means outperform other models in estimating NEE at AmeriFlux sites (Fig. 3a). Most NEE estimates have lower variabilities than site measurements. The CASA ensemble means are more highly correlated to AmeriFlux NEE than all other models ($r < 0.5$). The *RMSEs* of CASA ensemble means are lower than other models (minimum *RMSE* is 41.3 g C m⁻² mo⁻¹ from CT2017 b4 prior and max *RMSE* is 66.4 g C m⁻² mo⁻¹ from GTEC). The CASA L2 mean presents the best correspondence with site-level measurements of NEE compared to other modeled results at the monthly scale, showing the highest Taylor Skill score ($S_T = 6.9$, Table 5a) and relatively low *RMSE* (38.3 g C m⁻² mo⁻¹, Fig. 3a). All modeled results have positive mean error (ME) compared to the AmeriFlux-measured NEE in Table 6a. The ME between the CASA L2 mean and AmeriFlux NEE is 21.0 g C m⁻² mo⁻¹, which is slightly higher than other models. This indicates that both CASA and other modeled results do not have as large a sink as the observations indicate at the monthly scale.

The CASA ensembles vary three key metabolic parameters, and all pixels of a given biome type receive the same parameter setting in a given ensemble run, which results in different flux magnitudes but fairly uniform spatiotemporal distribution across the ensemble. Therefore, the correlation coefficients of measured NEE and modeled L2 members (and thus the mean) are all close to 0.60 (Fig. 3b). Said another way, the variation in the parameters only scales the magnitude of the fluxes, not the spatio-temporal distributions. Ensemble means tend to present less variability than AmeriFlux-measured NEE, evidenced by the larger standard deviations for AmeriFlux data. However, the 11 members in the L1 (Fig. S7) and 8 members in the L2 (Fig. 3b) achieved larger standard deviation than the observed variability in NEE, and the variabilities of some ensemble members (e.g., CASA L2 #para14-15 and 19-21) are close to the variability measured at the site-level. *RMSEs* of the L2 ensemble are generally less than that of the L1 ensemble with a narrower range from 37.2 g C m⁻² mo⁻¹ to 46.5 g C m⁻² mo⁻¹.

[Insert Figure 3 here]

[Insert Table 5 here]

[Insert Table 6 here]

The CASA ensemble estimates less net uptake of carbon compared to AmeriFlux measured NEE, especially in the cold seasons (November to March) across all five biome types (Fig. 4 for sites reporting gross and net fluxes, Fig 5 for all sites). Based first on only those sites reporting gross and net fluxes (Fig. 4), we find the averaged seasonal NEE bias of the CASA L2 mean is highest in ENF (37.3 g C m⁻² mo⁻¹), followed by DBF (34.6 g C m⁻² mo⁻¹), CRO (28.9 g C m⁻² mo⁻¹), OSH (8.8 g C m⁻² mo⁻¹), and lastly GRA (4.3 g C m⁻² mo⁻¹). Similar underestimates are found when we take all available sites into consideration (Fig. 5), and allowing us to include additional biome types. In this case, the averaged seasonal underestimates are most significant in ENF, DBF, CSH, and CRO (36.8, 35.7, 25.3, and 21.9 g C m⁻² mo⁻¹, respectively), and comparably less in MF, OSH, WSA, and GRA (7.2, 9.5, 2.1, 1.6 g C m⁻² mo⁻¹, respectively). This is particularly consistent during the cold season of November to March, when the CASA L2 underestimates net carbon uptake for all the biome types.

The L2 overestimates GPP in the most productive biomes (ENF, DBF, GRA, CRO), with the largest overestimations occurring in spring and early summer. GPP biases in the cold seasons are comparably smaller in all biome types. Constrained E_{max} is able to adjust the spread of monthly GPP bias to encompass a range around 0, however, there are some significant overestimates, e.g., June and July in the GRA sites, and underestimates, e.g., July and August in the CRO sites. The overestimate of GRA net carbon uptake ($-41.4 \text{ g C m}^{-2} \text{ mo}^{-1}$) in June-August mainly results from the overestimate of GPP ($85.0 \text{ g C m}^{-2} \text{ mo}^{-1}$). In contrast, OSH sites have a relatively small GPP bias and weaker seasonal pattern than other biome types, ranging from -15.9 to $-0.8 \text{ g C m}^{-2} \text{ mo}^{-1}$. CRO sites have a significant overestimate of GPP in May ($147.6 \text{ g C m}^{-2} \text{ mo}^{-1}$) and underestimate in July-August ($-85.1 \text{ g C m}^{-2} \text{ mo}^{-1}$), leading to the corresponding overestimate and underestimate of net uptake in May ($-42.4 \text{ g C m}^{-2} \text{ mo}^{-1}$) and in July-August ($118.7 \text{ g C m}^{-2} \text{ mo}^{-1}$), respectively.

Overestimation of respiration is highest in DBF ($70.5 \text{ g C m}^{-2} \text{ mo}^{-1}$ averaged), followed by ENF ($56.4 \text{ g C m}^{-2} \text{ mo}^{-1}$), CRO ($47.4 \text{ g C m}^{-2} \text{ mo}^{-1}$), GRA ($25.8 \text{ g C m}^{-2} \text{ mo}^{-1}$), and OSH ($0.005 \text{ g C m}^{-2} \text{ mo}^{-1}$). The seasonal pattern of RECO bias is similar to that of GPP bias. In addition to the GPP biases, the general underestimate of net uptake attributes to the respiration overestimation. This is more significant in the cold seasons when GPP biases are generally insignificant (at less than $15 \text{ g C m}^{-2} \text{ mo}^{-1}$ from October to March) across the five biome types.

[Insert Figure 4 here]

[Insert Figure 5 here]

3.3 Evaluating temporally downscaled CASA L2 ensemble at AmeriFlux sites

CASA L2 – flux tower temporal correlations at the 3-hourly time scale (Fig. S8) are comparable to those found at the monthly scale for both GPP ($r_{GPP} = 0.78$) and RECO ($r_{RECO} = 0.79$). Also, the 3-hourly downscaled L2 ensemble has a comparably narrower spread and improved estimate of biogenic carbon fluxes compared to the L1 ensemble (Fig. S8). Naturally, underestimation of GPP is more significant in the 3-hourly than monthly results, most notably at croplands.

All of the models have less skill in simulating diurnal NEE variation in the winter season (DJF of Fig. 6a) than in other climatic seasons. During winter, the correlation coefficients between modeled estimates and the site-level measurement are generally lower than 0.2, and most estimates achieve less than 60% of the observed variability. We note that all of the model results were temporally downscaled in the same way from their monthly values, leading to similar temporal correlations with flux tower data. In MAM, JJA, and SON seasons, the CASA L2 mean is more strongly correlated with observed NEE compared to other models, which correlation coefficients are 0.54, 0.74, and 0.61, respectively (Fig. 6b-d). In particular, their SDs are closest to the site-level measurement in MAM and JJA, indicating the diurnal variabilities of NEE are well represented by CASA ensemble means. The L1 and L2 means account for 92.1% and 100.7% of the observed standard deviation in MAM, respectively, and in JJA, they achieve 86.9% and 97.2% of the observed standard deviations. **CASA L2 mean gains the highest Taylor skill scores in SON (6.8), and second highest in MAM (6.4) and JJA (7.5) which have similarly good performance to the highest scored models (Table 5a).** The temporal downscaling results in a less strong diurnal cycle compared to the AmeriFlux NEE, but the mean errors of each climatic season are less than $0.1 \text{ g C m}^{-2} \text{ 3hr}^{-1}$ as positive and negative biases are balanced out (Table 6a).

This is also true for diurnal GPP and RECO biases (Table 6b-c), but in terms of Taylor skill scores, the CASA L2 mean outperforms SiBCSU in representing the diurnal GPP and RECO in all climatic seasons (Table 5b-c). However, the diurnal GPP is still underestimated in the CASA L2, especially during the summer (JJA of Table 6b). Overall, the 3-hourly carbon fluxes, which derived from a simple temporal downscaling of monthly CASA ensemble, has a comparable good performance in representing diurnal cycles.

[Insert Figure 6 here]

3.4 Seasonal net carbon exchange differences at biome level in North America

In addition to evaluating CASA L2 ensemble at flux tower sites, we also compare it with other models by aggregating pixel-level NEE to the biome-scale average. CASA L2 NEE generally agrees with the seasonal patterns from other modeled products, but with stronger carbon sinks in the summer season, particularly compared to the prior estimates used in Carbon Tracker 2017 (Fig. 7). In particular, for the net carbon uptake in the summer for most biome types except for the EBF, it encompasses and agrees well with the CT2017 posterior whose estimates are nudged to improve agreement with atmospheric CO₂ observations.

However, the CASA L2 NEE presents larger carbon sources in the winter and early spring compared to other models for several biomes. This is most significant in several dominant biome types, including ENF, DBF, GRA, CRO, and cropland and natural mosaic (CNM) that account for 19.6%, 10.1%, 17.6%, 8.5%, and 3.3% of vegetated areas in the North America. This is consistent with the site-level evaluation with flux tower data which also showed less net carbon uptake in the CASA ensemble for most biome types, as a result of overestimation of respiration (Fig. 4).

[Insert Figure 7 here]

3.5 Net carbon exchange differences at ecoregional and continental levels

CASA L2 generally agrees with the seasonal patterns from other evaluation datasets in most high- and mid-latitude ecoregions (Fig. 8a–i, and l), presenting the net carbon sink in the warm seasons and source in the cold seasons. Among these ecoregions, CASA L2 presents larger temporal fluctuations in seasonal NEE than MsTMIP means, CT2017 estimates, and SiB3CSU in Marine West Coast Forests (MWCF), Northwestern Forested Mountains (NFM), Hudson Plain (HP), Great Plains (GP), Northern Forests (NF), and Eastern Temperate Forests (ETF). However, for several tropical ecoregions (Fig. 8j, l, and m–o) the CASA L2 and other model datasets present less agreement with each other in both the seasonal pattern and the magnitude of net carbon exchange, including Mediterranean California (MC), Temperate Sierras (TS), Tropical Dry Forests (TDF), South Semiarid Highlands (SSH), and Tropical Wet Forests (TWF). The CASA L2 has stronger seasonality in these ecoregions than CT2017 priors and posterior, and MsTMIP means, but their temporal patterns are similar to the variation of SiB3CSU estimates, though the time of the carbon sink maxima differ from one to three months.

[Insert Figure 8 here]

Similar to the patterns among high- and mid-latitude ecoregions, the seasonal variation of CASA NEE ensembles agree with MsTMIP means, SiB3CSU estimates, and CT posterior for North America (Fig. 9a) and conterminous US (Fig. 9b), and all of them present stronger seasonality than the CT2017 priors. Aggregating the NEE estimates to the annual scale, CASA

ensembles generally present a less strong carbon sink than other model products from 2006 to 2010, except for SiB3CSU (Table 7). NEE estimates from the CASA ensemble show a net carbon release in 2008 and 2009 for North America (CASA L2 mean, 0.23 and 0.35 Pg C yr⁻¹), and for CONUS (CASA L2, 0.01 and 0.15 Pg C yr⁻¹), while MsTMIP means and CT2017 estimates indicate an opposite net carbon sink. **The annual less net carbon uptake is also found in both CASA L2 (ME = 253.3 g C m⁻² yr⁻¹) and other evaluation datasets (ME ranges from 71.7 to 249.7 g C m⁻² yr⁻¹) at the flux tower sites compared to AmeriFlux (Table 6).** At the annual scale, SiB3CSU and some model members in the MsTMIP, e.g., GTEC and CLM, scored higher ST than the CASA ensembles, MsTMIP means, and CT2017 priors and posterior. Errors in both GPP and RECO could contribute to the general underestimation of monthly net carbon uptake, while RECO explains most of the annual bias because GPP and RECO are overestimated by 155.4 and 514.5 g C m⁻² yr⁻¹, respectively.

[Insert Figure 9 here]

[Insert Table 7 here]

4. Discussion

4.1 CASA ensemble strengths

The strength of this model ensemble firstly comes from its spatial detail, describing surface biogenic carbon fluxes at 463m for the conterminous US and 5km for North America. This product provides estimates and uncertainties of primary productivity and net carbon uptake that has such large spatial coverage with relatively fine spatial resolutions and two temporal scales (monthly and 3-hourly). The parameter varied ensemble has good performance in capturing monthly GPP variations, though it shows larger biases in representing RECO and hence NEE. We found that use of default, unconstrained parameters for simulating GPP with the CASA light-use-efficiency approach would lead to large overall uncertainty and bias, similar to findings by others (e.g. Zheng et al. 2018). By adjusting E_{max} to a biome-specific range, GPP was well adjusted year-around, with the highest Taylor skill score among the models compared here. In addition to adjusted E_{max} , satellite-derived $fPAR$ is a good indicator of productive potential, and has been widely used to provide powerful information on the spatial and temporal detail of primary productivity. It improves seasonal phasing of productivity increases and decreases (Running et al. 2000; Christian et al. 2015; Xin et al. 2015), and representing features such as agricultural planting and harvest (Bradford et al. 2005; Xin et al. 2015), deciduous leaf loss (Xiao et al. 2004), and complex foliage heterogeneity (Shabanov et al. 2003). Somewhat surprisingly, the simple temporal downscaling of monthly productivity and respiration provided similar or better skill in capturing diurnal carbon fluxes compared to models with a more detailed representation of fast processes such as 10 to 15 minute (e.g., SiB3CSU; Baker et al. 2008; Baker et al. 2013) variations in canopy physiology (canopy conductance and assimilation) as it responds to light, temperature, humidity, and soil water.

Our approach to developing this CASA ensemble involved initial, broad sampling of a wide range for key parameters, and the pruning these based on evaluation with observations from flux towers. CASA ensemble provides the uncertainty range that comes from the parameter uncertainty and encompasses a reasonable range compared with flux tower measurements (Figs. 3-5). While effective, the result is a rigid set of parameter values at regular intervals that do not optimally span the parameter space inferred from observations. Nonetheless, the approach taken

has been adequate for learning about model biases and for producing an ensemble that is tractable computationally and useful for atmospheric inversions and intercomparison with other models. An alternative approach would be to more directly sample the parameter space based on observations (e.g., Kaminski et al. 2002; Braswell et al. 2005; Nakatsuka and Maksyutov 2009; Xiao et al. 2014). The inverted parameters would be more flexible to match the observations. However, it is also more computationally intensive, and has the risk of over-fitting the model to observations (Prihodko et al. 2008; Keenan et al. 2013). Moreover, observation sites are likely to underestimate the true spread in the real world, potentially calling for widening the inferred parameter spreads obtained from limited observation sites.

4.2 Detected ensemble biases and potential solutions

E_{max} is adjusted based on the flux tower results for each biome type, and the estimated GPP in the L2 ensemble performed better than the L1 ensemble. However, the seasonal bias is still significant, presenting an overestimation at the monthly scale (Table 6), especially in summer (Fig. 4). In the cold seasons, there is a slight underestimate of monthly GPP. This is not likely due to overestimation of E_{max} given the way that E_{max} was parameterized with the same data used for model evaluation (flux tower data). Therefore, we surmise that the model's climate limitation of GPP may be too modest, principally cold and hot limitations and plant water stress. In addition to the excessively modest climatic control in simulating GPP, there is an inconsistency of seasonal patterns between satellite-derived $fPAR$ and flux tower GPP, noticed particularly at cropland sites. Flux tower GPP reaches the peak in cropland only in July and August (above $500 \text{ g C m}^{-2} \text{ mo}^{-1}$), and suddenly declines to less than $300 \text{ g C m}^{-2} \text{ mo}^{-1}$ in the early and late growing seasons and to less than $50 \text{ g C m}^{-2} \text{ mo}^{-1}$ in the fallow seasons. This could result from a mismatch of spatial scale when using satellite-derived mixed reflectance to represent site-level information when flux towers are set up downwind of a particular land cover type to measure the carbon exchange over an upwind fetch (Horst and Weil 1992; Schmid 1994), potentially leading to a mismatch between the vegetation characteristics centered around flux site coordinates and the ecological and physiological conditions that are actually sampled by the flux tower instrumentation. In addition, the satellite data product appears to be representing a mixed signal across several cropland sites in different conditions, potentially mixing across fields planted with different crop types or left fallow, planted with a cover crop, or even recently harvested, and possibly also surrounding non-crop land cover types. Also, we found that satellite-derived $fPAR$ at these cropland sites has a comparably flatter pattern than flux tower GPP for which peak values show up in July and August (~ 0.8) but gradually decrease until the fallow season (~ 0.2). As E_{max} is adjusted by growing season flux tower GPP, and $fPAR$ is linearly correlated to GPP in the light-use-efficiency model, the inconsistent seasonal patterns between flux tower GPP and $fPAR$ also contribute to the GPP overestimation in early and late growing seasons and underestimation in middle growing seasons (Fig. 4e). Allowing for seasonal variation in E_{max} as crops develop may be necessary to improve model performance in these environments.

Net ecosystem carbon uptake is underestimated by the CASA ensemble, as well as other models included here. This is not a surprising result considering that the CASA model and other models in the MstMIP ensemble are typically either set to have a mean annual balance of productivity and respiration, or run to an equilibrium carbon-pool state as an initial condition, often referred to as a “balanced biosphere”. This is true even while net uptake of atmospheric carbon by terrestrial ecosystems is widely reported (e.g. Keenan and Williams 2018). Such sinks

result from harvest removals in croplands, accumulation of live biomass stocks in forests, and soil carbon sinks in diverse ecosystem types, particularly grasslands. All of those processes are missing from the CASA ensemble, and many other biosphere models, at present (see Limitations and Challenges).

The main culprit of this underestimation was determined to be overestimation of ecosystem respiration, particularly in winter and early spring, and perhaps year-round. This bias has several candidate explanations. First and foremost, the model is missing the net sink processes noted above. Additionally, cold limitation of the decomposer community may be too weak in the CASA model, allowing too much respiration in the cold seasons. However, we note that enhancing cold-limitation of respiration while retaining a balanced biosphere condition (as noted above) would shift the annual release of respiration from winter and spring to summer and fall, but would not fix the annual bias. Another candidate explanation for overestimation of cold-season respiration is an error in the seasonal timing of carbon supply to heterotrophs from litterfall and woody debris. However, it would be unrealistic to delay litterfall and woody debris supply for the entire cold season, so this explanation does not generate a plausible fix for the observed biases.

Correcting the prevailing over-estimation of cold-season carbon sources could be achieved in the model with several mechanistic adjustments, each with unique impacts on carbon fluxes. 1) Cold-season respiration could be suppressed to a greater degree at low temperatures with the model's temperature-dependent abiotic scalar that is applied to respiration. The higher Q_{10} results in a less proportion of carbon released from dead carbon pools. This would displace cold-season carbon releases to warmer seasons, with little to no effect on total annual respiration. Notably, bottom-up and top-down approaches concluded significantly different Q_{10} ranges. Nakatsuka and Maksyutov (2009) inverted Q_{10} by vertical profiles of CO₂ concentration, which Q_{10} are greater than 1.7 for most biome types, However, the Q_{10} inferred from flux tower in Mahecha et al. (2010) has a convergence of 1.4 ± 0.1 across different biome types, which is much smaller than 1.7. As the prior estimate is crucial and largely affect the atmospheric inversion, we expect an adjusted parameter range by taking the CASA ensemble as a prior in the top-down approach. 2) Plant mortality and soil and litter carbon turnover times could be decreased to delay respiration. This would have a similar effect as above, but by delaying carbon supply to decomposers rather than by decreasing their metabolic rate. 3) Respiration could be reduced seasonally or year-round by imposing a net carbon sink in the soil. 4) Respiration could be reduced seasonally or year-round by imposing a net carbon sink in vegetation. 5) Primary productivity could be increased in the cold season, by decreasing cold-limitation of productivity in the model's temperature-dependent biotic scalar. Regarding this last mechanism we note that doing so would not alter warm-season productivity, and thus its effects on NEE would be concentrated during the cold-season with less effect on warm-season productivity. Candidates 3 and 4 appear to be most likely according to intercomparison with flux tower data presented here.

4.3 Limitations and challenges

For the CASA model and many other carbon cycle models, management activities such as age-related wood accumulation legacies in forests (Williams et al. 2014) as well as crop harvest and residue removals are not integrated into the model. In the case of forests, these CASA ensembles have wood pools that start with equilibrium, fully mature stocks, but forests regularly accumulate carbon as they recover from past disturbances, with carbon uptake

exceeding carbon release and thus a net carbon sink (Williams et al. 2016). In the case of crops, harvesting and post-harvest management of crop residues can involve direct removals off-site, but carbon cycle models typically lack this process and instead transfer material to the dead carbon pools where it is then respired via heterotrophic respiration. This overestimates respiration and therefore underestimates the net carbon sink. Imposing sinks and sources from disturbance, land use, climate, and atmospheric composition effects is difficult because knowledge is lacking in the distribution of the forcings in space and time and also on ecosystem responses. In the next level of our CASA ensemble, we intend to include fine-scale information on forest age and associated carbon sinks and also to include harvest-induced crop sinks with spatiotemporal crop patterns (e.g., Portmann et al. 2010) to improve the estimation of carbon fluxes in croplands.

The constrained E_{max} range is still relatively large (0.70 ± 0.30) for grasslands compared to other biome types. This resulted from the large span of grassland in the North America, and the eddy covariance flux sites that exhibit a wide range in their physiology and climatic settings. Therefore, a more stratified parameterization for specific ecoregions or by climate groupings may be needed for more accurate estimation of carbon fluxes. We also found that seasonal biases of CASA L2 NEE vary not only among biome types, but also within each biome type. Questions also have been raised about the ability of pre-defined biome types to capture variability in ecosystem fluxes. Another approach would be to evaluate the model at individual sites, given sufficient data, and test the whether or not the model performance is grouped according to the pre-defined biome types, allowing for the potential to re-classify the land cover types. Recently released datasets may help address these challenges. For example, a solar-induced chlorophyll fluorescence (SIF) product (Li and Xiao 2019) could provide a valuable source for a more stratified parameterization of carbon cycle models for future studies (MacBean et al. 2018), and thus might lead to more realistic complexity in the spatial structures captured within the modeling ensemble.

It is difficult to know how uncertainty scales from pixel level to landscape and regional and national levels. It is also hard to know how uncertainty is correlated in space in time, and thus how to impose uncertainties in these dimensions. Furthermore, as current available flux tower network under-represents the flux variations in space and time (Kumar et al. 2016), it is unclear how uncertainties should be translated from site-level training data (e.g. parameter constraints and tuning) to spatiotemporal fields. Technical and even theoretical advances are needed to advance all of these concerns to generate a useful and useable model ensemble beyond what has been introduced here.

5. Conclusions

The CASA ensembles introduced in this study offer a finer spatial resolution than other currently available model products, and provide the up-to-date retrospective of carbon flux anomalies at monthly and 3-hourly time scales since 2003. We identify spread in the model ensemble by perturbing the three most sensitive parameters in the model, including the maximum light use efficiency (E_{max}), the optimal temperature of photosynthesis (T_{opt}), and the temperature response of respiration (Q_{10}). In the second level of CASA ensemble, we further constrained E_{max} range by biome, leading to a better agreement between site-based and modeled carbon fluxes which correlation coefficients are 0.88, 0.82, and 0.66 for monthly GPP, RECO, and NEE, respectively, and 0.78, 0.79, and 0.68 for 3-hourly GPP, RECO, and NEE,

respectively. Each ensemble spread encompasses the pooled AmeriFlux observations, with CASA ensemble means capturing more than 70% of observed NEE variabilities at monthly and 3-hourly scales, except for in the winter. Although a seasonal bias of net carbon uptake is detected in the CASA ensemble (especially in the cold seasons), this new CASA ensemble product outperforms most other models in terms of the Taylor skill score, including members in the MsTMIP suite of models and their naïve and optimal means, CT2017 priors and posterior, and SiB3CSU at diurnal, monthly, and annual temporal scales. Regarding spatial patterns, CASA ensembles agree with the seasonal NEE variations from other evaluating datasets for most biome types, ecoregions in high- and mid-latitudes of North America, and at the continental scale, but model products show diverse behaviors in the tropical ecoregions / biomes. The amount of annual net carbon exchange aggregated from CASA gridded ensembles generally shows a smaller net carbon sink compared to that from the MsTMIP means and CT2017 priors and posterior of the conterminous US and North America.

This data product is available through the Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC) at <https://doi.org/10.3334/ORNLDAAC/1675>.

Acknowledgement

This work was primarily funded by the Atmospheric Carbon and Transport (ACT) - America project, a NASA Earth Venture Suborbital 2 project supported by NASA's Earth Science Division. Funding for this work came from the NASA ACT-America Project under awards #NNX16AN17G and #NNX15AG76G. This work used eddy covariance data acquired and shared by the FLUXNET community, including AmeriFlux and Fluxnet-Canada. Funding for AmeriFlux data resources was provided by the U.S. Department of Energy's Office of Science. CarbonTracker (CT2017) results provided by NOAA ESRL, Boulder, Colorado, USA from the website at <http://carbontracker.noaa.gov>. The three-hourly output from Multi-scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP; <http://nacp.ornl.gov/MsTMIP.shtml>) can be found the Modeling and Synthesis Thematic Data Center at Oak Ridge National Laboratory (ORNL; <http://nacp.ornl.gov>). We thank Klaus Keller at the Pennsylvania State University for the helpful suggestions on the model-data comparison. We thank the anonymous reviewers and the associate editor for the constructive comments on the manuscript.

References

- Asrar, G., Myneni, R., & Choudhury, B. 1992. Spatial heterogeneity in vegetation canopies and remote sensing of absorbed photosynthetically active radiation: a modeling study. *Remote Sensing of Environment*, 41, 85-103
- Aubinet, M. 2008. Eddy covariance CO₂ flux measurements in nocturnal conditions: an analysis of the problem. *Ecological Applications*, 18, 1368-1378
- Baker, D., Law, R.M., Gurney, K., Rayner, P., Peylin, P., Denning, A., Bousquet, P., Bruhwiler, L., Chen, Y.H., & Ciais, P. 2006. TransCom 3 inversion intercomparison: Impact of transport model errors on the interannual variability of regional CO₂ fluxes, 1988–2003. *Global Biogeochemical Cycles*, 20
- Baker, I., Harper, A., da Rocha, H., Denning, A., Araújo, A., Borma, L., Freitas, H., Goulden, M., Manzi, A., & Miller, S. 2013. Surface ecophysiological behavior across vegetation and moisture gradients in tropical South America. *Agricultural and Forest Meteorology*, 182, 177-188
- Baker, I., Prihodko, L., Denning, A., Goulden, M., Miller, S., & Da Rocha, H. 2008. Seasonal drought stress in the Amazon: Reconciling models and observations. *Journal of Geophysical Research: Biogeosciences*, 113
- Baldocchi, D.D. 2003. Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change Biology*, 9, 479-492
- Barr, A., Richardson, A., Hollinger, D., Papale, D., Arain, M., Black, T., Bohrer, G., Dragoni, D., Fischer, M., & Gu, L. 2013. Use of change-point detection for friction–velocity threshold evaluation in eddy-covariance studies. *Agricultural and Forest Meteorology*, 171, 31-45
- Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M.A., Baldocchi, D., & Bonan, G.B. 2010. Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. *Science*, 329, 834-838
- Berchet, A., Pison, I., Chevallier, F., Bousquet, P., Bonne, J.-L., & Paris, J.-D. 2014. Objectified quantification of uncertainties in Bayesian atmospheric inversions. *Geoscientific Model Development Discussions*, 7, 4777-4827
- Bradford, J., Hicke, J., & Lauenroth, W. 2005. The relative importance of light-use efficiency modifications from environmental conditions and cultivation for estimation of large-scale net primary productivity. *Remote Sensing of Environment*, 96, 246-255
- Braswell, B.H., Sacks, W.J., Linder, E., & Schimel, D.S. 2005. Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. *Global change biology*, 11, 335-355
- Canadell, J.G., Ciais, P., Gurney, K., Le Quéré, C., Piao, S., Raupach, M.R., & Sabine, C.L. 2011. An international effort to quantify regional carbon fluxes. *Eos, Transactions American Geophysical Union*, 92, 81-82
- Chevallier, F., Viovy, N., Reichstein, M., & Ciais, P. 2006. On the assignment of prior errors in Bayesian inversions of CO₂ surface fluxes. *Geophysical research letters*, 33
- Christian, B., Joshi, N., Saini, M., Mehta, N., Goroshi, S., Nidamanuri, R.R., Thenkabail, P., Desai, A.R., & Krishnappa, N. 2015. Seasonal variations in phenology and productivity of a tropical dry deciduous forest from MODIS and Hyperion. *Agricultural and Forest Meteorology*, 214, 91-105
- Cosgrove, B.A., Lohmann, D., Mitchell, K.E., Houser, P.R., Wood, E.F., Schaake, J.C., Robock, A., Marshall, C., Sheffield, J., & Duan, Q. 2003. Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *Journal of Geophysical Research: Atmospheres*, 108
- Cukier, R., Levine, H., & Shuler, K. 1978. Nonlinear sensitivity analysis of multiparameter model systems. *Journal of computational physics*, 26, 1-42
- Denman, K.L., Brasseur, G., Chidthaisong, A., Ciais, P., Cox, P.M., Dickinson, R.E., Hauglustaine, D., Heinze, C., Holland, E., Jacob, D., Lohmann, U., Ramachandran, S., da Silva Dias, P.L., Wofsy, S.C., & Zhang, X. 2007. Couplings between changes in the climate system and biogeochemistry. In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen,

M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Dimiceli, C., Carroll, M., Sohlberg, R., Kim, D.H., Kelly, M., & Townshend, J.R.G. 2015. MOD44B MODIS/Terra Vegetation Continuous Fields Yearly L3 Global 250m SIN Grid V006. NASA EOSDIS Land Processes DAAC. Available online: https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod44b_v006

Enting, I., Trudinger, C., & Francey, R. 1995. A synthesis inversion of the concentration and $\delta^{13}C$ of atmospheric CO₂. *Tellus B*, 47, 35-52

Enting, I.G. 2002. *Inverse problems in atmospheric constituent transport*. Cambridge University Press

Epstein, E.S. 1969. Stochastic dynamic prediction. *Tellus*, 21, 739-759

Fisher, J.B., Sikka, M., Huntzinger, D.N., Schwalm, C., & Liu, J. 2016a. 3-hourly temporal downscaling of monthly global terrestrial biosphere model net ecosystem exchange. *Biogeosciences*, 13, 4271-4277

Fisher, J.B., Sikka, M., Huntzinger, D.N., Schwalm, C.R., Liu, J., Wei, Y., Cook, R.B., Michalak, A.M., Schaefer, K., Jacobson, A.R., Arain, M.A., Ciais, P., El-Masri, B., Hayes, D.J., Huang, M., et al. 2016b. CMS: Modeled Net Ecosystem Exchange at 3-hourly Time Steps, 2004-2010. ORNL DAAC, Oak Ridge, Tennessee, USA. <https://doi.org/10.3334/ORNLDAAC/1315>

Friedl, M.A., McIver, D.K., Hodges, J.C., Zhang, X.Y., Muchoney, D., Strahler, A.H., Woodcock, C.E., Gopal, S., Schneider, A., & Cooper, A. 2002. Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of Environment*, 83, 287-302

Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., & Huang, X. 2010. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, 114, 168-182

Friedlingstein, P., Meinshausen, M., Arora, V.K., Jones, C.D., Anav, A., Liddicoat, S.K., & Knutti, R. 2014. Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *Journal of Climate*, 27, 511-526

Garbulsky, M.F., Peñuelas, J., Papale, D., Ardö, J., Goulden, M.L., Kiely, G., Richardson, A.D., Rotenberg, E., Veenendaal, E.M., & Filella, I. 2010. Patterns and controls of the variability of radiation use efficiency and primary productivity across terrestrial ecosystems. *Global Ecology and Biogeography*, 19, 253-267

Gitelson, A.A., & Gamon, J.A. 2015. The need for a common basis for defining light-use efficiency: Implications for productivity estimation. *Remote Sensing of Environment*, 156, 196-201

Goward, S.N., Huang, C., Masek, J.G., Cohen, W.B., Moisen, G.G., & Schleeuwis, K. 2012. NACP North American Forest Dynamics Project: Forest Disturbance and Regrowth Data. ORNL DAAC, Oak Ridge, Tennessee, USA. <http://dx.doi.org/10.3334/ORNLDAAC/1077>

Gurney, K.R., Law, R.M., Denning, A.S., Rayner, P.J., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., & Fan, S. 2002. Towards robust regional estimates of CO₂ sources and sinks using atmospheric transport models. *Nature*, 415, 626

Horst, T., & Weil, J. 1992. Footprint estimation for scalar flux measurements in the atmospheric surface layer. *Boundary-Layer Meteorology*, 59, 279-296

Huntzinger, D., Schwalm, C., Michalak, A., Schaefer, K., King, A., Wei, Y., Jacobson, A., Liu, S., Cook, R., & Post, W. 2013. The north american carbon program multi-scale synthesis and terrestrial model intercomparison project—part 1: Overview and experimental design. *Geoscientific Model Development*, 6, 2121-2133

Huntzinger, D., Schwalm, C., Wei, Y., Cook, R., Michalak, A., Schaefer, K., Jacobson, A., Arain, M., Ciais, P., Fisher, J., Hayes, D.J., Huang, M., Huang, S., Ito, A., Jain, A.K., et al. 2018. NACP MsTMIP: Global 0.5-deg Terrestrial Biosphere Model Outputs (version 1) in Standard Format. ORNL DAAC, Oak Ridge, Tennessee, USA. <https://doi.org/10.3334/ORNLDAAC/1225>.

Huntzinger, D.N., Michalak, A., Schwalm, C., Ciais, P., King, A., Fang, Y., Schaefer, K., Wei, Y., Cook, R., & Fisher, J. 2017. Uncertainty in the response of terrestrial carbon sink to environmental drivers undermines carbon-climate feedback predictions. *Scientific reports*, 7, 4765

Huntzinger, D.N., Post, W.M., Wei, Y., Michalak, A., West, T.O., Jacobson, A., Baker, I., Chen, J.M., Davis, K., & Hayes, D. 2012. North American Carbon Program (NACP) regional interim synthesis: Terrestrial biospheric model intercomparison. *Ecological Modelling*, 232, 144-157

IPCC 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp.

Kaminski, T., Knorr, W., Rayner, P.J., & Heimann, M. 2002. Assimilating atmospheric data into a terrestrial biosphere model: A case study of the seasonal cycle. *Global Biogeochemical Cycles*, 16, 14-11-14-16

Keenan, T.F., Davidson, E.A., Munger, J.W., & Richardson, A.D. 2013. Rate my data: quantifying the value of ecological data for the development of models of the terrestrial carbon cycle. *Ecol Appl*, 23, 273-286

Keenan, T.F., & Williams, C.A. 2018. The Terrestrial Carbon Sink. *Annual Review of Environment and Resources*, 43, 219-243

Kountouris, P., Gerbig, C., Totsche, K.-U., Dolman, A.-J., Meesters, A.-G.-C.-A., Broquet, G., Maignan, F., Gioli, B., Montagnani, L., & Helfter, C. 2015. An objective prior error quantification for regional atmospheric inverse applications. *Biogeosciences*, 12, 7403-7421

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S., & Prentice, I.C. 2005. A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, 19

Kumar, J., Hoffman, F.M., Hargrove, W.W., & Collier, N. 2016. Understanding the representativeness of FLUXNET for upscaling carbon flux from eddy covariance measurements. In: Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States)

Lasslop, G., Reichstein, M., Papale, D., Richardson, A.D., Arneth, A., Barr, A., Stoy, P., & Wohlfahrt, G. 2010. Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. *Global change biology*, 16, 187-208

Lauvaux, T., Pannekoucke, O., Sarrat, C., Chevallier, F., Ciais, P., Noilhan, J., & Rayner, P. 2009. Structure of the transport uncertainty in mesoscale inversions of CO₂ sources and sinks using ensemble model simulations. *Biogeosciences*, 6, 1089-1102

Lauvaux, T., Schuh, A., Uliasz, M., Richardson, S., Miles, N., Andrews, A., Sweeney, C., Diaz, L., Martins, D., & Shepson, P. 2012. Constraining the CO₂ budget of the corn belt: exploring uncertainties from the assumptions in a mesoscale inverse system. *Atmospheric Chemistry & Physics*, 12

LDAS 2016. NLDAS-2 Forcing Dataset, Land Data Assimilation Systems, at <https://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing.php>

Leonardo Di G, S., Sira, E., Klapp, J., & Trujillo, L. 2014. Environmental fluid mechanics: Applications to weather forecast and climate change. *Computational and Experimental Fluid Mechanics With Applications to Physics, Engineering and the Environment* (pp. 3-36): Springer

Li, X., & Xiao, J. 2019. A Global, 0.05-Degree Product of Solar-Induced Chlorophyll Fluorescence Derived from OCO-2, MODIS, and Reanalysis Data. *Remote Sensing*, 11, 517

Liu, S., Wei, Y., Post, W., Cook, R., Schaefer, K., & Thornton, M. 2014. NACP MsTMIP: Unified North American soil map. Available online: <http://dx.doi.org/10.3334/ORNLDAAAC/1242>, Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA (accessed on 27 Jan 2017)

Lobell, D., Hicke, J., Asner, G., Field, C., Tucker, C., & Los, S. 2002. Satellite estimates of productivity and light use efficiency in United States agriculture, 1982–98. *Global change biology*, 8, 722-735

Lynch, P. 2008. The origins of computer weather prediction and climate modeling. *Journal of computational physics*, 227, 3431-3444

MacBean, N., Maignan, F., Bacour, C., Lewis, P., Peylin, P., Guanter, L., Köhler, P., Gómez-Dans, J., & Disney, M. 2018. Strong constraint on modelled global carbon uptake using solar-induced chlorophyll fluorescence data. *Scientific reports*, 8

- Madani, N., Kimball, J.S., Affleck, D.L.R., Kattge, J., Graham, J., van Bodegom, P.M., Reich, P.B., & Running, S.W. 2014. Improving ecosystem productivity modeling through spatially explicit estimation of optimal light use efficiency. *Journal of Geophysical Research: Biogeosciences*, 119, 1755-1769
- Mahecha, M.D., Reichstein, M., Carvalhais, N., Lasslop, G., Lange, H., Seneviratne, S.I., Vargas, R., Ammann, C., Arain, M.A., & Cescatti, A. 2010. Global convergence in the temperature sensitivity of respiration at ecosystem level. *Science*, 329, 838-840
- Meroni, M., Rossini, M., Guanter, L., Alonso, L., Rascher, U., Colombo, R., & Moreno, J. 2009. Remote sensing of solar-induced chlorophyll fluorescence: Review of methods and applications. *Remote Sensing of Environment*, 113, 2037-2051
- Miller, D.A., & White, R.A. 1998. A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling. Available online: <http://EarthInteractions.org>. *Earth interactions*, 2, 1-26
- Myneni, R., Knyazikhin, Y., & Park, T. 2015. MCD15A2H MODIS/Terra+Aqua Leaf Area Index/FPAR 8-day L4 Global 500m SIN Grid V006. NASA EOSDIS Land Processes DAAC. Available online: https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mcd15a2h_v006
- Myneni, R.B., Hoffman, S., Knyazikhin, Y., Privette, J., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y., & Smith, G. 2002. Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sensing of Environment*, 83, 214-231
- Nakatsuka, Y., & Maksyutov, S. 2009. Optimization of the seasonal cycles of simulated CO₂ flux by fitting simulated atmospheric CO₂ to observed vertical profiles. *Biogeosciences*, 6, 2733-2741
- NCEP 2005. National Centers for Environmental Prediction (NCEP) North American Regional Reanalysis (NARR). In. Boulder, CO: Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory
- Olsen, S.C., & Randerson, J.T. 2004. Differences between surface and column atmospheric CO₂ and implications for carbon cycle research. *Journal of Geophysical Research: Atmospheres*, 109
- Papale, D., Reichstein, M., Aubinet, M., Canfora, E., Bernhofer, C., Kutsch, W., Longdoz, B., Rambal, S., Valentini, R., & Vesala, T. 2006. Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: algorithms and uncertainty estimation. *Biogeosciences*, 3, 571-583
- Penman, J., Gytarsky, M., Hiraishi, T., Krug, T., Kruger, D., Pipatti, R., Buendia, L., Miwa, K., Ngara, T., Tanabe, K., & Wagner, F. 2003. *Good practice guidance for land use, land-use change and forestry*. Kanagawa Prefecture: Institute for Global Environmental Strategies
- Peters, W., Jacobson, A.R., Sweeney, C., Andrews, A.E., Conway, T.J., Masarie, K., Miller, J.B., Bruhwiler, L.M., Pétron, G., & Hirsch, A.I. 2007. An atmospheric perspective on North American carbon dioxide exchange: CarbonTracker. *Proceedings of the National Academy of Sciences*, 104, 18925-18930
- Pinker, R., & Laszlo, I. 1992. Global distribution of photosynthetically active radiation as observed from satellites. *Journal of Climate*, 5, 56-65
- Portmann, F.T., Siebert, S., & Döll, P. 2010. MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochemical Cycles*, 24
- Potter, C.S., Randerson, J.T., Field, C.B., Matson, P.A., Vitousek, P.M., Mooney, H.A., & Klooster, S.A. 1993. Terrestrial ecosystem production: a process model based on global satellite and surface data. *Global Biogeochemical Cycles*, 7, 811-841
- Prihodko, L., Denning, A., Hanan, N., Baker, I., & Davis, K. 2008. Sensitivity, uncertainty and time dependence of parameters in a complex land surface model. *Agricultural and Forest Meteorology*, 148, 268-287
- PRISM Climate Group 2016. PRISM Gridded Climate Data, Oregon State University, <http://prism.oregonstate.edu>
- Randerson, J.T., Thompson, M.V., Malmstrom, C.M., Field, C.B., & Fung, I.Y. 1996. Substrate limitations for heterotrophs: Implications for models that estimate the seasonal cycle of atmospheric CO₂. *Global Biogeochemical Cycles*, 10, 585-602

- Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., & Granier, A. 2005. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global Change Biology*, *11*, 1424-1439
- Ruefenacht, B., Finco, M., Nelson, M., Czaplewski, R., Helmer, E., Blackard, J., Holden, G., Lister, A., Salajanu, D., & Weyermann, D. 2008. Conterminous US and Alaska forest type mapping using forest inventory and analysis data. *Photogrammetric Engineering & Remote Sensing*, *74*, 1379-1388
- Running, S.W., Thornton, P.E., Nemani, R., & Glassy, J.M. 2000. Global terrestrial gross and net primary productivity from the Earth Observing System. *Methods in ecosystem science* (pp. 44-57): Springer
- Saltelli, A., Tarantola, S., & Chan, K.-S. 1999. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, *41*, 39-56
- Schimel, D.S., House, J.I., Hibbard, K.A., Bousquet, P., Ciais, P., Peylin, P., Braswell, B.H., Apps, M.J., Baker, D., Bondeau, A., Canadell, J., Churkina, G., Cramer, W., Denning, A.S., Field, C.B., et al. 2001. Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. *Nature*, *414*, 169-172
- Schmid, H. 1994. Source areas for scalars and scalar fluxes. *Boundary-Layer Meteorology*, *67*, 293-318
- Schuh, A., Denning, A., Corbin, K., Baker, I., Uliasz, M., Parazoo, N., Andrews, A., & Worthy, D. 2010. A regional high-resolution carbon flux inversion of North America for 2004. *Biogeosciences*, *7*, 1625-1644
- Schwalm, C.R., Huntzinger, D.N., Fisher, J.B., Michalak, A.M., Bowman, K., Ciais, P., Cook, R., El-Masri, B., Hayes, D., & Huang, M. 2015. Toward “optimal” integration of terrestrial biosphere models. *Geophysical research letters*, *42*, 4418-4428
- Schwalm, C.R., Williams, C.A., Schaefer, K., Anderson, R., Arain, M.A., Baker, I., Barr, A., Black, T.A., Chen, G., Chen, J.M., Ciais, P., Davis, K.J., Desai, A., Dietze, M., Dragoni, D., et al. 2010. A model-data intercomparison of CO₂ exchange across North America: Results from the North American Carbon Program site synthesis. *Journal of Geophysical Research*, *115*
- Shabanov, N., Wang, Y., Buermann, W., Dong, J., Hoffman, S., Smith, G., Tian, Y., Knyazikhin, Y., & Myneni, R. 2003. Effect of foliage spatial heterogeneity in the MODIS LAI and FPAR algorithm over broadleaf forests. *Remote Sensing of Environment*, *85*, 410-423
- Sitch, S., Huntingford, C., Gedney, N., Levy, P., Lomas, M., Piao, S., Betts, R., Ciais, P., Cox, P., & Friedlingstein, P. 2008. Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five Dynamic Global Vegetation Models (DGVMs). *Global change biology*, *14*, 2015-2039
- Tarantola, A. 2005. Inverse problem theory and methods for model parameter estimation. Society for Industrial and Applied Mathematics, Philadelphia, PA.
- Taylor, K.E. 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*, 7183-7192
- Taylor, K.E., Stouffer, R.J., & Meehl, G.A. 2012. An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*, 485-498
- Thornton, P.E., Law, B.E., Gholz, H.L., Clark, K.L., Falge, E., Ellsworth, D.S., Goldstein, A., Monson, R.K., Hollinger, D., & Falk, M. 2002. Modeling and measuring the effects of disturbance history and climate on carbon and water budgets in evergreen needleleaf forests. *Agricultural and Forest Meteorology*, *113*, 185-222
- Turner, D.P., Urbanski, S., Bremer, D., Wofsy, S.C., Meyers, T., Gower, S.T., & Gregory, M. 2003. A cross-biome comparison of daily light use efficiency for gross primary production. *Global change biology*, *9*, 383-395
- Wang, H., Jia, G., Fu, C., Feng, J., Zhao, T., & Ma, Z. 2010. Deriving maximal light use efficiency from coordinated flux measurements and satellite data for regional gross primary production modeling. *Remote Sensing of Environment*, *114*, 2248-2258
- Wei, Y., Liu, S., Huntzinger, D.N., Michalak, A.M., Viogy, N., Post, W.M., Schwalm, C.R., Schaefer, K., Jacobson, A.R., & Lu, C. 2014. The North American carbon program multi-scale synthesis and terrestrial

model intercomparison project–Part 2: environmental driver data. *Geoscientific Model Development Discussions*, 7, 2875-2893

West, T.O., Gurwick, N.P., Brown, M.E., Duren, R., Mooney, S., Paustian, K., McGlynn, E., Malone, E.L., Rosenblatt, A., Hultman, N., & Ocko, I.B. 2018. Chapter 18: Carbon cycle science in support of decision making. In Second State of the Carbon Cycle Report (SOCCR2): A Sustained Assessment Report [Cavallaro, N., G. Shrestha, R. Birdsey, M. A. Mayes, R. G. Najjar, S. C. Reed, P. RomeroLankao, and Z. Zhu (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, pp. 728-759, <https://doi.org/10.7930/SOCCR2.2018.Ch18>.

Williams, C.A., Gu, H., MacLean, R., Masek, J.G., & Collatz, G.J. 2016. Disturbance and the carbon balance of US forests: A quantitative review of impacts from harvests, fires, insects, and droughts. *Global and Planetary Change*, 143, 66-80

Williams, C.A., Vanderhoof, M.K., Khomik, M., & Ghimire, B. 2014. Post-clearcut dynamics of carbon, water and energy exchanges in a midlatitude temperate, deciduous broadleaf forest environment. *Glob Chang Biol*, 20, 992-1007

Xiao, J., Davis, K.J., Urban, N.M., & Keller, K. 2014. Uncertainty in model parameters and regional carbon fluxes: A model-data fusion approach. *Agricultural and Forest Meteorology*, 189, 175-186

Xiao, X., Zhang, Q., Braswell, B., Urbanski, S., Boles, S., Wofsy, S., Moore III, B., & Ojima, D. 2004. Modeling gross primary production of temperate deciduous broadleaf forest using satellite images and climate data. *Remote Sensing of Environment*, 91, 256-270

Xin, Q., Broich, M., Suyker, A.E., Yu, L., & Gong, P. 2015. Multi-scale evaluation of light use efficiency in MODIS gross primary productivity for croplands in the Midwestern United States. *Agricultural and Forest Meteorology*, 201, 111-119

Yang, W., Tan, B., Huang, D., Rautiainen, M., Shabanov, N.V., Wang, Y., Privette, J.L., Huemmrich, K.F., Fensholt, R., & Sandholt, I. 2006. MODIS leaf area index products: From validation to algorithm improvement. *IEEE Transactions on Geoscience and Remote Sensing*, 44, 1885-1898

Zaehle, S., Sitch, S., Smith, B., & Hatterman, F. 2005. Effects of parameter uncertainties on the modeling of terrestrial biosphere dynamics. *Global Biogeochemical Cycles*, 19

Zeng, N., Qian, H., Roedenbeck, C., & Heimann, M. 2005. Impact of 1998–2002 midlatitude drought and warming on terrestrial ecosystem and the global carbon cycle. *Geophysical research letters*, 32

Zheng, Y., Zhang, L., Xiao, J., Yuan, W., Yan, M., Li, T., & Zhang, Z. 2018. Sources of uncertainty in gross primary productivity simulated by light use efficiency models: Model structure, parameters, input data, and spatial resolution. *Agricultural and Forest Meteorology*, 263, 242-257

Zhu, Z., Bi, J., Pan, Y., Ganguly, S., Anav, A., Xu, L., Samanta, A., Piao, S., Nemani, R., & Myneni, R. 2013. Global data sets of vegetation leaf area index (LAI) 3g and fraction of photosynthetically active radiation (FPAR) 3g derived from global inventory modeling and mapping studies (GIMMS) normalized difference vegetation index (NDVI3g) for the period 1981 to 2011. *Remote Sensing*, 5, 927-948

Tables

Table 1. Driven datasets used for simulations at (a) conterminous US and (b) North America.

Model input	Dataset	Spatial resolution	Temporal resolution	Reference
(a) Conterminous US				
f PAR	MCD15A2H	463.31 m	8-day	Myneni et al. (2015)
Tree and herb covers	MOD44B	250 m	Yearly	Dimiceli et al. (2015)
Precipitation and T_{air}	PRISM	30 "	Monthly	PRISM Climate Group (2016)
DWSW and DWLW ¹	NDLAS -2 Forcing	0.125 °	Monthly	LDAS (2016)
DWSW ¹ and T_{air}	NARR	32 km	3hourly	NCEP (2005)
Biome type	National Forest Type	250 m	--	Ruefenacht et al. (2008)
	NAFD	30 m	--	Goward et al. (2012)
	MOD12Q1 IGBP	463.31 m	Yearly	Friedl et al. (2010)
Clay, silt, sand Fractions	CONUS-Soil	1000 m	--	Miller and White (1998)
(b) North America				
f PAR	MCD15A2	1000 m	8-day	Myneni et al. (2002)
Tree and herb covers	MOD44B	250 m	Yearly	Dimiceli et al. (2015)
Precipitation, T_{air} , DWSW, and DWLW ¹	NARR	32 km	Monthly	NCEP (2005)
DWSW ¹ and T_{air}	NARR	32 km	3hourly	NCEP (2005)
Biome type	National Forest Type	250 m	--	Ruefenacht et al. (2008)
	NAFD	30 m	--	Goward et al. (2012)
	MOD12Q1 IGBP	463.31 m	Yearly	Friedl et al. (2010)
Clay, silt, and sand fractions	NACP MsTMIP Soil Map	0.25 °	--	Liu et al. (2014)

Table 2. Datasets of CASA ensemble and datasets for comparison and evaluating.

Dataset	Time span	Components	Spatial Resolution	Temporal Resolution
CASA ensemble	2003-2017	NEE, GPP, RECO	500 m	3-hourly, Monthly
AmeriFlux	Site-specific	NEE, GPP, RECO	Flux tower	(Half-)hourly
SiB3CSU	2006-2015	NEE, GPP, RECO	1° × 1.25°	Hourly
CT2017	2000-2017	NEE	1° × 1°	3-hourly, Monthly
MsTMIP	2004-2010	NEE	0.5° × 0.5°	3-hourly, Monthly

Table 3. Description, default value and range of each parameter used in EFAST.

Parameter	Description	Default
E_{max}	Maximum light use efficiency	0.55
T_{opt}	Optimal temperature of photosynthesis	<i>Def</i>
Q_{10}	Parameter driving the exponential dependency of the heterotrophic respiration on temperature	1.40
$K_{surfmet}$	Turnover rate of surface metabolic carbon pool	14.80
$K_{surfstr}$	Turnover rate of surface structural carbon pool	3.90
$K_{surfmic}$	Turnover rate of surface microbial carbon pool	6.00
$K_{soilmet}$	Turnover rate of soil metabolic carbon pool	18.50
$K_{soilstr}$	Turnover rate of soil structural carbon pool	4.80
$K_{soilmic}$	Turnover rate of soil microbial carbon pool	7.30
K_{cwd}	Turnover rate of CWD carbon pool	0.10
K_{slow}	Turnover rate of slow carbon pool	0.20
$K_{armored}$	Turnover rate of armored carbon pool	0.0045

Table 4. Statistics of E_{max} inferred from flux tower data for biome types including evergreen needleleaf forest (ENF), deciduous broadleaf forest (DBF), mixed forest (MF), closed shrubland (CSH), and open shrubland (OSH), woody savanna (WSA), grassland (GRA), cropland (CRO).

Biome type	ENF	DBF	MF	CSH	OSH	WSA	GRA	CRO
E_{max} mean	0.65	0.69	0.51	0.49	0.41	0.51	0.70	1.01
E_{max} SD	0.23	0.14	0.12	0.26	0.16	0.04	0.30	0.37
E_{max} samples for full uncertainty [E_1, E_2, E_3]	[0.50, 0.75, 0.75]	[0.50, 0.75, 0.75]	[0.25, 0.50, 0.75]	[0.25, 0.50, 0.75]	[0.25, 0.50, 0.50]	[0.25, 0.50, 0.50]	[0.50, 0.75, 1.00]	[0.75, 1.00, 1.25]

Table 5. Taylor skill scores ($S_T \times 10$) across temporal scales for (a) NEE estimates from MsTMIP and their naïve and optimal means, two priors (bc and b4) and posterior estimates from CT2017 (CT), SiB3CSU, CASA L2 ensemble, (b) GPP, and (c) RECO from CASA L2 ensemble and SiB3CSU. The best model and models have similarly good performance (Section 2.5) for each column are shown in bold. Results of full ensemble members is included in Table S5.

Model	3-hourly diurnal				Monthly	Annual
	DJF	MAM	JJA	SON		
(a) NEE						
BIOME BGC	3.0	4.9	6.1	5.7	4.1	2.2
CLASS CTEM	3.0	4.0	6.2	5.6	3.7	2.3
CLM4VIC	2.9	4.6	4.2	4.1	3.3	2.0
CLM	3.8	6.0	6.0	6.1	4.7	3.3
DLEM	1.9	5.5	6.3	5.5	4.3	2.0
GTEC	2.7	5.5	7.2	6.8	4.9	3.4
ISAM	2.7	5.8	4.4	3.4	2.5	1.5
LPJ	2.4	5.6	6.7	6.7	4.3	3.3
ORCHIDEE	2.4	5.5	5.9	4.0	4.5	2.4
SIB3	3.8	5.2	6.6	5.9	4.7	2.1
SIBCASA	4.0	5.7	7.0	6.5	4.9	1.8
TEM6	1.9	5.4	5.6	4.3	3.9	2.5
TRIPLEX	3.6	6.3	5.8	6.1	3.4	2.9
VEGAS	2.2	6.1	6.5	4.4	5.1	1.6
VISIT	2.5	3.0	3.9	3.7	2.4	1.3
MsTMIP naïve	3.7	6.5	7.0	6.5	4.6	2.4
MsTMIP optimal	3.9	6.3	6.4	6.1	4.3	2.0
CT b4 prior	2.8	6.5	7.6	6.8	4.8	0.8
CT bc prior	2.7	6.4	7.0	6.6	4.9	1.0
CT posterior	3.1	5.6	6.3	5.3	5.0	2.8
SiB3CSU	1.5	5.6	5.9	5.2	4.9	3.4
CASA L2 mean	2.8	6.4	7.5	6.8	6.9	2.6
(b) GPP						
SiB3CSU	0.6	4.8	6.4	5.6	6.7	2.5
CASA L2 mean	2.7	5.1	6.5	6.5	8.6	4.5
(c) RECO						
SiB3CSU	1.3	3.1	3.2	2.4	7.1	3.4
CASA L2 mean	3.9	5.3	4.8	5.1	7.6	4.6

Table 6. Mean errors (*MEs*, g C m⁻² time⁻¹) across temporal scales for (a) NEE estimates from MsTMIP and their naïve and optimal means, two priors (bc and b4) and posterior estimates from CT2017 (CT), SiB3CSU, CASA L2 ensemble, (b) GPP, and (c) RECO from CASA L2 ensemble and SiB3CSU. Mean errors are calculated as model estimate minus flux tower. The best scores for each column are shown in bold. Results of full ensemble members is included in Table S6.

Model	3-hourly diurnal ($\times 1000$)				Monthly	Annual
	DJF	MAM	JJA	SON		
(a) NEE						
BIOME BGC	-35.0	-30.4	-63.6	-41.7	20.6	249.7
CLASS CTEM	-22.6	-27.2	-81.2	-50.8	10.8	124.6
CLM4VIC	-21.8	-18.0	-47.6	-29.6	18.8	228.8
CLM	-27.4	-23.6	-55.6	-36.5	18.7	222.7
DLEM	-9.0	-18.5	-45.1	-27.4	18.3	212.5
GTEC	-21.0	-30.2	-90.7	-46.8	15.9	195.8
ISAM	-16.1	-20.8	-35.8	-19.9	20.9	245.1
LPJ	-28.3	-31.7	-64.7	-58.3	18.0	222
ORCHIDEE	-17.0	-31.7	-45.6	-23.8	18.5	219.6
SIB3	-26.3	-34.8	-70.5	-38.4	20.8	248.5
SIBCASA	-25.7	-31.6	-57.9	-38.5	18.0	213.1
TEM6	-8.5	-14.2	-33.6	-19.8	17.6	211.9
TRIPLEX	-28.4	-26.1	-54.2	-40.1	6.1	71.7
VEGAS	-15.7	-20.8	-54.1	-27.6	19.2	231.1
VISIT	-17.7	-33.0	-63.2	-31.0	16.2	198.4
MsTMIP naïve	-21.6	-25.7	-56.4	-35.1	17.4	208
MsTMIP optimal	-20.1	-25.1	-55.5	-34.6	17.5	208.8
CT b4 prior	-11.5	-20.0	-58.0	-34.7	19.4	234.1
CT bc prior	-10.2	-17.7	-52.1	-31.3	18.6	222.2
CT posterior	-15.1	-23.7	-56.7	-24.8	18.0	207.7
SiB3CSU	-9.8	-14.0	-38.8	-24.1	20.4	245.6
CASA L2 mean	-9.6	-20.3	-71.8	-34.3	21.0	253.3
(b) GPP						
SiB3CSU	-2.5	-22.3	-92.7	-11.0	-43.4	-560.3
CASA L2 mean	-12.4	-63.6	-136.3	-37.8	19.7	155.4
(c) RECO						
SiB3CSU	3.0	-11.4	-66.8	-0.5	-14	-165.1
CASA L2 mean	-7.1	-62.4	-40.2	-0.4	43.4	514.5

Figures [will change to Figure Captions in the submission]

Figure 1. Spatial maps of NEE in August 2010 across the North American domain from (a) CASA L2 mean and (b) spread which is the half difference between maximum and minimum, (c) SiB3CSU, (c) posterior and two priors (d) b4 and (f) bc estimates from CT2017, (g) naïve and (h) optimal means from MsTMIP.

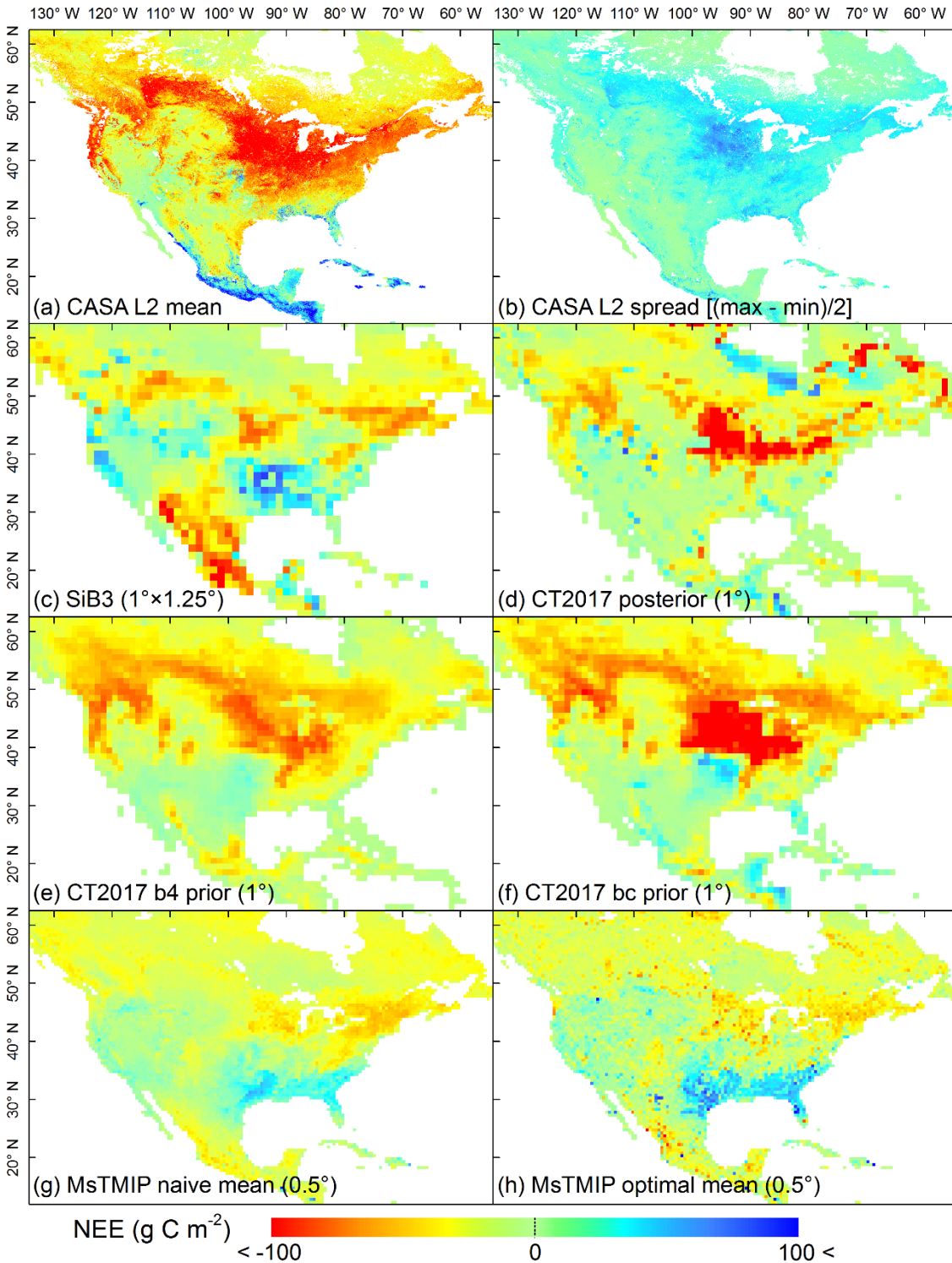


Figure 2. The first-order sensitivities of (a) GPP, (b) RECO, and (c) NEE to 12 CASA parameters with a 20% perturbation of each default value. The boxes represent the values for the 25th and 75th percentiles, the horizontal red line gives the median, and the whiskers show the 1th and 99th percentiles, and the outliers are plotted individually using the red cross symbol.

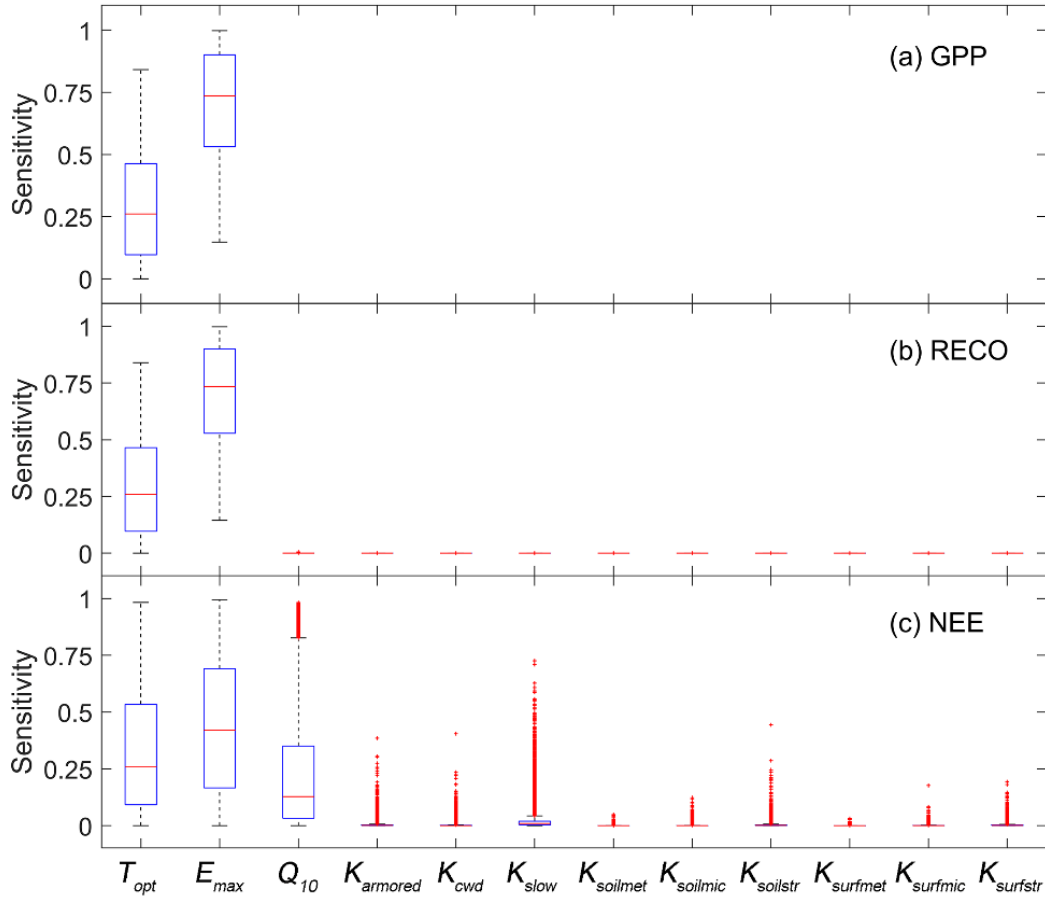


Figure 3. Taylor diagrams showing the monthly scale correspondence between AmeriFlux-measured NEE and simulated NEE from (a) CASA L2 ensemble mean, SiB3CSU, two priors (bc and b4) and posterior estimates from CarbonTracker2017 (CT2017), modeled ensemble from MsTMIP and their naïve and optimal means, and (b) among CASA L2 27 ensemble members. The red dot represents the AmeriFlux NEE. Correlation coefficient (r , blue), standard deviation (SD , $\text{g C m}^{-2} \text{ mo}^{-1}$, black) and root mean square error ($RMSE$, $\text{g C m}^{-2} \text{ mo}^{-1}$, green) are averaged across AmeriFlux sites.

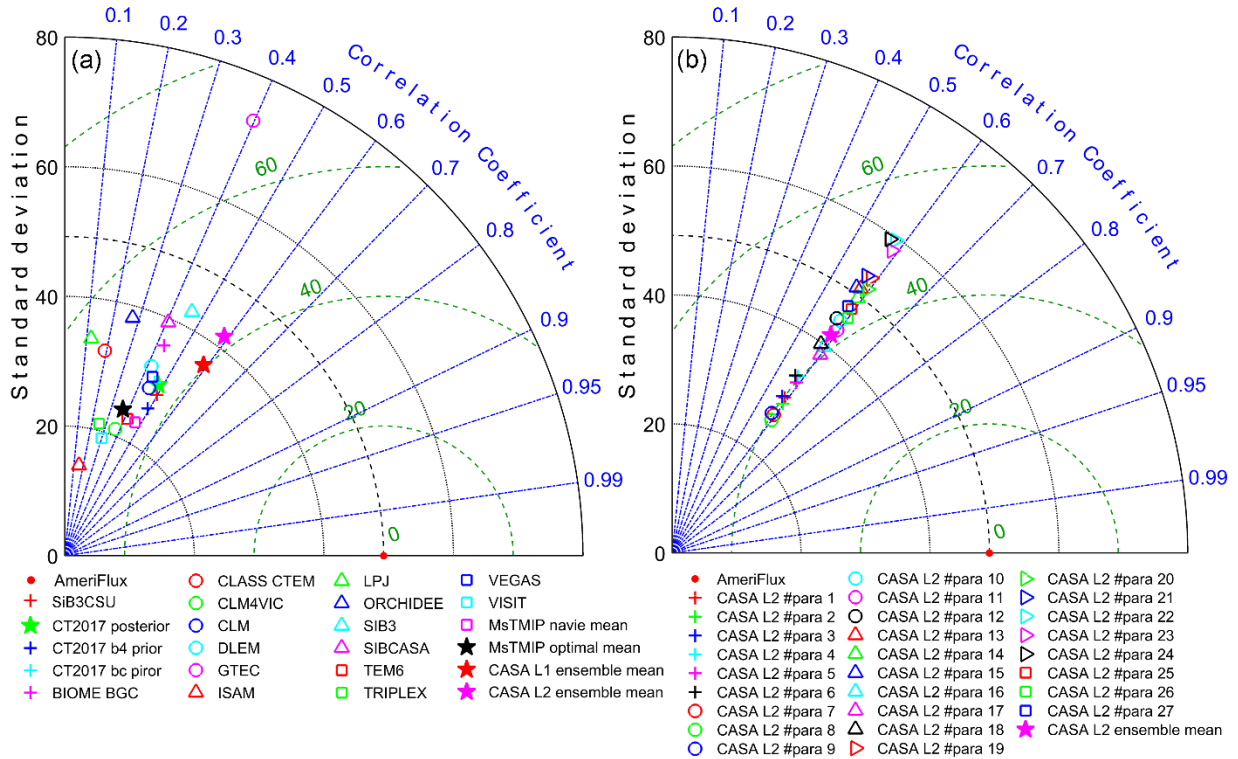


Figure 4. Mean monthly bias of CASA L2 GPP (green), RECO (red), and NEE (blue) compared to AmeriFlux results for five biome types, including (a) evergreen needleleaf forests (ENF), (b) deciduous broadleaf forests (DBF), (c) open shrublands (OSH), (d) grasslands (GRA), and (e) croplands (CRO). The solid line represents CASA L2 ensemble mean, and the shaded area are max and min biases.

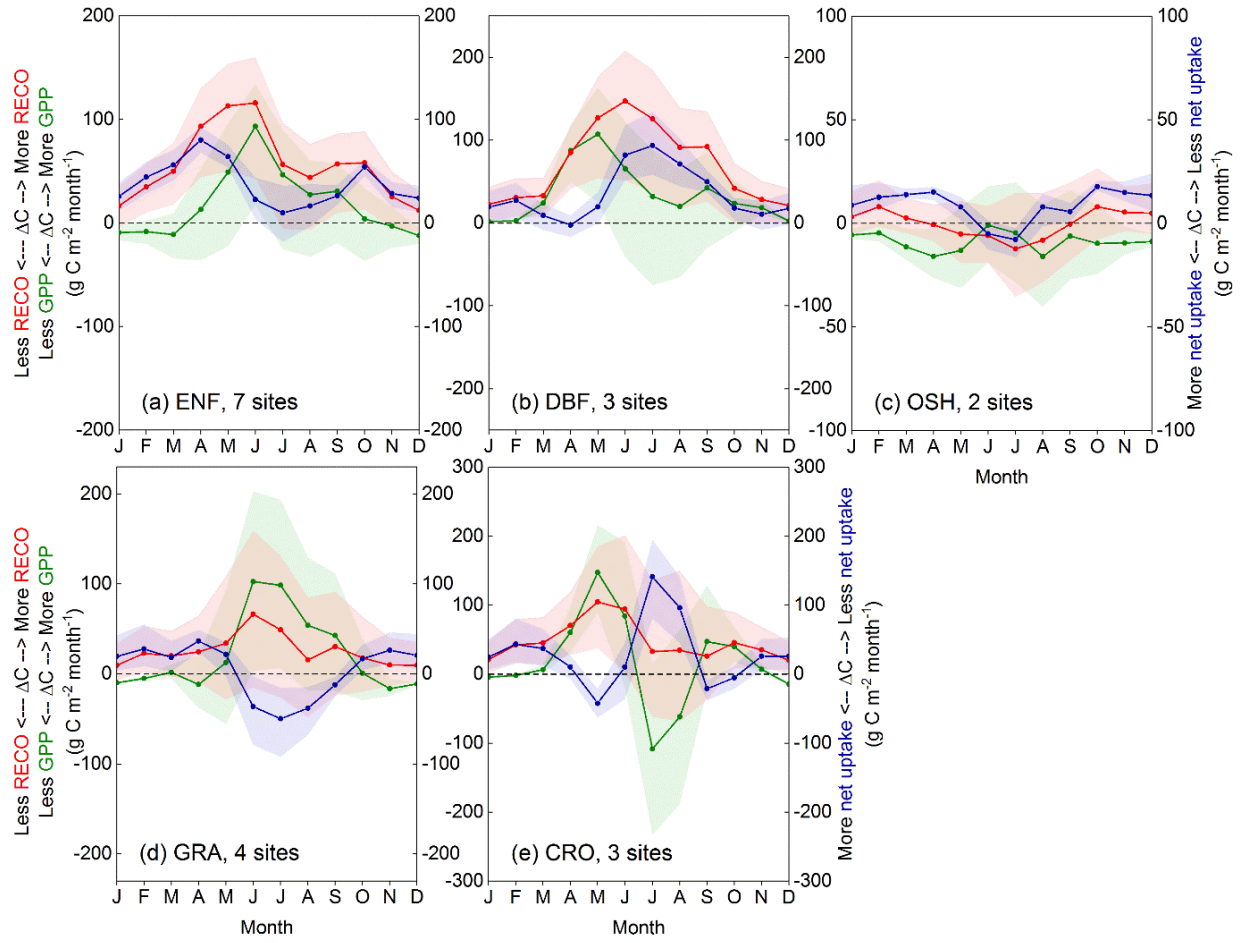


Figure 5. Mean monthly bias of CASA L2 NEE compared to AmeriFlux results (all available sites) for biome types, including (a) evergreen needleleaf forests (ENF), (b) deciduous broadleaf forests (DBF), (c) mixed forest (MF), (d) closed shrublands, (e) open shrublands (OSH), (f) woody savannas, (g) grasslands (GRA), and (h) croplands (CRO). The solid line represents CASA L2 ensemble mean, and the shaded area are max and min biases.

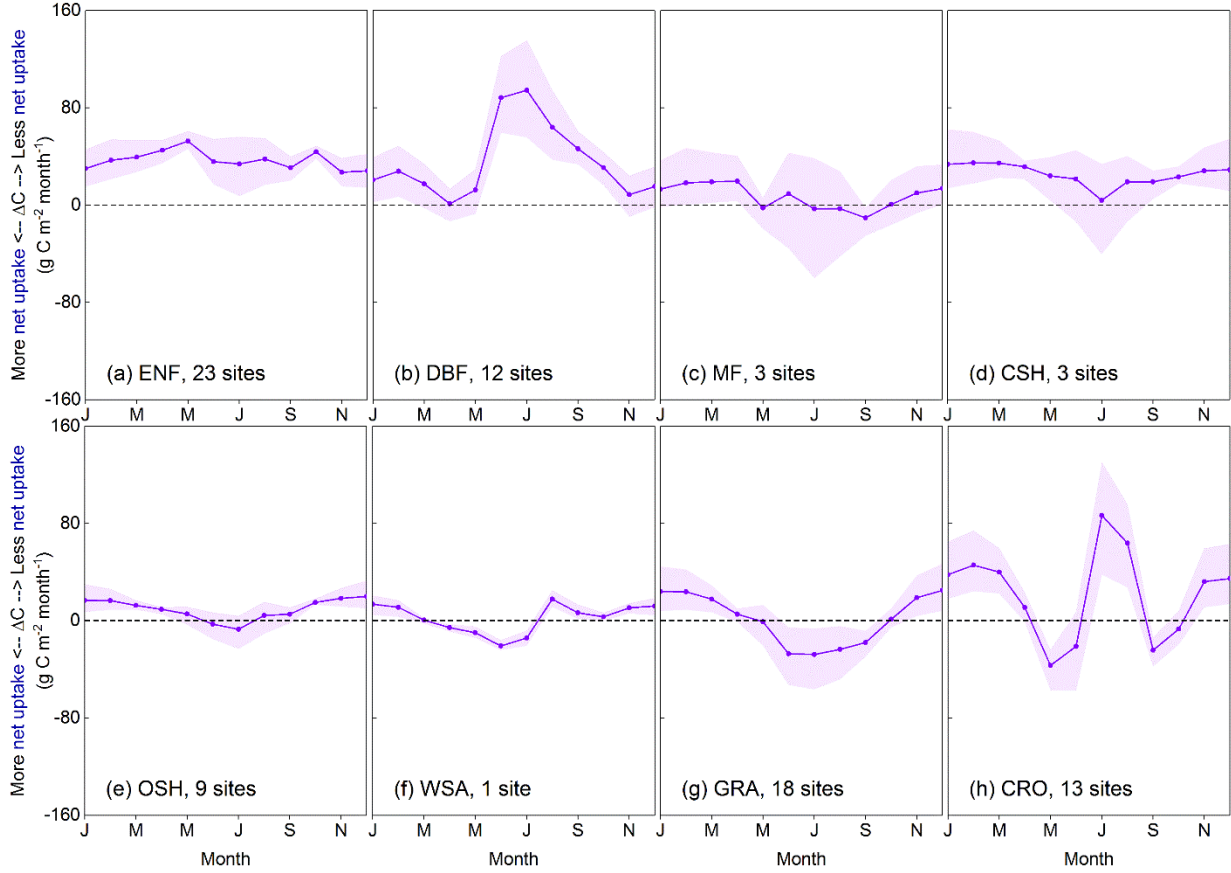


Figure 6. Taylor diagrams showing the 3-hourly correspondence in four climatic seasons (a) December to February (DJF), (b) March to May (MAM), (c) June to August (JJA), and (d) September to November (SON) between deseasoned AmeriFlux-measured NEE and deseasoned, simulated NEE from means of CASA L1 and L2 ensembles, SiB3CSU, two priors (bc and b4) and posterior estimates from CarbonTracker2017 (CT2017), modeled ensemble from MsTMIP and their naïve and optimal means. Correlation coefficient (r , blue), standard deviation (SD , $\text{g C m}^{-2} \text{ 3h}^{-1}$, black) and root mean square error ($RMSE$, $\text{g C m}^{-2} \text{ 3h}^{-1}$, green) are averaged across AmeriFlux sites.

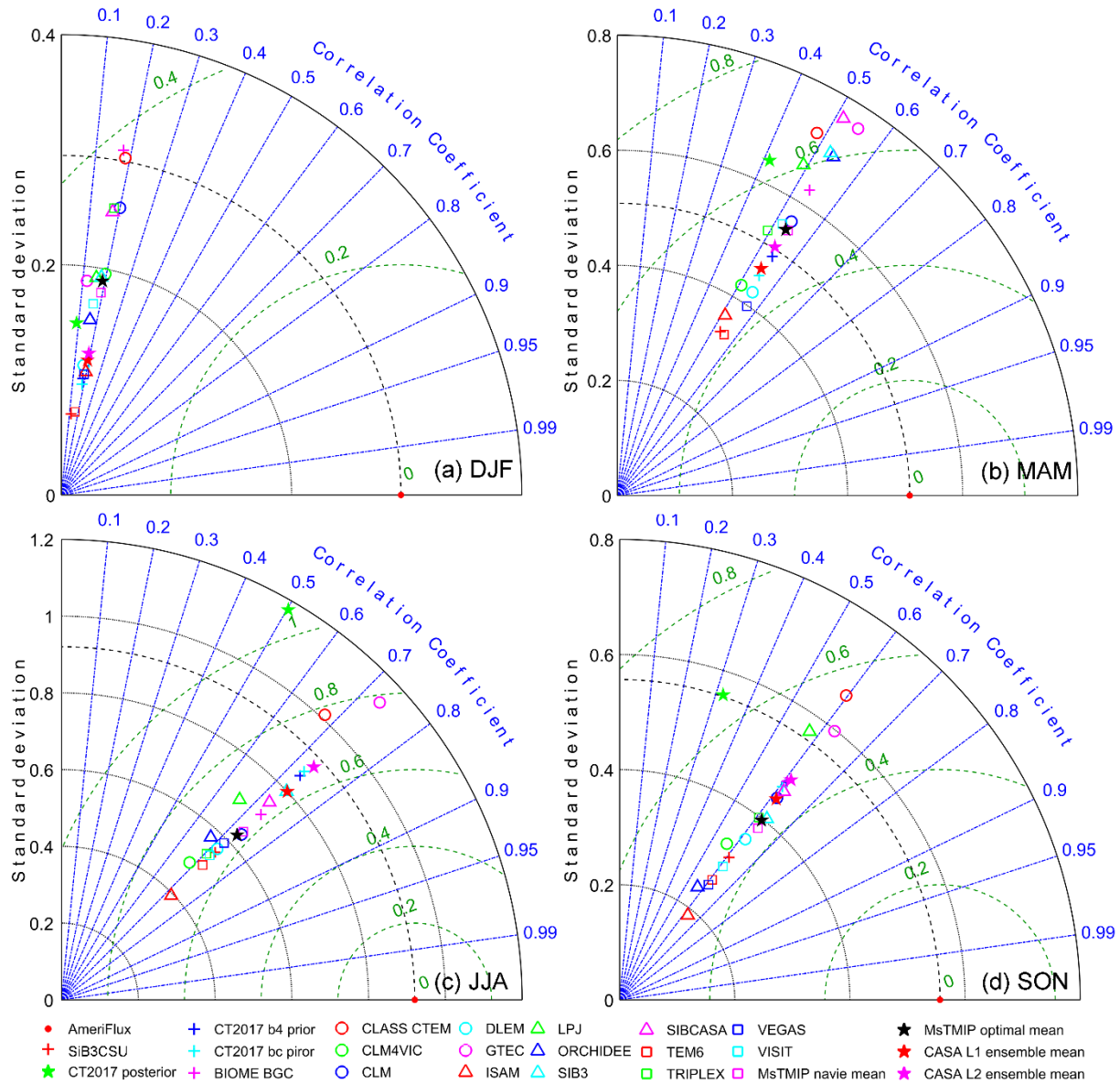


Figure 7. Monthly mean NEE (g C m^{-2}) variations in 2006 – 2010 derived from CASA L2 ensemble mean (solid line) and spread (shaded area), MsTMIP naïve and optimal means, CT2017 bc and b4 priors, CT2017 posterior, and SiBCSU of 12 biome types. The percentage of each biome type accounts for all vegetated area is included in each subplot.

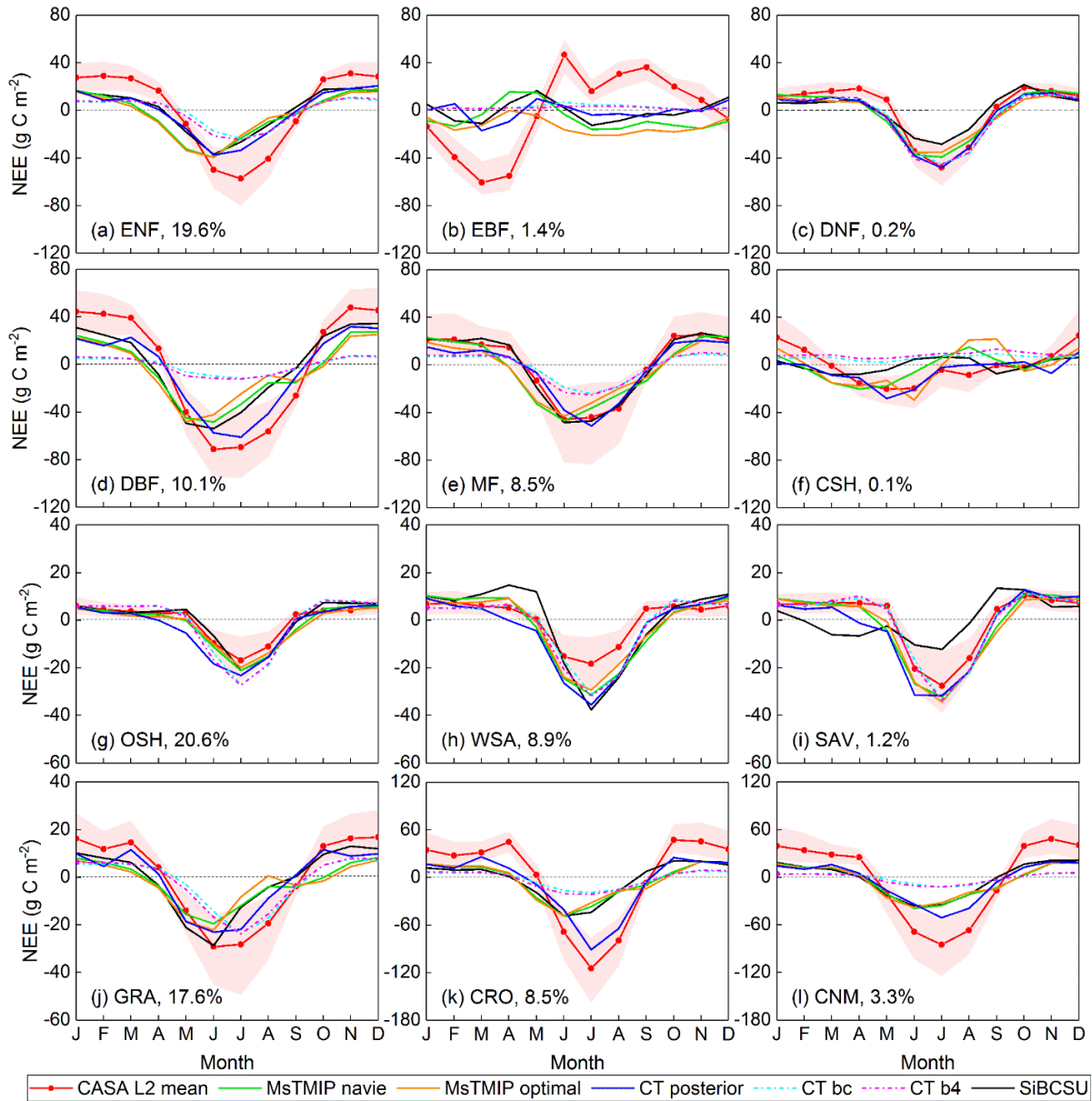


Figure 8. Monthly mean NEE (g C m^{-2}) variations in 2006 – 2010 derived from CASA L2 ensemble mean (solid line) and spread (shaded area), MsTMIP naïve and optimal means, CT2017 bc and b4 priors, CT2017 posterior, and SiBCSU of 15 North American ecoregions. The subfigures are ranked by geolocations from north (top) to south (bottom) and from east (left) to west (right). The area of each ecoregion is also included in each subplot.

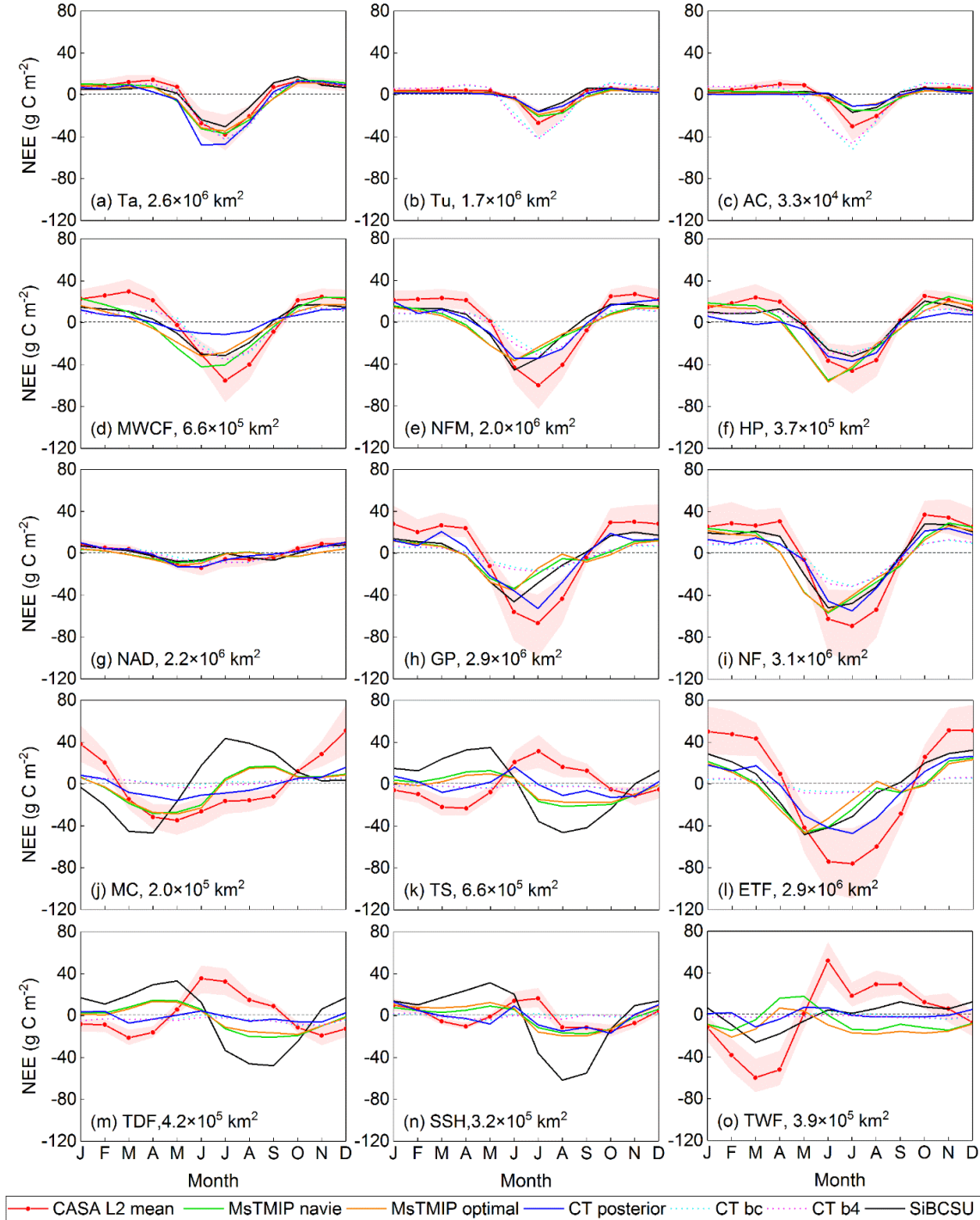


Figure 9. Monthly NEE variations in 2006 – 2010 of (a) North America and (b) conterminous US derived from CASA L2 ensemble mean (solid line) and spread (shaded area), MsTMIP naïve and optimal means, CT2017 (CT) bc and b4 priors, CT2 017 posterior, and SiB3CSU.

