A modified Vegetation Photosynthesis and Respiration Model (VPRM) for the eastern USA and Canada, evaluated with comparison to atmospheric observations and other biospheric models

Sharon M. Gourdji[1], Anna Karion[1], Israel Lopez-Coto[1], Subhomoy Ghosh[1,2], Kim Mueller[1], Yu Zhou[3], Chris Williams[3], Ian Baker[4], Katharine Haynes[4], James Whetstone[1]

1. National Institute of Standards & Technology, Gaithersburg, MD, USA, 20899, USA
2. Center for Research Computing, University of Notre Dame, South Bend, IN, 46556, USA
3. Graduate School of Geography, Clark University, Worcester, MA, 01610, USA
4. Cooperative Institute for Research in the Atmosphere (CIRA), Colorado State University, Fort Collins, CO, 80521, USA

Key points

1. VPRM was customized for eastern North America with a respiration model including EVI, quadratic temperature and water stress factors.
2. The modified VPRM improves agreement with atmospheric CO_2 observations by increasing growing season respiration.
3. The modified VPRM better explains hourly atmospheric CO_2 variability from April to October compared to other models in 2016/ 2017.

Plain Language Summary

Monitoring fossil fuel emissions with atmospheric CO_2 measurements can provide an independent check on bottom-up estimates and support mitigation policies by tracking emission trends over time and identifying unknown sources. However, atmospheric CO_2 is influenced by anthropogenic emissions and the natural carbon cycle from plants and soils, which contributes a strong hourly-varying signal in the atmosphere during the growing season. Here we implement a relatively simple model of the biosphere, i.e., the Vegetation Photosynthesis and Respiration Model, at high spatiotemporal resolution in eastern North America. The equation describing sources to the atmosphere from respiration (i.e., "breathing" from plants and decaying organic matter) is modified to make it more physiologically realistic by accounting for increases in leaf respiration during summer. Model estimates are compared with output from other similar biospheric models and with atmospheric CO_2 observations, and results show that the new VPRM better explains CO_2 fluctuations in the atmosphere during the growing season compared to other models. The model improvements shown here demonstrate promise for helping to isolate the biospheric signal in atmospheric CO_2 measurements and thus improve estimation of fossil fuel emissions year-round in areas with significant nearby and upwind vegetation.

Abstract

Increasing atmospheric CO_2 measurements in North America, especially in urban areas, may help enable the development of an operational CO_2 emission monitoring system. However, isolating the fossil fuel emission signal in the atmosphere requires factoring out CO_2 fluctuations due to the biosphere, especially during the growing season. To help improve simulations of the biosphere, here we customize the Vegetation Photosynthesis and Respiration Model (VPRM) at high-resolution for an eastern North American domain, upwind of coastal cities from Washington D.C. to Boston, MA, optimizing parameters using domain-specific flux tower data from 2001 to the present. We run three versions of VPRM from November 2016 to October 2017 using i) annual ($VPRM_{ann}$) and ii) seasonal parameters ($VPRM_{seas}$), and then iii) modifying the respiration equation to include the Enhanced Vegetation Index (EVI), a squared temperature term and interactions between temperature and water stress ($VPRM_{new}$). VPRM flux estimates are evaluated by comparison with other models (the Carnegie-Ames-Stanford Approach model, or CASA, and the Simple Biosphere Model v4), and with comparison to atmospheric CO_2 mole fraction data at 21 surface towers. Results show that $VPRM_{new}$ is relatively unbiased and outperforms all other models in explaining CO_2 variability from April to October, while $VPRM_{ann}$ overestimates growing season sinks by underestimating summertime respiration. Despite unknown remaining errors in $VPRM_{new}$, and uncertainties associated with other components of the atmospheric CO_2 comparisons, $VPRM_{new}$ appears to hold promise for more effectively separating anthropogenic and biospheric signals in atmospheric inversion systems in eastern North America.

1. Introduction

Carbon dioxide (CO_2) surface fluxes from the terrestrial biosphere produce a large and variable signal in the atmosphere during the growing season (due to both photosynthetic uptake and ecosystem respiration) that can dwarf the signal from fossil fuel emissions in biologically productive areas (Shiga et al., 2014). Even in the dormant season (e.g., December to March in the northern hemisphere), biospheric sources from ecosystem respiration have the same sign and an accumulated signal in the atmosphere potentially as large as that from fossil fuel emission sources, especially when integrated over large areas. Within cities and their suburban surroundings there is also an active biosphere year-round within parks, lawns, and gardens, and from urban street trees and city-scale agriculture (Buyantuyev & Wu, 2009; Golubiewski, 2006; Nowak & Crane, 2002; Raciti et al., 2014).

Recent efforts to estimate anthropogenic CO_2 emissions using inversion models with atmospheric measurements collected in urban and suburban areas close to fossil fuel emission sources are confounded by the difficulty in separating out the biospheric from the fossil fuel signal (Lauvaux et al., 2021; Miller et al., 2020; Sargent et al., 2018; Shiga et al., 2014; Yadav et al., 2016). Many previous inversion studies (Gurney et al., 2002; Hu et al., 2019; Rödenbeck et al., 2003) at continental and global scales factored out the influence of fossil fuel emissions (assumed as well-known) from atmospheric observations and then used the inversion to optimize biospheric flux estimates. However, in order to estimate emissions with atmospheric inversions, the problem must be reversed by assuming biospheric fluxes as known or separately estimating both biospheric and anthropogenic fluxes with additional tracers like C^{14} (e.g., Basu et al., 2020). Regardless, any errors in biospheric CO_2 flux estimates will be directly aliased onto emission estimates, and thus, a high-quality biospheric model at fine spatial and

temporal scales consistent with the variability of CO_2 in the atmosphere can further help to isolate the emissions signal. Put another way, the better that the biospheric CO_2 signal can be modeled in the atmosphere (at surface observing locations, along aircraft trajectories and/ or in total columns as seen from satellites), the more statistical power that inversion models will have to adjust the emission signal using observed atmospheric CO_2 mole fractions. In fact, the terrestrial biosphere is considered as one of the largest sources of uncertainty in atmospheric CO_2 inversions during the growing season for the North American and smaller regional domains (Feng et al., 2019a; Feng et al., 2019b; Sargent et al., 2018).

A further complication with inversions designed to isolate the anthropogenic CO_2 signal, especially in urban areas, is that biospheric CO_2 fluxes need to be appropriately modeled not just inside the specified domain, but also in upwind areas outside the domain (also known as background conditions). For urban areas downwind of significant vegetation (e.g., agriculture or forests), determining the background CO_2 contribution to observed mole fractions can be a substantial challenge (Karion et al., 2021; Lauvaux et al., 2021; Sargent et al., 2018). This is exemplified by recent efforts to expand the urban and suburban atmospheric CO_2 monitoring network in the Northeast Corridor (NEC) of the USA (from Washington D.C. to Boston, MA; Karion et al., 2020; Lopez-Coto et al., 2017; Pitt et al., 2020; Sargent et al., 2018), given that regions upwind of the NEC include the biologically productive Appalachian deciduous forests, northern mixed forests and southern pine plantations, croplands in the Midwestern Corn Belt and Mississippi river valley, grasslands in Kentucky and Tennessee and coastal and northern wetlands (Figure 1). Furthermore, the influence of background conditions vs. fluxes inside the domain is largely a function of how the domain is defined. Future efforts to combine multiple urban areas into nested domain inversions (with high-resolution fluxes estimated in urban areas and coarser resolution outside, e.g., Schuh et al., 2019; Turner & Jacob, 2015), would increase the signal to noise inside the domain by using towers with overlapping constraints and reduce the influence of background uncertainty. With such a setup, it will become even more important to appropriately model the biosphere at high spatial and temporal resolution in both urban and surrounding rural areas.

Towards these ends, we focus here on improving and evaluating biospheric models specifically for a domain in the eastern USA and Canada upwind of the NEC (Figure 1). For this region, we create a customized version of the Vegetation Photosynthesis and Respiration Model (VPRM; Mahadevan et al., 2008) for a single year (November 2016 to October 2017), optimizing model parameters with data from flux towers in and near the domain (Figure 1), and using both annual ($VPRM_{ann}$) and seasonally-varying ($VPRM_{seas}$) parameters. Given the relatively simplistic respiration model in the original Mahadevan et al. (2008) formulation (with a baseline value and linear temperature dependence), we also introduce a modified respiration model in VPRM (i.e., $VPRM_{new}$) that includes additional covariates: i) a quadratic temperature term, ii) a vegetation index to better capture seasonality in autotrophic respiration, and iii) a water stress scaling factor and its interactions with temperature to capture drought and soil moisture effects.

By comparing three versions of VPRM optimized for this domain, we evaluate different aspects of model improvement, and identify their contributions to improved performance relative to two types of observations: flux tower direct measurements of Net Ecosystem Exchange (NEE) and near-surface atmospheric CO_2 mole fractions. We also evaluate VPRM in comparison to two other commonly-used terrestrial biosphere models in North America: the Carnegie-Ames Stanford Approach (CASA) model (Potter et al., 1993; Randerson et al., 1996; Zhou et al., 2020) and the Simple Biosphere model, version 4

(SiB4; Haynes et al., 2019; Sellers et al., 1986, 1996). VPRM, CASA and SiB4 represent a range of biospheric modeling approaches that vary from the most empirical (VPRM) to the most process-based and mechanistic (SiB4), with the CASA model of intermediate complexity. Additional details of the models and their principal differences are discussed in Section 2.1.

The seasonal cycles, diurnal cycles, and gridded spatial patterns of CO_2 flux estimates are first compared across models, with the goal to identify differences and commonalities in model inputs, formulation and outputs, and potential mechanisms contributing to differences. Next, we compare modeled fluxes to flux tower NEE observations at 22 sites in the domain with relatively complete data in 2016/ 2017, examining how each model's diurnal and seasonal cycle compares to observations at these discrete locations. Finally, we couple modeled CO_2 fluxes with two different transport and dispersion models and compare simulated atmospheric CO_2 mole fractions to measurements collected at 21 surface tower locations in our domain (Figure 2). Fossil fuel emissions and background conditions are factored out to isolate atmospheric CO_2 enhancements due to the biosphere, and simulations are compared to observations using both monthly means and hourly time series to assess how well modeled fluxes reproduce spatial gradients, seasonality and finer-scale (diurnal and synoptic) temporal variability in the atmosphere. The comparison with atmospheric CO_2 measurements helps to evaluate the models using a regional integrated signal (albeit with intermittent sampling in space and time and dilution due to atmospheric mixing and transport), and thus identify how biological flux models manifest in reproducing atmospheric CO_2 variability.

2. Methods

2.1 Biospheric models

All three biospheric models (VPRM, CASA and SiB4) separately estimate carbon release to the atmosphere due to ecosystem respiration (R_e, or the sum of autotrophic, R_a, and heterotrophic, R_h, respiration) and carbon uptake due to photosynthesis (also known as Gross Primary Production, or GPP), with Net Ecosystem Exchange (NEE) defined as their difference (i.e., R_e - GPP). Each model also differentiates flux dynamics across land cover categories, referred to as Plant Functional Types (PFTs), and then merges the flux estimates from each category into a total flux for each pixel. However, the models differ in their physiological representations of GPP and R_e, their input datasets, PFT classifications and phenology (or seasonal timing) schemes, as will be discussed further below (and shown in Table S1).

Each biospheric model was run for a single year from Nov 1, 2016 to Oct 31, 2017 for our domain of interest: -92 °W to -68 °W, 33 °N to 47 °N (Figure 1), with all models run at or downscaled to hourly temporal resolution. The spatial resolution for each model varies, with VPRM run at 0.02° (approximately 2 km^2), CASA at 500 m in the coterminous USA (and 5 km in Canada), and SiB4 at 0.5° (approximately 50 km). Flux estimates from all models are aggregated/ disaggregated to 0.1° for further analysis, a scale which allows for comparison of fine-scale spatial variability and computational tractability. The availability of transport model runs from two models limited the study to this single year; however, weather patterns in this year were within the range of 20 years of interannual variability in most parts of the domain (Figure S1).

In this section, we review the original VPRM model from Mahadevan et al. (2008), and then describe the modified version developed for this study (Table 1). We then briefly describe the model structure and implementations of CASA and SiB4 included in the model inter-comparison, with further details and inputs for all three models also included in Table S1.

2.1.1 VPRM

The Vegetation Photosynthesis and Respiration Model (VPRM) is a diagnostic light-use efficiency model that relies on remote-sensing inputs to calculate GPP and simulate phenology. It was implemented here using the original equations from Mahadevan et al. (2008), as well as a version with a respiration equation modified to include additional covariates, as described below. PFT-specific parameters for both versions are optimized using hourly NEE observations from 69 flux towers in and near our domain with data at any time from 2001 to the present (Figure 1; Table S2), excluding data in 2017 for model evaluation. (More details on the flux tower data is included in the supplemental material.)

The GPP equation is the same for all VPRM model versions, and is defined as:

$$GPP = \lambda * T_{scale} * P_{scale} * W_{scale1} \frac{1}{\left(1 + \frac{PAR}{PAR_0}\right)} * PAR * EVI \quad \textit{Eq. 1}$$

where PAR is Photosynthetically Active Radiation, EVI is the Enhanced Vegetation Index, and T_{scale}, P_{scale}, and W_{scale1} are temperature, phenology and water stress scaling factors, as defined in Mahadevan et al. (2008). Parameters optimized with flux tower data are λ (a potential light-use efficiency factor) and PAR_0 (the half-saturation constant of PAR, or photosynthetic efficiency at high light levels). PAR*EVI represents the amount of absorbed radiation, with GPP modeled as potential uptake (i.e., λ*PAR*EVI) downregulated by each of the scaling factors.

W_{scale1} is defined for all PFTs as:

$$W_{scale1} = \frac{1 + LSWI}{1 + LSWI_{max}} \quad \textit{Eq. 2}$$

where LSWI is the remotely-sensed Land Surface Water Index (Chandrasekar et al., 2010) from MODIS, and $LSWI_{max}$ is the site-specific maximum daily LSWI from a multi-year mean for May to October (Xiao et al., 2004).

The original VPRM model formulation estimates R_e as a baseline value plus linear function of temperature:

$$R_e = \beta + \alpha * T_{air}. \quad \textit{Eq. 3}$$

Here, T_{air} is the surface air temperature (in °C), and β (baseline level) and α (temperature sensitivity of respiration) are optimized parameters. In Mahadevan et al. (2008), T_{air} below a threshold (T_{low}) is set

equal to T_{low} to account for continued respiration in winter, when soils remain warmer than air temperatures. Here instead, we just set predicted negative R_e values to zero.

For the original VPRM model, GPP (λ and PAR_0) and R_e (β and α) parameters are optimized simultaneously by minimizing least squares across all 24-hours of hourly NEE observations. This optimization procedure ensures zero bias on an average basis across the full time period of the data but does not guarantee that optimized parameters will explain all variability, e.g. the peaks of the diurnal or seasonal cycles. We first optimize a set of annual (i.e., time-invariant) parameters for each PFT, and then four sets of seasonally varying parameters (for December to February, March to May, June to August and September to November; Table S3), given that the optimal grouping of parameters (in time and space) in VPRM remains an open question (T. W. Hilton et al., 2013, 2014). These VPRM runs using the original VPRM equations are henceforth referred to as $VPRM_{ann}$ (annual parameters) and $VPRM_{seas}$ (seasonal parameters) throughout the paper (Table 1).

To help improve the respiration model, the equation for R_e is then updated to include additional predictor variables:

$$R_e = \alpha_1 * T' + \alpha_2 * T'^2 + \gamma * EVI + \theta_1 * W_{scale2} + \theta_2 * W_{scale2} * T' + \theta_3 * W_{scale2} * T'^2 \quad \text{Eq. 4}$$

where α_1, α_2, γ, θ_1, θ_2 and θ_3 are optimized parameters (discussed below), and T' are modified low temperatures designed to simulate soil temperatures that remain warmer than air temperatures in winter:

$$For\ T_{air} < T_{crit} \qquad T' = (T_{air} - T_{crit}) * T_{scale} + T_{crit} \quad \text{Eq. 5}$$

$$For\ T_{air} \geq T_{crit} \qquad T' = T_{air} \quad \text{Eq. 6}$$

where T_{crit} is a low temperature threshold (in °C) and T_{scale} is a scalar from 0 to 1, which is multiplied by air temperatures below T_{crit}, and both are optimized parameters. This modification of low air temperatures for R_e is like the fixed T_{low} threshold in Mahadevan et al (2008), but here T' can still slope downward in winter. A slightly different water stress scaling factor is also used in the updated R_e equation, with W_{scale2} defined as a normalized LSWI:

$$W_{scale2} = \frac{LSWI - LSWI_{min}}{LSWI_{max} - LSWI_{min}}. \quad \text{Eq. 7}$$

The new respiration model formulation (Equation 4) was chosen by running various multivariate regressions against observed nighttime average NEE at the flux towers to determine a single equation that consistently improves model fit across PFTs. Models were evaluated by comparing adjusted R^2s across different sets of potential covariates, which, like other model selection algorithms, penalizes the addition of spurious predictor variables. W_{scale2} gave a slightly better model fit for the grasslands,

soybean/ other crops and shrubland PFTs compared to W_{scale1}, and equivalent fits for the other PFTs; hence this definition of W_{scale} was chosen for the R_e equation.

In the modified respiration equation, the squared temperature term introduces a non-linear temperature response, while EVI introduces realistic seasonality and spatial patterns into R_e estimates. The W_{scale2} parameter and its interactions with temperature help to account for water stress, especially at high temperatures when soils tend to dry out. Literature supports the use of these additional factors to help explain R_e fluxes, given that autotrophic respiration has large seasonal increases associated with canopy development (Jassal et al., 2007), current assimilation (i.e., photosynthetic uptake) is known to account for a large portion of above and below-ground autotrophic respiration during the growing season (Amthor, 2000; Högberg et al., 2001), and that soil moisture limits both autotrophic and heterotrophic respiration during drought periods (Flexas et al., 2006; Meir et al., 2008; Molchanov, 2009). These additional factors have also previously been suggested as needed improvements to the VPRM R_e equation in Li et al. (2020).

With the new R_e model, all parameters (α_1, α_2, γ, θ_1, θ_2, and θ_3) in Equation 4 are estimated as coefficients from the nighttime NEE regressions using flux tower data for each PFT, while T_{crit} and T_{scale} are optimized by maximizing R^2 values in the regressions across a range of realistic values. GPP parameters are then optimized using partitioned daytime GPP observations (determined by subtracting predicted daytime respiration from observed daytime NEE). Only one set of time-invariant parameters are optimized for each PFT with this updated R_e model given that EVI and LSWI help to account for seasonal changes (Table S4). This version of VPRM is henceforth referred to as $VPRM_{new}$ (Table 1).

For all three versions of VPRM ($VPRM_{new}$, $VPRM_{ann}$ and $VPRM_{seas}$), the original PFT classification from Mahadevan et al. (2008) and the AmeriFlux database was re-examined to see how tower-specific optimized parameters cluster across and between PFTs. Based on this preliminary analysis (Figure S2), corn is separated from other crop categories, evergreen needleleaf and mixed forests are merged and then divided into north and south at 40°N (the halfway latitude in the domain, where fast-growing pine plantations in the south behave differently from more mature forests in boreal areas, Figure 1). This is consistent with previous work (Hilton et al., 2013; J. Xiao et al., 2011) showing that stand age and disturbance history may be equally as important as climate and PFTs for understanding NEE variations at large regional scales. The optimal temperature for corn (which goes into the T_{scale} parameter) was set higher than for other crops to match literature values and reflects higher temperature optima for C4 relative to C3 crops (Tables S3 and S4). Minimum, maximum and optimal temperatures for other PFTs are taken from previous implementations of VPRM and based on literature values (T. W. Hilton et al., 2013; Mahadevan et al., 2008).

For all versions of VPRM, developed land with low, medium and high intensity in the National Land Cover Database (NLCD; Jin et al., 2019; Yang et al., 2018) in the USA (and "urban/ developed" in the Canadian land cover product, Table S1, Figure 1) is classified as an urban PFT. However, the "developed-open" category in the NLCD, i.e., cemeteries, gardens, lawns, and parks, is instead classified as grasslands, with the assumption (in the absence of other information) that they behave like other unmanaged grasslands. Parameters for the urban PFT are assumed to be the same as for deciduous broadleaf forests (as in Hardiman et al., 2017), presumably the native vegetation of most cities in our domain. A correction was then applied to reduce heterotrophic respiration (assumed as half the total) by the fraction of impervious surfaces in the urban PFT, but the autotrophic respiration correction in

Hardiman et al (2017) was not applied, given the difficulty in identifying reference pixels outside of every city in the domain. For $VPRM_{new}$, R_e is also lower in urban areas due to lower EVI values.

Gridded VPRM fluxes are estimated across the domain using the single GPP equation and two different R_e equations using PFT-specific parameters, with total fluxes derived as weighted averages of PFT-specific fluxes using fractional gridded land cover maps. Fluxes are estimated hourly, using daily EVI and LSWI inputs (interpolated from overlapping 16-day and 8-day composite products from the MODIS Terra and Aqua satellites), and hourly air temperature and radiation data from the High Resolution Rapid Refresh model (HRRR; Benjamin et al., 2016). VPRM inputs, including land-cover maps, are described in more detail in the supplemental material.

2.1.2 CASA

The CASA model was first developed in the 1990s to take advantage of remote-sensing data from NASA satellites (Potter et al., 1993; Randerson et al., 1996) and to probe scientific questions about the global carbon cycle at coarse spatial scales. Since then, it has been used extensively as a biospheric prior in global and continental inversion studies, e.g. in the North American CarbonTracker CO_2 inversion system from NOAA-ESRL (https://www.esrl.noaa.gov/gmd/ccgg/carbontracker/). CASA is a diagnostic light-use efficiency model, which incorporates remotely-sensed data, meteorological inputs and light-use efficiency factors to estimate GPP. Unlike VPRM, CASA also includes a process-based respiration model, originally based on a simplified version of the CENTURY model (Parton et al., 1988), which tracks carbon across three live pools (leaves, stems, roots), three litter pools, five soil pools and two coarse woody debris pools (Zhou et al., 2020, Table S1). Respiration fluxes are then determined from each pool as a function of carbon supply from photosynthetic uptake, pool-specific turnover and decay-rate constants, and environmental stress factors.

The CASA model operates at a monthly timestep but an algorithm was introduced by Fisher et al. (2016) to downscale monthly fluxes to 3-hour resolution using temperature and radiation data (further linearly interpolated to an hourly resolution here). This downscaling algorithm has proven useful for inversion studies that need to account for diurnal variability in biospheric fluxes in order to avoid biasing flux estimates at coarser temporal and spatial scales (Gourdji et al., 2010; L. Hu et al., 2019). Even with the temporal downscaling, monthly fPAR (fraction of photosynthetically active radiation) is still used to track phenology in CASA, which can exacerbate phenological errors during times of rapid seasonal transitions (Guindin-Garcia et al., 2012; Zhou et al., 2020). A comparison between the monthly fPAR used in CASA and 8-day EVI for VPRM in cropland and deciduous broadleaf forest pixels (indicated in Figure 1) is shown in Figure S3 in the supplemental material.

The implementation of CASA used here was run as an ensemble for the ACT-America project (Zhou et al., 2020a; Zhou et al., 2020b) at relatively fine spatial scales (500 m in the coterminous US and 5 km in Canada). Here, we use the Level 2 pruned version of the ensemble with 27 members, which contains PFT-specific parameters calibrated with flux tower data. We then use the ensemble mean with downscaled 3-hour fluxes interpolated to hourly resolution throughout the analysis. Ensemble means typically have superior performance compared to individual ensemble members due to a reduction in random errors (Elder, 2018; Schwalm et al., 2010). Zhou et al. (2020) also demonstrated that the CASA

ensemble mean included in this study agrees well with flux tower measurements compared to other models, with a reasonable downscaled diurnal cycle.

For this study, we merge the 500 m fluxes in the US (about ¾ of land area in the domain) with the 5 km resolution fluxes in Canada. Given that CASA is run using the dominant land cover in each pixel, we expect that the high resolution of the 500 m runs in the USA will help to improve model performance in this part of the domain by better representing patchy land covers relative to the 5 km product. Table S1 and Zhou et al. (2020) show other details of the CASA implementation used here.

2.1.3 SiB4

The Simple Biosphere Model (SiB), despite the name, is a complex process-based, fully prognostic model of the carbon cycle which can be used to predict future carbon dynamics. Unlike CASA and VPRM, SiB simulates both the carbon and energy cycles, and was originally developed for coupling with general circulation models to improve their boundary conditions (Sellers et al., 1986, 1996). Like CASA, respiration in SiB4 is calculated by tracking carbon pools, although with five live pools (including products for agriculture) and six dead pools (two for litter, three for soil and one for coarse woody debris; Table S1).

GPP in SiB4 is estimated using the Farquhar et al. (1980) enzyme-kinetic photosynthesis algorithm (unlike CASA and VPRM) every 10 minutes, with explicit leaf and canopy-level dynamics and daily updating of phenological variables and carbon pools. Thus, SiB4 has finer temporal resolution than either the CASA or VPRM models. However, given the computational cost associated with running this complex model with high temporal resolution, fluxes were estimated here at the spatial resolution of 0.5° using weighted average land-cover, which can partly account for sub-pixel variability in PFTs. The weighted land-cover approach (also implemented in VPRM) has been shown to improve performance of land surface models and to make model performance less sensitive to the spatial resolution of the estimates (Li et al, 2013).

Unlike previous versions of SiB, SiB4 (Haynes et al., 2019) has prognostic phenology, with internally calculated leaf area temporal dynamics. This prognostic phenology scheme has been shown to perform well in croplands across distinct crop types (i.e., winter and spring wheat, corn, soybeans and generic C3 and C4 crops; Lokupitiya et al., 2009) and grasslands (Haynes et al., 2019), two ecosystems with sharp seasonal transitions that may be difficult to detect with remote-sensing data having sparse sampling frequency (Guindin-Garcia et al., 2012). However, predicted phenology can still become decoupled from actual phenology due to unmodeled management effects (e.g., fertilization, irrigation, planting and harvest), and other non-climatic factors such as disturbances. Table S1 shows other details of the SiB4 implementation used here.

2.2 Model evaluation with observations

In order to evaluate model performance, NEE flux estimates from the three VPRM versions, CASA and SiB4 are compared to two kinds of observations: flux tower observations of NEE at 22 flux towers (shown in Figure 1) and atmospheric CO_2 measurements at 21 towers in the domain (Figure 2, Table S5).

Flux towers directly measure NEE in relatively homogeneous terrain in a ~1 km^2 footprint around each tower, and thus can be used for biospheric model evaluation. Towers are selected here to have relatively complete data coverage in 2016/ 2017, but still only cover only a small fraction of area in the domain given the limited number of sites and their small footprint, which is not necessarily representative of overall land cover patterns. However, a range of PFT's across the selected towers helps to represent land cover heterogeneity in the domain, with towers sited in deciduous broadleaf forest and mixed forest (6), evergreen needleleaf forest (4), wetlands (3), croplands (4) and grasslands (4). Modeled NEE estimates are extracted in 2016/ 2017 at each flux tower location at the finest possible spatial scale to match the flux tower footprint: 0.02° for VPRM, 500 m for CASA in the coterminous USA and 5 km in Canada, and 0.5° for SiB4, but for the specific PFT of the flux tower. Comparisons of flux estimates to observations across towers are principally made to assess biases in the seasonal and diurnal cycles corresponding to each model. For individual towers, mismatches may occur due to the spatial scale mismatch and errors in the underlying land cover maps for each model, although previous work comparing regional and site-specific biospheric model runs to observations at flux tower sites has found that most of the mismatch occurs because of model structure, parameters and inputs, not differences in spatial scale (Raczka et al., 2013).

A complementary analysis was also performed using atmospheric CO_2 mole fractions measured at 21 surface towers (Figure 2, 3a; Table S5) to help evaluate biospheric model performance, as these observations reflect the influence of regional CO_2 fluxes diluted by atmospheric mixing and transport. The footprint (or sensitivity to fluxes in space and time) of each atmospheric observation varies by tower location, inlet height and weather patterns, although the towers generally see the influence of high-resolution fluxes (in space and time) near the tower, and a more diffuse signal coming from farther away. With changes in wind direction and synoptic weather conditions, what towers "see" in the near-field may also be sparse and variable, but on average, most parts of the domain and the full flux diurnal cycle are sampled, albeit intermittently (Figure 2). Thus, atmospheric CO_2 data can help to evaluate CO_2 flux estimates at coarser spatial and temporal scales over more parts of the domain compared to the flux tower comparison, although atmospheric measurements are not a direct measurement of CO_2 flux and the comparison is subject to errors in other components (i.e., transport, background, fossil fuel emissions) in the analysis, as described below.

2.2.1 Atmospheric CO_2 tower observations, footprints and fossil fuel emission products

In order to simulate atmospheric CO_2 mole fractions for comparison with observations, NEE flux estimates from the models and fossil fuel emission estimates at 0.1° are multiplied by footprints from an atmospheric transport model (Lin et al., 2003). This multiplication is referred to as a convolution where flux units of $\mu mol * m^{-2} * s^{-1}$ are converted to $\mu mol/\ mol$, representing the mole fraction enhancement of CO_2 at the observation location and time due to modeled fluxes in the domain. Fossil fuel (FF) emissions are similarly convolved with footprints and then subtracted from CO_2 observations, to remove the enhancement due to FF and isolate the biospheric influence in the atmosphere. Finally, the influence of background conditions (CO_2 in air masses originating outside the domain) is also subtracted from observations in order to isolate the influence of biospheric fluxes occurring solely within the domain of interest. Throughout the study, we thus compare convolved NEE fluxes (or

simulated CO_2 enhancements) with observed biologic enhancements, defined as total observed CO_2 mole fractions – FF convolutions – background influence. We use the term "enhancement" here, although technically these can be other positive or negative due to both biospheric sources and sinks.

Hourly average atmospheric CO_2 observations are used here from 21 towers in our domain (Figures 2 and 3, Table S5), which are primarily sited in rural areas and from a variety of data providers (Karion et al., 2020; Miles et al., 2018; Mitchell et al., 2019; NOAA ESRL, 2019; Richardson et al., 2017). Other potential towers were excluded from the study due to challenging topography for modeling transport or urban locations where fossil fuel emissions have a stronger impact on observations. We use CO_2 observations at the tallest sampling inlet on each tower during afternoon hours for the analysis, with "afternoon" as defined in Section 2 of the supplemental material.

Footprints corresponding to each atmospheric observation were generated from two different transport models: WRF-STILT and NAMS-STILT. WRF (Weather Research and Forecasting model, Skamarock et al., 2008)) was run with a 1 km, 3 km, and 9 km nest, with the finer scale nests centered around the Washington DC/ Baltimore area (Figure 2). NAMS (North American Mesoscale System; NCEI et al., 2020) is a meteorological product with a spatial resolution of 12 km made publicly available by NOAA/ARL ((ftp://arlftp.arlhq.noaa.gov/nams). The Stochastic Time-Inverted Lagrangian Transport model (STILT, Lin et al., 2003) was then used to generate footprints on a 0.1° grid across the domain by releasing particles from each observation point and tracing them back in time and space using winds from the two different meteorological products. Further details of the custom WRF runs and footprint generation is included in Section 3 of the supplemental material.

WRF-STILT with its custom setup and higher spatial resolution may help to better model transport in the Appalachian mountain range that crosses most of our domain (Pillai et al., 2011) and within the two inner nests from Washington D.C. to Philadelphia, as compared to NAMS-STILT. However, without a more in-depth study evaluating the two transport models in our domain, it is difficult to know which set of footprints is more accurate in different regions and at different times of the year. Averaging across process-based models is known to help reduce the influence of systematic and random errors (Elder, 2018); therefore, for this study, we average convolutions using WRF-STILT and NAMS-STILT footprints and present these averaged results in the main text. Corresponding figures using convolutions from just WRF-STILT or NAMS-STILT footprints alone are included in the supplemental material, and results from the atmospheric CO_2 comparison are only highlighted which are robust across both transport models.

In order to remove the influence of fossil fuel emissions from atmospheric CO_2 observations, we pre-subtract convolved fossil fuel flux estimates from the Vulcan 3.0 product (Gurney et al., 2020) in the USA merged with the FFDAS product (Asefi-Najafabady et al., 2014) in Canada. Both products are defined hourly for the year 2015, with FFDAS at 0.1° and Vulcan 3.0 at 1 km resolution. The merged fossil fuel product at 0.1° is adjusted to match the days of week in our year of interest (2016/ 2017), given that fossil fuel emissions are known to behave differently on weekdays vs. weekends (Gurney et al., 2020).

Uncertainty associated with fossil fuel emission estimates is generally considered to be lower than that from biospheric flux estimates (Lauvaux et al., 2021), especially during the growing season; however, emission magnitudes and their fine spatiotemporal patterns are not perfectly known, especially if the emission product was developed for a year other than the one of interest, as in this study. In order to simplify the analysis, we pick what we consider to be the highest-quality emission product defined at fine spatiotemporal scales in our domain for the atmospheric CO_2 comparison. This choice is supported

by minimal differences between fossil fuel convolutions using different emission products (e.g., FFDAS in the USA rather than Vulcan 3.0) compared to the spread associated with varying transport, biospheric model and background conditions, as seen in other studies in eastern North America as well (e.g., Martin et al., 2019). However in winter, fossil fuel and biospheric enhancements for towers in the domain are similar in magnitude and have the same sign (Figure 3c); thus small errors in the emissions estimates could have a larger relative impact on the CO_2 analysis in winter months.

2.2.2 Background conditions

Atmospheric CO_2 observations at the towers are influenced by CO_2 fluxes occurring within the domain of interest, as well as by air masses flowing into the domain from outside (predominantly from the west and northwest in this case.) Thus, in order to factor out the "background" influence from atmospheric CO_2 observations (Karion et al., 2021; Mueller et al., 2018), STILT virtual particles are traced backwards from each observation location and time period to the points at which they exit the domain, and then a 4-dimensional CO_2 mole fraction field is sampled and averaged at those exit locations and time periods. Modeled CO_2 mole fractions at exit points are taken from two optimized data assimilation products for comparison: CarbonTracker v2019B (CT19B; Jacobson et al., 2020) and CarbonTracker Europe (CTE; Peters et al., 2010).

Background conditions at the towers from the two products (CT19B and CTE) differ throughout the year, with mean monthly differences ranging from 0.5 µmol/mol (or parts per million, ppm) in November and December to 1.4 µmol/mol in August and 1.8 µmol/mol in September (Figure 3b), in part due to differences in the underlying biospheric models used as priors in each optimization system. In fact, the difference between the background conditions is larger in magnitude than the biologic enhancements for about 23 % of observations in winter months (November to April) and 12 % of observations from May to October, with even higher percentages in early spring (e.g., 37 % in March and 29 % in April).

Therefore, in order to minimize biases associated with background conditions (and better isolate the influence of biospheric fluxes occurring inside the domain), we compare modeled atmospheric CO_2 at the surface level (i.e., the total mole fraction, not their background contribution) from CT19B and CTE and their mean to hourly afternoon observations at the tower locations in our year of interest. We then visually inspect the monthly mean biases across towers to select the product with the smallest bias and lowest spread (Figure S4) in each month. Only surface CO_2 is compared here, whereas many particles exit the domain at higher altitudes and in the free troposphere; also, errors in CT19B or CTE fluxes inside the domain could affect these model-data comparisons. However, 1) the CarbonTracker fluxes are optimized with atmospheric CO_2 data and 2) systematic biases across towers are likely to be at least in part influenced by background air flowing into the domain.

This analysis resulted in selecting CT19B in July, CTE in October and the mean of the background products in all other months as "optimal", with these monthly selections then used throughout the atmospheric CO_2 comparisons. Using the mean of the two products in most months helps to cancel out opposing biases, especially during the growing season, although residual biases in all months will still affect the atmospheric CO_2 comparisons.

2.2.3 Statistical metrics

To compare variability in model simulations to flux tower NEE or biologic atmospheric CO_2 enhancements, we use the coefficient of determination (R^2) and the Nash-Sutcliffe coefficient (or NSC, Moriasi et al., 2007):

$$NSC = 1 - \frac{\sum_{i=1}^{n}(pred_i - obs_i)^2}{\sum_{i=1}^{n}\left(obs_i - \overline{obs}\right)^2}.$$

The R^2 is used to assess how much variability in the observations (obs_i) can be explained by model predictions ($pred_i$), after accounting for any biases. The NSC is calculated using the same equation as a coefficient of determination (R^2), but instead of using a fitted regression model (that is guaranteed to have zero mean bias), the NSC metric uses the actual model simulations, or in this case, the NEE convolutions with atmospheric footprints. Therefore, the NSC can vary from -∞ to 1, with values < 0 indicating that the model performs worse than the observational mean for prediction (due to biases in the model), and values > 0 implying a better fit than the mean. Thus, this metric allows us to assess variability, while also penalizing model fits that have particularly biased flux estimates. We also include comparisons between hourly convolutions and observations using other statistical metrics (like root-mean squared error) by tower and season in the supplemental material.

3. Results and discussion

3.1 VPRM respiration model improvement

The modified respiration model in $VPRM_{new}$ (dark green points in Figures 4, S5) is shown to substantially improve the fit of predicted R_e fluxes to nighttime average NEE (or R_e) observations, as compared to the fit using annual ($VPRM_{ann}$) or seasonal ($VPRM_{seas}$) parameters with the original model formulation. In the temperate, humid climate that covers most of our domain, including EVI was seen to be the single most important factor improving model fit for $VPRM_{new}$ across all PFTs, except in evergreen needleleaf and mixed forests (Table S6). This is consistent with the fact that canopy development during the growing season for crops, deciduous forests and other non-evergreen ecosystems induces large increases in autotrophic respiration. In particular, the cropland PFTs (maize and soybean/ other crops) have a dramatic improvement in model performance with $VPRM_{new}$, where e.g., the NSC in the maize PFT for $VPRM_{new}$ is 0.62 compared to 0.28 for $VPRM_{seas}$ and 0.21 for $VPRM_{ann}$. In more water-limited ecosystems (e.g., shrublands) and times of the year (i.e., late summer, early autumn), the water stress scaling factor and its interactions with temperature also help to improve performance. For example, the NSC for shrublands with $VPRM_{new}$ is 0.33, while for $VPRM_{ann}$ and $VPRM_{seas}$ the NSC is negative, implying that these latter models perform worse than using the observational mean for prediction.

In forested ecosystems, $VPRM_{new}$ also has better performance than either $VPRM_{ann}$ or $VPRM_{seas}$, but the increase in NSC is less dramatic. This may be because seasonal increases in leaf area (particularly in evergreen forests) and/ or water stress are relatively less important predictors of R_e fluxes in these ecosystems compared to air temperature. In deciduous broadleaf forests and southern evergreen and mixed forests, relatively low NSC values even with $VPRM_{new}$ (0.33 and 0.19 respectively), may be because the respiration model does not account for inputs to dead carbon pools (e.g. from leaf litter at the end of the season or events like storms and logging).

In contrast to $VPRM_{new}$, night-time respiration estimates using the standard VPRM model with annual parameters (i.e., $VPRM_{ann}$, yellow lines in Figures 4 and S5) are biased low in all PFTs at the highest temperatures, as compared to historical NEE observations and to a linear regression model fit to only nighttime data (purple lines). This is because the only seasonality in the original respiration model comes from temperature, which is not enough to explain changes in R_e associated with seasonality in biomass. Also, there is no guarantee that $VPRM_{ann}$ will produce unbiased flux estimates for any portion of the diurnal or seasonal cycle, as explained in Section 2.1.1. NSC values for $VPRM_{ann}$ are lowest among the three VPRMs across all PFTs.

$VPRM_{seas}$ shows slightly higher NSC values than $VPRM_{ann}$ for all PFTs, in part by estimating a different baseline respiration in each season. In particular, the bias at high temperatures seen in $VPRM_{ann}$ is alleviated with seasonal parameters (i.e., light green lines at high temperatures in Figures 4 and S5). Spring and fall show similarly positive relationships between R_e and temperature for most PFTs, but a negative relationship is inferred in summer for 7 of 8 PFTs (with evergreen needleleaf and mixed forests > 40°N the only exception; Table S3). This relationship apparent in the data shows that water limitations play a role in limiting R_e at the height of the growing season when soil moisture has been depleted by spring and early-summer growth. Nevertheless, the negative inferred relationship between temperature and R_e in summer is unlikely to explain the relationship well in areas and time periods that are not water-limited, and therefore provides further justification for an improved respiration model that can include additional mechanistic detail.

3.2 Gridded CO_2 flux comparison across models

Flux estimates for GPP, R_e and NEE are first compared across the full domain by examining gridded fluxes (at 0.1°) averaged across 3-monthly seasons (i.e., December/ January/ February, or DJF, March/ April/ May, or MAM, June/ July/ August, or JJA, and September/ October/ November, or SON). SiB4, CASA and $VPRM_{new}$ fluxes are principally compared in the main text (with corresponding plots comparing $VPRM_{new}$, $VPRM_{seas}$ and $VPRM_{ann}$ in the supplemental material.) The seasonal and diurnal cycles are then compared for two aggregated spatial groupings: deciduous broadleaf forest and croplands (Figure 1, right panel; Figure S6). Together, these two land cover types make up about half of the land area in the domain and are the predominant land covers upwind of the Northeast Corridor, thus disproportionately influencing atmospheric CO_2 in many of the urbanized areas of the domain. Moreover, the Corn Belt is known to be one of the most biologically productive ecosystems on Earth during the height of the growing season (Gray et al., 2014; Hilton et al., 2017; Zeng et al., 2014), and therefore has a strong influence on CO_2 across the domain, especially in June, July and August.

3.2.1 GPP, R_e and NEE spatial patterns by season

Gridded 3-monthly mean GPP, R_e and NEE vary across models in terms of flux magnitude and spatial patterns (and with SiB4 having a coarser spatial resolution; Figures 5, 6, 7, 8, S7, S8, S9, S10). In winter and spring, all the models generally capture the north-south gradient with dormant conditions in the north and a more active biosphere in the south of the domain, whereas in summer and fall, the spatial patterns are more distinct across models (Figure S11). Spatial patterns are, not surprisingly, influenced by the underlying land cover maps for each model, such that e.g., in summer, SiB4 has a larger area of

peak uptake in croplands relative to the other models (Figure 7), given more extensive cropping areas in Michigan, Wisconsin and the Carolinas (Figure S6). Also, CASA has strong R_e fluxes in the spring and fall across large parts of the domain (Figures 6 and 8), in areas principally defined as deciduous broadleaf forests, which have a larger spatial extent in this model compared to the definitions for VPRM and SiB4 (Figure S1). The higher R_e fluxes for CASA in spring across most parts of the domain are also related to differences in timing across the models, as discussed further in Section 3.2.2.

In summer, despite differences in flux magnitude, spatial patterns for GPP, R_e and NEE fluxes are the most correlated between SiB4 and $VPRM_{new}$, with both capturing strong net uptake in the midwestern Corn Belt extending north into Wisconsin, neutral uptake in the Mississippi River valley, Alabama and Georgia, and strong uptake in the northeastern USA forests (Figure 7, Figure S11). CASA also has strong net uptake in cropping areas from the Corn Belt through the Mississippi River valley, but GPP is relatively homogeneous throughout the domain and R_e is slightly lower in croplands compared to other areas, unlike in $VPRM_{new}$ and SiB4 which show the opposite pattern. Given that CASA and $VPRM_{new}$ share the same light-use efficiency formulation for photosynthesis, whereas SiB4 uses a more physiologically realistic enzyme-kinetic formulation, the closer correspondence in GPP spatial patterns between SiB4 and $VPRM_{new}$ in summer show that spatial patterns may be more influenced by the underlying land cover maps, PFT-specific parameters and sub-pixel weighting in each model than the underlying photosynthesis algorithm.

In addition to a larger cropland spatial extent, SiB4 fluxes also differ in timing in cropping areas compared to the other models. In the Mississippi River Valley which has substantial soybean production, SiB4 shows an earlier start to the cropping season in spring (Figure 6) and an earlier end in summer (Figure 7). In the fall, SiB4 GPP has already shut down in all cropping areas, while R_e continues at a reduced rate in the Midwestern Corn Belt, leading to large net sources from Indiana, Illinois and Ohio in these months (Figure 8), consistent with results from previous atmospheric CO_2 inversions in these areas (Gourdji et al., 2012). CASA and $VPRM_{new}$ also show net sources to the atmosphere from cropping areas in October and November, although lower in magnitude and less evident at the 3-monthly average timescale shown here because cropland GPP in CASA extends through the end of September and the VPRM respiration models do not explicitly consider inputs of dead biomass to carbon pools in these months. Across the domain, SiB4 also shows the strongest R_e fluxes in the fall months from forested areas in Wisconsin and Pennsylvania, whereas both CASA and VPRM show stronger R_e fluxes towards the south of the domain, perhaps due to a stronger sensitivity of R_e to temperature in these models.

In winter, all the models capture the north to south gradient for GPP, with non-zero productivity in the south and correlations > 0.8 across models (Figure 5). Spatial patterns for winter R_e and NEE are more distinct, with CASA and $VPRM_{new}$ showing higher correlations than SiB4 and $VPRM_{new}$ (Figure S11). In this season, SiB4 shows relatively homogeneous R_e fluxes throughout the domain (as in fall), whereas CASA and $VPRM_{new}$ show reduced sources in the north compared to the south. NEE spatial patterns also show opposite north-south gradients between models, with stronger net sources for SiB4 in the north of the domain and for CASA and $VPRM_{new}$ in the southern half. CASA also shows anomalously low R_e and NEE fluxes in parts of Ontario during winter months, which could be related to the 5 km spatial resolution (along with the dominant land cover formulation) in this part of the domain.

Across the VPRM versions, correlations in GPP spatial patterns are > 0.95 in all seasons, whereas correlations in R_e spatial patterns are > 0.8, with slightly lower correlations in summer (Figures S7, S8, S9, S10). $VPRM_{seas}$ has the most distinct spatial patterns across the three versions of VPRM for R_e in summer and NEE in winter, and it also shows the strongest net sources in cropping areas in the fall months. Interestingly, the R_e spatial patterns for $VPRM_{new}$ in spring, summer and fall are marginally more correlated to those in CASA and SiB4 than are the correlations with $VPRM_{ann}$ or $VPRM_{seas}$ (Figure S11), providing some evidence of improvement in skill for the new respiration model.

Finally, the three biospheric models differ in how they represent fluxes in urban areas. CASA in particular shows weaker fluxes near cities (Figures 5, 6, 7, 8), e.g., in Indianapolis and Atlanta in summer, given that GPP and R_e are set to zero in this model for pixels at the native spatial resolution (500 m or 5 km) where urban is the dominant land-cover. $VPRM_{new}$ has somewhat lower R_e and GPP fluxes in urban areas due to lower EVI and the impervious surface correction to heterotrophic respiration in this model. In contrast, SiB4 does not have an urban PFT, and therefore does not show any reduced R_e or GPP fluxes in pixels dominated by urban land use.

3.2.2 **GPP, R_e and NEE seasonal and diurnal cycles**

CASA, SiB4 and the three versions of VPRM differ in the magnitude of GPP and R_e fluxes during the growing season (Figure 9) in both croplands and deciduous broadleaf forests. CASA and SiB4 have similarly strong GPP at their peak in both ecosystems, and then a lower peak magnitude (by about 25%) for $VPRM_{new}$ followed by $VPRM_{seas}$ and $VPRM_{ann}$ (about 30 to 40% lower). GPP timing is relatively consistent across the models with CASA having a slightly longer growing season, especially in croplands, where the peak occurs slightly earlier in CASA and SiB4 compared to the VPRM models. R_e fluxes also differ across models in terms of both magnitude and timing, with CASA having an earlier ramp-up and peak (by about a month) compared to the other models, and the strongest peak magnitude, followed by SiB4, $VPRM_{new}$, $VPRM_{seas}$ and then $VPRM_{ann}$.

For the NEE seasonal cycle during the growing season, the models generally agree well in terms of timing and magnitude in deciduous broadleaf forests, but they have more distinct patterns in croplands. SiB4 has a peak NEE uptake shifted about a month earlier compared to the other models, despite a GPP seasonal cycle similar in timing to CASA and a R_e seasonal cycle more similar to the VPRMs. In contrast, the similarity in NEE timing between CASA and the VPRMs in croplands, despite differences in timing for GPP and R_e, is likely due to some combination of both models using remote-sensing inputs and calibration with flux tower NEE.

$VPRM_{ann}$ and SiB4 have the strongest peak uptake in July and August in both ecosystems, and CASA and $VPRM_{new}$ the weakest (with $VPRM_{seas}$ in the middle.) It is interesting to note that $VPRM_{ann}$ has the strongest net uptake in summer months, despite having the lowest magnitude of GPP and R_e fluxes, which is because R_e in $VPRM_{ann}$ is most depressed relative to GPP during summer months. Also, CASA and SiB4 show stronger net sources to the atmosphere in winter than the VPRMs for both ecosystems, given the predominance of R_e in this season.

The mean diurnal cycle in July for GPP and R_e (Figure S12) shows a similar pattern as that seen for the seasonal cycle, i.e., CASA has the strongest component fluxes, then SiB4, followed by the three versions

of VPRM (with $VPRM_{new}$ stronger than $VPRM_{ann}$ and $VPRM_{seas}$). For the NEE diurnal cycle, CASA has by far the strongest net afternoon drawdown and nighttime sources compared to the other models, about 40% stronger than SiB4 and the VPRMs, which are more similar in shape. At night in this month, $VPRM_{new}$ has the second highest sources after CASA, followed by SiB4 and $VPRM_{seas}$, and then $VPRM_{ann}$.

Larger component fluxes for CASA and SiB4 compared to the VPRM models could be due to several reasons. First, CASA and SiB4 are (mostly) neutral biosphere models that start from equilibrium carbon pools in this year, which could lead to overestimation of R_e in areas with net sinks, e.g., in forests recovering from disturbances (e.g. storms, insect attacks or harvest) or in croplands where harvested products are transferred to other areas for consumption (Zhou et al., 2020). (As a side note, net sources and sinks in VPRM primarily reflect the influence of the flux tower observations used in the parameter optimization, which in this case results in a net annual sink in the domain 4.5 times greater than that for CASA and SiB4 in this year.) Secondly, the flux tower observations used for VPRM parameter optimization (especially at night) and/ or the partitioning algorithms for separating component fluxes could potentially be biased, leading to biased flux estimates (Aubinet, 2008; Lasslop et al., 2010; Reichstein et al., 2005; Yi et al., 2000). Third, the parameter optimization using 24 hours of NEE data for $VPRM_{seas}$ and $VPRM_{ann}$ does not ensure unbiased fluxes for any portion of the diurnal cycle, as mentioned previously.

The differing seasonal timing in the models are also likely explained by several reasons. First, the longer growing season for CASA in both GPP and R_e relative to the other models, is likely influenced by the monthly fPAR inputs used to drive the model, which are coarser in time than the 8-day average EVI in VPRM and daily internally-calculated Leaf Area Index in SiB4 (Figure S3; Zhou et al, 2020). Longer growing seasons for process-based models compared to observations at flux towers in deciduous forests was also seen in Richardson et al. (2012). The shorter growing seasons in VPRM could be more accurate, but could also be influenced by long gaps in the overlapping 16-day EVI composites for each sensor, due to the satellite passing too early or too late within adjacent composites (Corbin et al., 2010; Guindin-Garcia et al., 2012). GPP phenology could potentially be modeled most accurately with an algorithm incorporating remotely-sensed solar induced fluorescence (Joiner et al., 2011; Parazoo et al., 2019; Shiga et al., 2018; X. Wang et al., 2020), which does not rely solely on "greenness" to sense growing season transitions.

Uncertainty in the timing of R_e across models may be more difficult to diagnose, although the earlier onset of R_e in CASA compared to $VPRM_{new}$ and SiB4 is also seen in Messerschmidt et al (2013), with the latter timing more consistent with atmospheric column CO_2 and flux tower observations (Falge et al., 2002). In future work, the timing of GPP and R_e fluxes could potentially be investigated further by using additional atmospheric tracers like carbonyl sulfide measurements to separate GPP and R_e at regional scales (e.g., Berry et al., 2013; Hilton et al., 2017; Wang et al., 2016). The magnitude and timing of NEE across the models in this year will also be evaluated with comparison to flux tower data in Section 3.3 and to atmospheric CO_2 observations in Section 3.4.

3.3 Model evaluation with flux tower observations

The comparison of modeled hourly NEE to flux tower observations at 22 locations in 2016 - 2017 shows that all biospheric models perform relatively well at capturing observed seasonal and diurnal variability,

with mean NSC values for the full year > 0.5 for most towers and models, and with CASA performing marginally better at the hourly timescale than the VPRM models and SiB4, perhaps due to its finer spatial scale in the USA.

Mean biases in nighttime R_e fluxes across towers (Figure 10a) are consistently different across models, particularly during the height of the growing season, with $VPRM_{new}$ relatively unbiased throughout the year (albeit with a weak negative bias in June and July), $VPRM_{ann}$ biased low from April to October, and $VPRM_{seas}$ with intermediate negative biases. The negative biases in $VPRM_{ann}$ from June to August are large relative to R_e fluxes, i.e. ~2 to 4 μmol/m^2/s compared to a mean R_e of ~6 μmol/m^2/s in these months, while the negative biases are alleviated in $VPRM_{seas}$ and $VPRM_{new}$ due to higher baseline summertime respiration in $VPRM_{seas}$ and the EVI covariate in $VPRM_{new}$ (as discussed in Section 3.1). In contrast, nighttime R_e for CASA is biased high for most months of the year, particularly in April, May and June, consistent with findings in Zhou et al. (2020). SiB4 R_e fluxes at night are also biased high in winter from December to April, and with a large spread during the growing season but small negative biases in June and July.

Mean biases in daytime NEE fluxes (Figure 10b) show a large spread across towers for each model from April to November, with the three VPRM models showing similar weak source biases from June through October. Although 2016/ 2017 fluxes were left out of the VPRM parameter optimization, many of the same sites included in the evaluation shown here were also included in the historical optimization. Thus, the daytime source biases for VPRM at the evaluation towers may be worsened by the spatial scale mismatch and mixed land-cover in the full 0.02° pixel, which may be less productive overall than the ~1 km^2 flux tower footprint. CASA and SiB4 both have daytime source biases in winter months (December to March), consistent with the nighttime R_e biases in these same months. CASA also shows a sink bias at most towers in September and October, perhaps due to an overly extended growing season in this model associated with monthly phenology based on fPAR (Figure S3; Figure S9 in Zhou et al., 2020).

The mean diurnal cycle in July for models and observations is also shown for two specific towers: US-IB1, a cropland site in Illinois growing corn in this year, and US-UMB, a deciduous broadleaf forest site in the northern lower peninsula of Michigan. At US-IB1 (Figure 10c), SiB4 overestimates peak uptake in the early afternoon by about 20%, whereas the VPRM models underestimate it by about 50%. CASA most closely matches the magnitude and timing of daytime drawdown, albeit with a slight underestimate of peak uptake. For VPRM, the spatial scale mismatch likely affects this comparison, as this 0.02° pixel is only 32% corn, but also 44% other crops (which have weaker uptake), and 24% other PFT's. The CASA landcover for this 500 m pixel is cropland, but CASA also parameterizes a single crop type, thus not allowing for the relatively stronger uptake in some corn fields relative to soybean or other crops, which is less apparent at this particular tower. At night, CASA overestimates R_e fluxes for this tower and month, whereas $VPRM_{new}$ closely matches nighttime R_e and $VPRM_{seas}$, $VPRM_{ann}$ and SiB4 slightly underestimate it.

At US-UMB (Figure 10d), the magnitude of the diurnal cycle is overestimated for CASA during both nighttime and mid-day, whereas the daytime uptake for the VPRM models and SiB4 is underestimated by about 20%. At night, $VPRM_{new}$, $VPRM_{seas}$ and SiB4 closely match the magnitude of observed R_e, whereas $VPRM_{ann}$ underestimates R_e by about 60%. The period of peak mid-day uptake in SiB4 is longer

than in the observations and the other models, thus leading to a similar total uptake during daylight hours compared to the observations and more compared to the VPRM models. The spatial scale mismatch could also contribute to the daytime mismatch for the VPRM models at this tower, although to a lesser extent than at US-IB1, given that the 0.02° pixel surrounding US-UMB is 62% deciduous broadleaf forest, but also 12% wetlands and open water.

3.4 Model evaluation with atmospheric CO_2 observations

Observed total CO_2 mole fractions (Figure 3a) and biologic enhancements (Figure 3b) show both a strong seasonal cycle and a large spatial variability across towers. Mean integrated footprints summed across towers in July show that most of the domain is "seen" on average by this network of towers (Figure 2, left panel), with slightly less sensitivity along the edges of the domain, and stronger sensitivities in summer compared to winter (when wind speeds are higher). Some towers have a stronger influence from croplands (e.g., BRI, TPD, S01), wetlands (e.g., LEF, SCT), forests (e.g., AMT, DNH, HAF) or urban areas (e.g., MSH, HCT), while almost all towers have some sensitivity to deciduous broadleaf forests, crops and grass/pasture (including developed-open space), showing the broad influence of these land covers throughout the domain (Figure 1, Table S5). The tower network is also sensitive to both day and nighttime fluxes on average, with afternoon receptors typically sensitive to nearby fluxes in the morning of the same day and previous night (Figure 2, right panel).

Simulated and observed biologic atmospheric CO_2 enhancements are compared across towers in several different ways in order to evaluate both bias and variability. First, we compare weekly mean enhancements throughout the year at two specific towers, with one predominantly influenced by croplands (S01 in Indiana) and the other by deciduous and mixed forests (DNH in New Hampshire), the two predominant land covers in the domain upwind of the NEC (Figure 1; Table S5). Then, monthly mean biases are examined across all towers to see how errors in the biospheric models (in terms of phenology, flux magnitudes and spatial patterns) translate into biased CO_2 across the domain and throughout the year. Next, we compare hourly variability in the afternoon enhancements across towers within each month, which tests the biospheric models' ability to match the diurnal cycle, synoptic variability, sub-monthly seasonality and spatial gradients seen in the observations at discrete time periods. These analyses are meant to inform how incorrect representations in the biospheric models would affect inversions estimating fossil fuel emissions, with biased biospheric flux estimates translating directly into biased emission estimates at averaged scales, and incorrect fine-scale variability (both spatial and temporal) leading to emission estimates incorrectly attributed in space and time.

3.4.1 Weekly mean comparison at S01 and DNH

Observed weekly mean biologic CO_2 enhancements at the S01 (Indiana) and DNH (New Hampshire) towers show net uptake from the atmosphere from May through October (Figure 11), and net releases in the rest of the year. At the S01 tower, which samples CO_2 from upwind crops, observations show a narrow and strong peak drawdown briefly reaching -19 µmol/mol for one week in mid-July, whereas at DNH, influenced primarily by deciduous and mixed forests (Table S5), observations show a longer period of peak uptake (about 2 months in June and July) which only reaches ~-8 µmol/mol. All biospheric

models (CASA, SiB4 and $VPRM_{new}$) capture the broad seasonality seen in the observations, although with differences in timing and magnitude.

At S01, all models perform reasonably well during the growing season in matching weekly variability, with $VPRM_{new}$ having a slightly higher adjusted R^2 compared to the other models (i.e., 0.82 compared to 0.78 for CASA and 0.73 for SiB4). SiB4 best captures the peak drawdown in July with averaged WRF and NAMS-STILT convolutions and WRF-STILT alone but is too strong with NAMS-STILT convolutions (Figure 11a, Figure S13). The SiB4 convolutions also match the timing of the observed drawdown well in May, June and July, but are too weak in August and September, and transition to a net source a couple weeks early compared to observations in the fall. For CASA and $VPRM_{new}$, the growing season drawdown starts a few weeks late in June, but both models match the observed timing well from August into September. In the dormant season from November to April, SiB4 modeled enhancements show the closest correspondence to observations in terms of both the NSC and adjusted R^2 (Figure 11, left panel).

At DNH, $VPRM_{new}$ better explains weekly variability during the growing season substantially compared to the other models (with an adjusted R^2 of 0.80 for $VPRM_{new}$, compared to 0.55 and 0.50 for CASA and SiB4). In the winter months, SiB4 best captures the weekly variability (adjusted R^2 of 0.47 compared to 0.30 for $VPRM_{new}$ and 0.13 for CASA) although with the lowest NSC values due to flux estimates which are biased high (Figure 11b). Both CASA and SiB4 start the growing season slightly late in May and end the growing season too late in September and October compared to observations, whereas $VPRM_{new}$ matches the timing of the observations much more closely in these seasonal transition months. In June and July, $VPRM_{new}$ and SiB4 have too much net uptake compared to observations, although the magnitude of peak uptake is sensitive to which transport model is used (Figure S14). At DNH, NAMS-STILT transport makes all biospheric models look more biased in June and July, perhaps pointing to errors in summertime footprint strength upwind of this tower. In general, the summertime biases in the biospheric models at both S01 and DNH are sensitive to transport, demonstrating, as in Feng et al. (2019), that transport model uncertainty tends to amplify biospheric model uncertainty during the growing season, although biases in the dormant season for each model are also seen to be sensitive to transport here.

3.4.2 **Monthly mean biases across towers**

Simulated biologic CO_2 enhancements show seasonal biases across towers and biospheric models, with all three VPRMs biased low in December and January, SiB4 and CASA biased high in February and May, $VPRM_{ann}$ biased low from June to August, SiB4 and $VPRM_{seas}$ biased low in July, and all models biased somewhat low in October (Figure 12, Table 2). On the positive side, $VPRM_{new}$ and CASA both look relatively unbiased during summer months (i.e., June to August, albeit with some sensitivity to transport, Figure S15), and all the VPRM versions look relatively unbiased from February to May and in October and November, pointing to skill in simulating growing season transitions. SiB4 is the least biased model on average across towers in December and January, although it also has the largest spread in biases throughout the year, particularly in April, July and September, which could be associated with its coarse spatial resolution in this study.

The biases in atmospheric enhancements seen here for each of the biospheric models (which could also be influenced by biased transport, background and/ or fossil fuel emissions) point to weaknesses in each

of the biospheric model setups. First, it appears that the low bias in $VPRM_{ann}$ in summer months is likely due to not enough increase in seasonal R_e in this model, particularly at night (as seen in the flux tower comparison, Figure 10), with this problem alleviated by the use of seasonal parameters in $VPRM_{seas}$ or the EVI covariate in the new respiration model (for $VPRM_{new}$). This result also shows the strong influence of nighttime fluxes on afternoon CO_2 observations (Figure 2), given that the VPRM models differ most in terms of R_e fluxes, which dominate the signal at night (Figure 9), as also seen in Hu et al. (in review) and T. Lauvaux et al. (2008, 2012). The negative bias for $VPRM_{seas}$ in July could be due to the negative relationship between temperature and R_e inferred for this model setup, which unrealistically lowers R_e in this month.

Secondly, the negative bias in flux estimates in December and January across all VPRM setups could point to a bias in the flux tower nighttime NEE observations used for optimization (Barr et al., 2013). While systematic errors in other components of the analysis, i.e., transport, background conditions, fossil fuels, cannot be ruled out, a bias in flux tower observations themselves is supported by studies suggesting that drainage loss due to horizontal advection in eddy-covariance systems, particularly at night with a stratified stable boundary layer, is non-negligible and represents flux to the atmosphere that is not measured (Aubinet, 2008; Nicolini et al., 2018). In addition, the biases seen in the VPRM enhancements in these months is evident with both sets of footprints and background conditions, as well as with fossil fuel convolutions using FFDAS across the entire domain. If the problem is in fact an observational bias in the flux tower data, this could point to systematic underestimation of respiration in VPRM throughout the year, which would be more difficult to detect with atmospheric CO_2 in months when GPP fluxes are stronger.

The biases in atmospheric CO_2 seen for CASA and SiB4 also have several plausible mechanisms. The positive biases in February and May and for most towers in November and March (also Figure S16) are likely due to over-estimated R_e in these models associated with balanced biosphere assumptions which do not account for vegetative and soil carbon sinks (Haynes et al., 2019; Zhou et al., 2020). Similarly, Zhou et al. (2020) also found an NEE source bias in the CASA ensemble mean compared to flux tower data from November to March in all biome types. The biases in May and October for SiB4 and CASA also point to difficulties in simulating phenology during seasonal transition months, but potentially also errors in the relative extent of crops and forests in each model (Figure S6), given that these two land cover types have different timings of seasonal drawdown (Figure 9). Difficulty simulating autumn phenological transitions in these two process-based models, particularly in deciduous forest ecosystems (e.g., near DNH), is also consistent with results from the model inter-comparison study of Richardson et al. (2012).

For SiB4, the biases in July, August and September are especially strong for the three towers with large cropland influence (Figure S16; Table S5), i.e., TPD and BRI, with negative biases in July of -5 µmol/mol and -7 µmol/mol respectively, and S01, with positive biases of 4 µmol/mol in August and September. This could point to errors in cropland fluxes for SiB4 associated with an over-estimate of the uptake rate, the prognostic phenology and/ or their spatial extent (Figure 7, Figure S6). These errors then propagate throughout the domain, given that all towers have some amount of cropland influence in their footprints (Table S5) due to the horizontal advection of air masses towards the east of the domain. CASA and $VPRM_{new}$ also show anomalous behavior at these same towers during summer months, which points to the difficulty in accurately simulating the strong drawdown in the Corn Belt for crop PFTs across models.

3.4.3 **Comparison of hourly CO_2 variability across towers**

The hourly comparisons of simulated to observed biologic CO_2 enhancements across towers by month show that $VPRM_{new}$ generally outperforms all other models in reproducing CO_2 variability during the growing season in terms of both NSC's and R^2s, with an average NSC from May to October of 0.38 compared to 0.32 for $VPRM_{seas}$, 0.25 for $VPRM_{ann}$, 0.24 for CASA, and 0.05 for SiB4 (Figures 13, S17). Tower-specific RMSE, NSC and R^2 metrics calculated with hourly enhancements (Tables S7, S8, S9, S10) are also highest for $VPRM_{new}$ at most towers in the spring, summer and fall, and even in winter with the R^2 metric, although CASA is least biased in this season.

All three versions of VPRM also have significantly higher R^2 values compared to CASA and SiB4 in May, June, September and October, and higher NSC values in May, September and October (due to the summer-time sink biases in $VPRM_{ann}$). This result points to strengths in VPRM across versions associated with the high temporal resolution phenology and potentially the domain-specific parameter optimization for this study, and not just the improved R_e model in $VPRM_{new}$. In the winter months, performance across the VPRM versions and CASA is more equivalent, with some month-to-month variation, and SiB4 again showing the lowest correspondence with observations. For SiB4, NSC values are negative in February, April, July, September and October and R^2s are < 0.1 in April, September and October, pointing to problems simulating phenology and high-resolution spatial variability in this model, as well as having biased flux estimates, e.g., in July. CASA NSC values are positive for all months of the year except February.

For all biospheric models, the NSC and adjusted R^2 metrics are somewhat higher during the growing season from May to October when the biospheric signal is stronger (Figures 3b, 3d), compared to winter months (November to February), when the uncertainties associated with transport, fossil fuel emissions and background become proportionally larger due to weaker biologic enhancements. In fact, the mean biologic enhancements in January across towers are of similar magnitude to the fossil fuel emission enhancements (Figure 3c), given the large extent of the domain over which R_e fluxes accumulate in the atmosphere. The biospheric models also better explain hourly variability in atmospheric CO_2 when biases are accounted for, as seen by higher R^2 values across months compared to NSC values (Figure 13). However, NSC and adjusted R^2 values are no higher than 0.4 and 0.5 respectively in any month throughout the year, pointing to substantial unexplained variability in modeled CO_2 compared to observations. The unexplained variability could be due to errors in biospheric flux estimates, but also from errors in transport, boundary conditions and/ or fossil fuel emissions.

Interestingly, the NSC and adjusted R^2 metrics are higher when using averaged WRF-STILT and NAMS-STILT convolutions, compared to using either transport model alone (Figure S17); e.g., the mean $VPRM_{new}$ NSC from May to October goes up to 0.38 with mean transport convolutions compared to 0.33 with WRF-STILT and 0.29 with NAMS-STILT transport alone, and up to 0.13 with the mean convolutions from November to April, compared to 0.05 using WRF-STILT and NAMS-STILT footprints alone. This suggests that convolution averaging helps to reduce random errors and potentially cancel out some biases in the transport simulations. NSC values calculated using NAMS-STILT in general are lower during the growing season compared to the metrics with WRF-STILT for all biospheric models (Figure S17), which points to potential biases in footprint strength in NAMS-STILT during these months, as also seen in the CO_2 comparison at DNH.

4. Conclusions, future work and recommendations

Three versions of VPRM were run for this study in a single year from November 2016 to October 2017 using customized parameters for an eastern North American domain: i) the original Mahadevan et al. (2008) implementation with annual parameters ($VPRM_{ann}$), ii) the same model formulation with seasonally-varying parameters ($VPRM_{seas}$), and iii) a modified respiration model that accounts for seasonality in biomass and water stress ($VPRM_{new}$). Flux estimates from the three VPRM versions were compared to flux tower observations and atmospheric CO_2 observations at 21 towers across the domain, and with flux estimates from the CASA and SiB4 biospheric models in this year.

Results show that the new respiration model in $VPRM_{new}$ increases the magnitude of nighttime R_e fluxes during the growing season, and thereby helps to realistically strengthen the diurnal and seasonal cycles of NEE compared to using the original respiration model which has a baseline value and a simple linear function of temperature. $VPRM_{new}$ thus better reproduces spatiotemporal variability in hourly atmospheric CO_2 observations, in part due to the strong influence of nighttime fluxes on afternoon CO_2 enhancements. In contrast, using annual parameters with the original VPRM model is seen to underestimate R_e during the height of the growing season, which results in a sink bias relative to atmospheric observations in these months. Using seasonal parameters removes the summertime bias but also puts an unrealistic negative relationship between temperature and R_e for 7 of 8 PFTs in this season.

The inter-comparison of the VPRMs with SiB4 and CASA sheds some light on the relative strengths and weaknesses of each model, at least as seen in this single year. Overall, $VPRM_{new}$ and $VPRM_{seas}$ are less biased and better explain the variability in hourly atmospheric CO_2 during the growing season compared to CASA and SiB4. The strong diurnal cycle in CASA NEE is seen to be somewhat overestimated compared to flux tower observations during the growing season, and with R_e fluxes overestimated year-round. However, CASA NEE fluxes are relatively unbiased compared to afternoon atmospheric CO_2 observations at the height of the growing season, although the growing season overly extended into the spring and fall. SiB4 has the most mechanistic detail among the three models but does not closely match observed variability in the atmosphere in this particular year. This could be because of its coarse spatial resolution in this study, its prognostic phenology (with errors in the timing of either GPP or R_e), or errors in the underlying land cover maps.

The strong performance from a careful implementation of a relatively simple model like VPRM shown here suggests that some of the mechanistic detail in more complex models may not be needed to represent fine spatiotemporal variability of CO_2 in the atmosphere due to the terrestrial biosphere (as also found in Raczka et al., 2013; Schwalm et al., 2010). Across all versions of VPRM, the custom parameter optimization using domain-specific historical flux tower data, fine temporal resolution observed phenology, high spatial resolution flux estimation with weighted average PFTs and high-resolution met drivers all help to improve model performance. Each of these factors could help to explain the improved performance of $VPRM_{new}$ relative to CASA and SiB4, with these latter models likely having more realistic model structures, particularly for R_e, but coarser spatial or temporal resolutions. At the height of the growing season however, none of the models explain more than 50% of the variability in atmospheric observations, which could be due to errors in modeled transport, background conditions, or fossil fuel estimates, but also due to errors in NEE estimates across biospheric models.

Potential future improvements to VPRM include i) incorporating SIF into the GPP equation (e.g., Luus & Lin, 2015; Turner et al., 2020a) to better simulate growing season transitions and water stress (when CO_2 uptake can become decoupled from "greenness" observed by satellites), ii) further modifying the respiration equation to incorporate accumulated EVI and/ or simultaneous GPP (to represent biomass, as in Xiao et al., 2011 and to account for the large contribution of recently assimilated carbon to autotrophic respiration), disturbance maps (to potentially improve spatial patterns), and week-to-week changes in EVI to account for inputs to surface litter pools at the end of the growing season, and iii) using pre-partitioned GPP and respiration data in the parameter optimization with more sophisticated algorithms for separating component fluxes, thus improving simulation of the flux diurnal cycle. In addition, the siting of new flux towers, especially in the southern half of the domain, in urban areas and across disturbance gradients, would help to improve the representativeness of optimized model parameters and flux estimates (as well as for CASA which also uses flux tower data for parameter calibration).

Along with parallel and continuing development for each biospheric model, the results of this study point towards what we might need in an "optimal" biospheric model for use in high-resolution CO_2 inversions in eastern North America, whether the biospheric signal is pre-subtracted from atmospheric observations or the inversion setup allows for the estimation of fossil fuel and biospheric fluxes simultaneously. Such a model should ideally include some or all of the following items: multiple land covers within each pixel weighted by fractional coverage or else very high spatial resolution (e.g., less than 100 m x 100 m), separation of different crop types to account for the strong uptake of corn relative to other crops, sub-monthly diagnostic phenology using EVI or SIF to better account for uptake during early and late growing-season transitions and water-stressed periods, more accurate land cover maps, improved mechanistic representation of R_e fluxes, perhaps aided by remote-sensing inputs, and the inclusion of processes allowing for net annual vegetative and soil sinks in models that track carbon pools. Future work using the tower CO_2 data in an atmospheric inversion model will also help to further identify needed improvements in the biospheric models that will help to enable an operational atmospheric emission monitoring system in North America.

Acknowledgements

The authors thanks Hratch Semerjian and David Allen (NIST) for their review and advice on the manuscript. This work was partially funded by NIST's Greenhouse Gas Measurements Program. Chris Williams and Yu Zhou were funded by NASA award #NNX16AN17G from the Atmospheric Carbon and Transport (ACT) - America project of the NASA Earth Venture Suborbital 2 program. Ian Baker was supported by NASA ACT-America subcontract 80NSSC20K0924. Funding for AmeriFlux data resources was provided by the U.S. Department of Energy's Office of Science.

SMG conceived of the study, ran VPRM, analyzed results and wrote the paper. ILC generated the footprints and provided initial model code for the gridded VPRM runs. AK generated the background conditions. YZ and CW provided the CASA model runs, while KH and IB provided the SiB4 runs. All co-authors gave ideas for analysis and helped edit the paper.

Data availability

Scripts to generate the results shown here, along with input data and summary files, are archived on the NIST server data.nist.gov at *doi:10.18434/mds2-2362*. DOIs for other datasets used in the paper are listed below.

Atmospheric CO_2 data:

Karion, Anna, Prinzivalli, Steve, Fain, Clayton, Stock, Michael, DiGangi, Elizabeth, Biggs, Bryan, Draper, Charlie, Baldelli, Seth, Veseshta, Uran, Salameh, Peter, Callahan, William, Whetstone, James (2019), Observations of CO_2, CH_4, and CO mole fractions from the NIST Northeast Corridor urban testbed, National Institute of Standards and Technology, https://doi.org/10.18434/M32126 (Accessed 2020-12-23)

Flux tower data:

Arain, M. Altaf (2003-) AmeriFlux CA-TP1 Ontario - Turkey Point 2002 Plantation White Pine, Dataset. https://doi.org/10.17190/AMF/1246009

Arain, M. Altaf (2003-) AmeriFlux CA-TP3 Ontario - Turkey Point 1974 Plantation White Pine, Dataset. https://doi.org/10.17190/AMF/1246011

Arain, M. Altaf (2012-) AmeriFlux CA-TPD Ontario - Turkey Point Mature Deciduous, Dataset. https://doi.org/10.17190/AMF/1246152

Baker, John, Tim Griffis (2003-2010) AmeriFlux US-Ro3 Rosemount- G19, Dataset. https://doi.org/10.17190/AMF/1246093

Baker, John, Tim Griffis (2003-2017) AmeriFlux US-Ro2 Rosemount- C7, Dataset. https://doi.org/10.17190/AMF/1418683

Baker, John, Tim Griffis (2014-) AmeriFlux US-Ro4 Rosemount Prairie, Dataset. https://doi.org/10.17190/AMF/1419507

Baker, John, Tim Griffis, Timothy Griffis (2003-2017) AmeriFlux US-Ro1 Rosemount- G21, Dataset. https://doi.org/10.17190/AMF/1246092

Bernacchi, Carl (2004-2008) AmeriFlux US-Bo2 Bondville (companion site), Dataset. https://doi.org/10.17190/AMF/1246037
Biraud, Sebastien, Marc Fischer, Stephen Chan, Margaret Torn (2002-) AmeriFlux US-ARM ARM Southern Great Plains site- Lamont, Dataset. https://doi.org/10.17190/AMF/1246027
Bohrer, Gil (2011-2016) AmeriFlux US-ORv Olentangy River Wetland Research Park, Dataset. https://doi.org/10.17190/AMF/1246135
Bohrer, Gil, Janice Kerns (2015-2016) AmeriFlux US-OWC Old Woman Creek, Dataset. https://doi.org/10.17190/AMF/1418679
Chen, Jiquan (2002-2002) AmeriFlux US-Wi8 Young hardwood clearcut (YHW), Dataset. https://doi.org/10.17190/AMF/1246023
Chen, Jiquan (2002-2005) AmeriFlux US-Wi4 Mature red pine (MRP), Dataset. https://doi.org/10.17190/AMF/1246019
Chen, Jiquan (2003-2003) AmeriFlux US-Wi1 Intermediate hardwood (IHW), Dataset. https://doi.org/10.17190/AMF/1246015
Chen, Jiquan (2004-2004) AmeriFlux US-Wi5 Mixed young jack pine (MYJP), Dataset. https://doi.org/10.17190/AMF/1246020
Chen, Jiquan (2004-2005) AmeriFlux US-Wi9 Young Jack pine (YJP), Dataset. https://doi.org/10.17190/AMF/1246024
Chen, Jiquan (2005-2005) AmeriFlux US-Wi7 Red pine clearcut (RPCC), Dataset. https://doi.org/10.17190/AMF/1246022
Chen, Jiquan, Housen Chu (2011-2013) AmeriFlux US-CRT Curtice Walter-Berger cropland, Dataset. https://doi.org/10.17190/AMF/1246156
Chen, Jiquan, Housen Chu (2011-2013) AmeriFlux US-WPT Winous Point North Marsh, Dataset. https://doi.org/10.17190/AMF/1246155
Chen, Jiquan, Housen Chu, Asko Noormets (2004-2013) AmeriFlux US-Oho Oak Openings, Dataset. https://doi.org/10.17190/AMF/1246089
Clark, Ken (2004-) AmeriFlux US-Slt Silas Little- New Jersey, Dataset. https://doi.org/10.17190/AMF/1246096
Clark, Ken (2005-2008) AmeriFlux US-Dix Fort Dix, Dataset. https://doi.org/10.17190/AMF/1246045
Clark,Ken (2005-) AmeriFlux US-Ced Cedar Bridge, Dataset. https://doi.org/10.17190/AMF/1246043
Desai, Ankur (1996-) AmeriFlux US-PFa Park Falls/WLEF, Dataset. https://doi.org/10.17190/AMF/1246090
Desai, Ankur (1999-) AmeriFlux US-WCr Willow Creek, Dataset. https://doi.org/10.17190/AMF/1246111
Desai, Ankur (2001-) AmeriFlux US-Los Lost Creek, Dataset. https://doi.org/10.17190/AMF/1246071
Desai, Ankur (2001-) AmeriFlux US-Syv Sylvania Wilderness Area, Dataset. https://doi.org/10.17190/AMF/1246106
Drake, Bert, Ross Hinkle (2000-2007) AmeriFlux US-KS2 Kennedy Space Center (scrub oak), Dataset. https://doi.org/10.17190/AMF/1246070
Gough, Christopher, Gil Bohrer, Peter Curtis (1999-) AmeriFlux US-UMB Univ. of Mich. Biological Station, Dataset. https://doi.org/10.17190/AMF/1246107
Gough, Christopher, Gil Bohrer, Peter Curtis (2007-) AmeriFlux US-UMd UMBS Disturbance, Dataset. https://doi.org/10.17190/AMF/1246134
Hadley, Julian, J. William Munger (2004-) AmeriFlux US-Ha2 Harvard Forest Hemlock Site, Dataset. https://doi.org/10.17190/AMF/1246060
Heilman, Jim (2004-) AmeriFlux US-FR3 Freeman Ranch- Woodland, Dataset. https://doi.org/10.17190/AMF/1246055
Hollinger, David (1996-) AmeriFlux US-Ho1 Howland Forest (main tower), Dataset. https://doi.org/10.17190/AMF/1246061
Hollinger, David (1999-) AmeriFlux US-Ho2 Howland Forest (west tower), Dataset. https://doi.org/10.17190/AMF/1246062
Hollinger, David (2000-) AmeriFlux US-Ho3 Howland Forest (harvest site), Dataset. https://doi.org/10.17190/AMF/1246063
Lee, Xuhui (1999-2004) AmeriFlux US-GMF Great Mountain Forest, Dataset. https://doi.org/10.17190/AMF/1246057
Matamala, Roser (2004-) AmeriFlux US-IB2 Fermi National Accelerator Laboratory- Batavia (Prairie site), Dataset. https://doi.org/10.17190/AMF/1246066

1049 Matamala, Roser (2005-) AmeriFlux US-IB1 Fermi National Accelerator Laboratory- Batavia (Agricultural site),
1050 Dataset. https://doi.org/10.17190/AMF/1246065
1051 McCaughey, Harry (2003-) AmeriFlux CA-Gro Ontario - Groundhog River, Boreal Mixedwood Forest, Dataset.
1052 https://doi.org/10.17190/AMF/1245996
1053 McFadden, Joe (2005-2009) AmeriFlux US-KUT KUOM Turfgrass Field, Dataset.
1054 https://doi.org/10.17190/AMF/1246145
1055 Meyers, Tilden (1995-1999) AmeriFlux US-WBW Walker Branch Watershed, Dataset.
1056 https://doi.org/10.17190/AMF/1246109
1057 Meyers, Tilden (1996-) AmeriFlux US-Bo1 Bondville, Dataset. https://doi.org/10.17190/AMF/1246036
1058 Meyers, Tilden (2002-2006) AmeriFlux US-Goo Goodwin Creek, Dataset. https://doi.org/10.17190/AMF/1246058
1059 Meyers, Tilden (2004-) AmeriFlux US-CaV Canaan Valley, Dataset. https://doi.org/10.17190/AMF/1246042
1060 Meyers, Tilden (2005-) AmeriFlux US-ChR Chestnut Ridge, Dataset. https://doi.org/10.17190/AMF/1246044
1061 Munger, J. William (1991-) AmeriFlux US-Ha1 Harvard Forest EMS Tower (HFR1), Dataset.
1062 https://doi.org/10.17190/AMF/1246059
1063 Noormets, Asko (2005-) AmeriFlux US-NC2 NC_Loblolly Plantation, Dataset.
1064 https://doi.org/10.17190/AMF/1246083
1065 Noormets, Asko (2005-2013) AmeriFlux US-NC1 NC_Clearcut, Dataset. https://doi.org/10.17190/AMF/1246082
1066 Noormets, Asko (2013-) AmeriFlux US-NC3 NC_Clearcut#3, Dataset. https://doi.org/10.17190/AMF/1419506
1067 Novick, Kim, Rich Phillips (1999-) AmeriFlux US-MMS Morgan Monroe State Forest, Dataset.
1068 https://doi.org/10.17190/AMF/1246080
1069 Oishi, Chris, Kim Novick, Paul Stoy (2001-2008) AmeriFlux US-Dk1 Duke Forest-open field, Dataset.
1070 https://doi.org/10.17190/AMF/1246046
1071 Oishi, Chris, Kim Novick, Paul Stoy (2001-2008) AmeriFlux US-Dk2 Duke Forest-hardwoods, Dataset.
1072 https://doi.org/10.17190/AMF/1246047
1073 Oishi, Chris, Kim Novick, Paul Stoy (2001-2008) AmeriFlux US-Dk3 Duke Forest - loblolly pine, Dataset.
1074 https://doi.org/10.17190/AMF/1246048
1075 Prueger, John, Tim Parkin (2001-) AmeriFlux US-Br1 Brooks Field Site 10- Ames, Dataset.
1076 https://doi.org/10.17190/AMF/1246038
1077 Prueger, John, Tim Parkin (2001-) AmeriFlux US-Br3 Brooks Field Site 11- Ames, Dataset.
1078 https://doi.org/10.17190/AMF/1246039
1079 Richardson, Andrew, David Hollinger (2004-) AmeriFlux US-Bar Bartlett Experimental Forest, Dataset.
1080 https://doi.org/10.17190/AMF/1246030
1081 Sturtevant, Cove, David Durden, Stefan Metzger (2016-) AmeriFlux US-xSC NEON Smithsonian Conservation Biology
1082 Institute (SCBI), Dataset. https://doi.org/10.17190/AMF/1671900
1083 Sturtevant, Cove, David Durden, Stefan Metzger (2016-) AmeriFlux US-xSE NEON Smithsonian Environmental
1084 Research Center (SERC), Dataset. https://doi.org/10.17190/AMF/1617734
1085 Sturtevant, Cove, David Durden, Stefan Metzger (2017-) AmeriFlux US-xDL NEON Dead Lake (DELA), Dataset.
1086 https://doi.org/10.17190/AMF/1579721
1087 Sturtevant, Cove, David Durden, Stefan Metzger (2017-) AmeriFlux US-xGR NEON Great Smoky Mountains National
1088 Park, Twin Creeks (GRSM), Dataset. https://doi.org/10.17190/AMF/1634885
1089 Sturtevant, Cove, David Durden, Stefan Metzger (2017-) AmeriFlux US-xST NEON Steigerwaldt Land Services (STEI),
1090 Dataset. https://doi.org/10.17190/AMF/1617737
1091 Sturtevant, Cove, David Durden, Stefan Metzger (2017-) AmeriFlux US-xTA NEON Talladega National Forest (TALL),
1092 Dataset. https://doi.org/10.17190/AMF/1671902
1093 Sturtevant, Cove, David Durden, Stefan Metzger (2017-) AmeriFlux US-xTR NEON Treehaven (TREE), Dataset.
1094 https://doi.org/10.17190/AMF/1634886
1095 Sturtevant, Cove, David Durden, Stefan Metzger (2017-) AmeriFlux US-xUK NEON The University of Kansas Field
1096 Station (UKFS), Dataset. https://doi.org/10.17190/AMF/1617740
1097 Sturtevant, Cove, David Durden, Stefan Metzger (2017-) AmeriFlux US-xUN NEON University of Notre Dame
1098 Environmental Research Center (UNDE), Dataset. https://doi.org/10.17190/AMF/1617741
1099 Suyker, Andy (2001-) AmeriFlux US-Ne2 Mead - irrigated maize-soybean rotation site, Dataset.
1100 https://doi.org/10.17190/AMF/1246085

Suyker, Andy (2001-) AmeriFlux US-Ne3 Mead - rainfed maize-soybean rotation site, Dataset. https://doi.org/10.17190/AMF/1246086
Torn, Margaret (2005-2006) AmeriFlux US-ARc ARM Southern Great Plains control site- Lamont, Dataset. https://doi.org/10.17190/AMF/1246026
Wood, Jeffrey, Lianhong Gu (2004-) AmeriFlux US-MOz Missouri Ozark Site, Dataset. https://doi.org/10.17190/AMF/1246081

References

A. E. Schuh, Otte, M. J., Walko, R., Oda, T., & Ott, L. (2019). Using the Refined-Mesh Ocean-Land-Atmosphere-Model (OLAM) to Quantify Expected Carbon Dioxide Variations Across Urban Landscapes. Presented at the American Meteorological Society, Phoenix, AZ.

Amthor, J. S. (2000). The McCree–de Wit–Penning de Vries–Thornley Respiration Paradigms: 30 Years Later. *Annals of Botany*, *86*(1), 1–20. https://doi.org/10.1006/anbo.2000.1175

Asefi-Najafabady, S., Rayner, P. J., Gurney, K. R., McRobert, A., Song, Y., Coltin, K., et al. (2014). A multiyear, global gridded fossil fuel CO2 emission data product: Evaluation and analysis of results. *Journal of Geophysical Research: Atmospheres*, *119*(17), 10,213-10,231. https://doi.org/10.1002/2013JD021296

Aubinet, M. (2008). Eddy Covariance Co2 Flux Measurements in Nocturnal Conditions: An Analysis of the Problem. *Ecological Applications*, *18*(6), 1368–1378. https://doi.org/10.1890/06-1336.1

Barr, A. G., Richardson, A. D., Hollinger, D. Y., Papale, D., Arain, M. A., Black, T. A., et al. (2013). Use of change-point detection for friction–velocity threshold evaluation in eddy-covariance studies. *Agricultural and Forest Meteorology*, *171–172*, 31–45. https://doi.org/10.1016/j.agrformet.2012.11.023

Basu, S., Lehman, S. J., Miller, J. B., Andrews, A. E., Sweeney, C., Gurney, K. R., et al. (2020). Estimating US fossil fuel CO2 emissions from measurements of 14C in atmospheric CO2. *Proceedings of the National Academy of Sciences*, *117*(24), 13300–13307. https://doi.org/10.1073/pnas.1919032117

Benjamin, S. G., Weygandt, S. S., Brown, J. M., Hu, M., Alexander, C. R., Smirnova, T. G., et al. (2016). A North American Hourly Assimilation and Model Forecast Cycle: The Rapid Refresh. *Monthly Weather Review*, *144*(4), 1669–1694. https://doi.org/10.1175/MWR-D-15-0242.1

Berry, J., Wolf, A., Campbell, J. E., Baker, I., Blake, N., Blake, D., et al. (2013). A coupled model of the global cycles of carbonyl sulfide and CO2: A possible new window on the carbon cycle. *Journal of Geophysical Research: Biogeosciences*, *118*(2), 842–852. https://doi.org/10.1002/jgrg.20068

Buyantuyev, A., & Wu, J. (2009). Urbanization alters spatiotemporal patterns of ecosystem primary production: A case study of the Phoenix metropolitan region, USA. *Journal of Arid Environments*, *73*(4), 512–520. https://doi.org/10.1016/j.jaridenv.2008.12.015

Chandrasekar, K., Sai, M. V. R. S., Roy, P. S., & Dwevedi, R. S. (2010). Land Surface Water Index (LSWI) response to rainfall and NDVI using the MODIS Vegetation Index product. *International Journal of Remote Sensing*, *31*(15), 3987–4005. https://doi.org/10.1080/01431160802575653

Corbin, K. D., Denning, A. S., Lokupitiya, E. Y., Schuh, A. E., Miles, N. L., Davis, K. J., et al. (2010). Assessing the impact of crops on regional CO2 fluxes and atmospheric concentrations. *Tellus B: Chemical and Physical Meteorology*, *62*(5), 521–532. https://doi.org/10.1111/j.1600-0889.2010.00485.x

Elder, J. (2018). Chapter 16 - The Apparent Paradox of Complexity in Ensemble Modeling*. In R. Nisbet, G. Miner, & K. Yale (Eds.), *Handbook of Statistical Analysis and Data Mining Applications (Second Edition)* (pp. 705–718). Boston: Academic Press. https://doi.org/10.1016/B978-0-12-416632-5.00016-5

Falge, E., Baldocchi, D., Tenhunen, J., Aubinet, M., Bakwin, P., Berbigier, P., et al. (2002). Seasonality of ecosystem respiration and gross primary production as derived from FLUXNET measurements. *Agricultural and Forest Meteorology*, *113*(1), 53–74. https://doi.org/10.1016/S0168-1923(02)00102-8

Farquhar, G. D., von Caemmerer, S., & Berry, J. A. (1980). A biochemical model of photosynthetic CO2 assimilation in leaves of C3 species. *Planta*, *149*(1), 78–90. https://doi.org/10.1007/BF00386231

Feng, S., Lauvaux, T., Keller, K., Davis, K. J., Rayner, P., Oda, T., & Gurney, K. R. (2019). A Road Map for Improving the Treatment of Uncertainties in High-Resolution Regional Carbon Flux Inverse

Estimates. *Geophysical Research Letters*, *46*(22), 13461–13469. https://doi.org/10.1029/2019GL082987

Feng, S., Lauvaux, T., Davis, K. J., Keller, K., Zhou, Y., Williams, C., et al. (2019). Seasonal Characteristics of Model Uncertainties From Biogenic Fluxes, Transport, and Large-Scale Boundary Inflow in Atmospheric CO2 Simulations Over North America. *Journal of Geophysical Research: Atmospheres*, *124*(24), 14325–14346. https://doi.org/10.1029/2019JD031165

Fisher, J. B., Sikka, M., Huntzinger, D. N., Schwalm, C., & Liu, J. (2016). Technical note: 3-hourly temporal downscaling of monthly global terrestrial biosphere model net ecosystem exchange. *Biogeosciences*, *13*(14), 4271–4277. https://doi.org/10.5194/bg-13-4271-2016

Flexas, J., Bota, J., Galmés, J., Medrano, H., & Ribas-Carbó, M. (2006). Keeping a positive carbon balance under adverse conditions: responses of photosynthesis and respiration to water stress. *Physiologia Plantarum*, *127*(3), 343–352. https://doi.org/10.1111/j.1399-3054.2006.00621.x

Golubiewski, N. E. (2006). Urbanization Increases Grassland Carbon Pools: Effects Of Landscaping In Colorado's Front Range. *Ecological Applications*, *16*(2), 555–571. https://doi.org/10.1890/1051-0761(2006)016[0555:UIGCPE]2.0.CO;2

Gourdji, S. M., Hirsch, A. I., Mueller, K. L., Yadav, V., Andrews, A. E., & Michalak, A. M. (2010). Regional-scale geostatistical inverse modeling of North American CO_2 fluxes: a synthetic data study. *Atmospheric Chemistry and Physics*, *10*(13), 6151–6167. https://doi.org/10.5194/acp-10-6151-2010

Gourdji, S. M., Mueller, K. L., Yadav, V., Huntzinger, D. N., Andrews, A. E., Trudeau, M., et al. (2012). North American CO_2 exchange: inter-comparison of modeled estimates with results from a fine-scale atmospheric inversion. *Biogeosciences*, *9*(1), 457–475. https://doi.org/10.5194/bg-9-457-2012

Gray, J. M., Frolking, S., Kort, E. A., Ray, D. K., Kucharik, C. J., Ramankutty, N., & Friedl, M. A. (2014). Direct human influence on atmospheric CO 2 seasonality from increased cropland productivity. *Nature*, *515*(7527), 398–401. https://doi.org/10.1038/nature13957

Guindin-Garcia, N., Gitelson, A. A., Arkebauer, T. J., Shanahan, J., & Weiss, A. (2012). An evaluation of MODIS 8- and 16-day composite products for monitoring maize green leaf area index. *Agricultural and Forest Meteorology*, *161*, 15–25. https://doi.org/10.1016/j.agrformet.2012.03.012

Gurney, Kevin R., Liang, J., Patarasuk, R., Song, Y., Huang, J., & Roest, G. (2020). The Vulcan Version 3.0 High-Resolution Fossil Fuel CO2 Emissions for the United States. *Journal of Geophysical Research: Atmospheres*, *125*(19), e2020JD032974. https://doi.org/10.1029/2020JD032974

Gurney, Kevin Robert, Law, R. M., Denning, A. S., Rayner, P. J., Baker, D., Bousquet, P., et al. (2002). Towards robust regional estimates of CO 2 sources and sinks using atmospheric transport models. *Nature*, *415*(6872), 626–630. https://doi.org/10.1038/415626a

Hardiman, B. S., Wang, J. A., Hutyra, L. R., Gately, C. K., Getson, J. M., & Friedl, M. A. (2017). Accounting for urban biogenic fluxes in regional carbon budgets. *The Science of the Total Environment*, *592*, 366–372. https://doi.org/10.1016/j.scitotenv.2017.03.028

Haynes, K. D., Baker, I. T., Denning, A. S., Stöckli, R., Schaefer, K., Lokupitiya, E. Y., & Haynes, J. M. (2019). Representing Grasslands Using Dynamic Prognostic Phenology Based on Biological Growth Stages: 1. Implementation in the Simple Biosphere Model (SiB4). *Journal of Advances in Modeling Earth Systems*, *11*(12), 4423–4439. https://doi.org/10.1029/2018MS001540

Haynes, Katherine D., Baker, I. T., Denning, A. S., Wolf, S., Wohlfahrt, G., Kiely, G., et al. (2019). Representing Grasslands Using Dynamic Prognostic Phenology Based on Biological Growth Stages: Part 2. Carbon Cycling. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4440–4465. https://doi.org/10.1029/2018MS001541

Hilton, T. W., Davis, K. J., Keller, K., & Urban, N. M. (2013). Improving North American terrestrial CO_2 flux diagnosis using spatial structure in land surface model residuals. *Biogeosciences*, *10*(7), 4607–4625. https://doi.org/10.5194/bg-10-4607-2013

Hilton, T. W., Davis, K. J., & Keller, K. (2014). Evaluating terrestrial CO_2 flux diagnoses and uncertainties from a simple land surface model and its residuals. *Biogeosciences*, *11*(2), 217–235. https://doi.org/10.5194/bg-11-217-2014

Hilton, Timothy W., Whelan, M. E., Zumkehr, A., Kulkarni, S., Berry, J. A., Baker, I. T., et al. (2017). Peak growing season gross uptake of carbon in North America is largest in the Midwest USA. *Nature Climate Change*, *7*(6), 450–454. https://doi.org/10.1038/nclimate3272

Högberg, P., Nordgren, A., Buchmann, N., Taylor, A. F., Ekblad, A., Högberg, M. N., et al. (2001). Large-scale forest girdling shows that current photosynthesis drives soil respiration. *Nature*, *411*(6839), 789–792. https://doi.org/10.1038/35081058

Hu, L., Andrews, A. E., Thoning, K. W., Sweeney, C., Miller, J. B., Michalak, A. M., et al. (2019). Enhanced North American carbon uptake associated with El Niño. *Science Advances*, *5*(6), eaaw0076. https://doi.org/10.1126/sciadv.aaw0076

Hu, X.-M., Gourdji, S., Davis, K. J., Wang, Q., Zhang, Y., Xue, M., et al. (in review). Implementation of improved parameterization of terrestrial flux in WRF-VPRM improves the simulation of nighttime CO2 peaks and a daytime CO2 band ahead of a cold front. *Journal of Geophysical Research: Atmospheres*.

Jacobson, A. R., Schuldt, K. N., Miller, J. B., Oda, T., Tans, P., Andrews, A., et al. (2020). CarbonTracker CT2019B. https://doi.org/10.25925/20201008

Jassal, R. S., Black, T. A., Cai, T., Morgenstern, K., Li, Z., Gaumont-Guay, D., & Nesic, Z. (2007). Components of ecosystem respiration and an estimate of net primary productivity of an intermediate-aged Douglas-fir stand. *Agricultural and Forest Meteorology*, *144*(1), 44–57. https://doi.org/10.1016/j.agrformet.2007.01.011

Jin, S., Homer, C., Yang, L., Danielson, P., Dewitz, J., Li, C., et al. (2019). Overall Methodology Design for the United States National Land Cover Database 2016 Products. *Remote Sensing*, *11*(24), 2971. https://doi.org/10.3390/rs11242971

Joiner, J., Yoshida, Y., Vasilkov, A. P., Yoshida, Y., Corp, L. A., & Middleton, E. M. (2011). First observations of global and seasonal terrestrial chlorophyll fluorescence from space. *Biogeosciences*, *8*(3), 637–651. https://doi.org/10.5194/bg-8-637-2011

Karion, A., Callahan, W., Stock, M., Prinzivalli, S., Verhulst, K. R., Kim, J., et al. (2020). Greenhouse gas observations from the Northeast Corridor tower network. *Earth System Science Data*, *12*(1), 699–717. https://doi.org/10.5194/essd-12-699-2020

Karion, A., Lopez-Coto, I., Gourdji, S. M., Mueller, K., Ghosh, S., Callahan, W., et al. (2021). Background conditions for an urban greenhouse gas network in the Washington, D.C. and Baltimore metropolitan region. *Atmospheric Chemistry and Physics Discussions*, 1–27. https://doi.org/10.5194/acp-2020-1256

Lasslop, G., Reichstein, M., Papale, D., Richardson, A. D., Arneth, A., Barr, A., et al. (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*(1), 187–208. https://doi.org/10.1111/j.1365-2486.2009.02041.x

Lauvaux, T., Uliasz, M., Sarrat, C., Chevallier, F., Bousquet, P., Lac, C., et al. (2008). Mesoscale inversion: first results from the CERES campaign with synthetic data. *Atmospheric Chemistry and Physics*, *8*(13), 3459–3471. https://doi.org/10.5194/acp-8-3459-2008

Lauvaux, T., Schuh, A. E., Uliasz, M., Richardson, S., Miles, N., Andrews, A. E., et al. (2012). Constraining the CO_2 budget of the corn belt: exploring uncertainties from the assumptions in a mesoscale

inverse system. *Atmospheric Chemistry and Physics*, *12*(1), 337–354. https://doi.org/10.5194/acp-12-337-2012

Lauvaux, Thomas, Gurney, K. R., Miles, N. L., Davis, K. J., Richardson, S. J., Deng, A., et al. (2021). Policy-Relevant Assessment of Urban CO2 Emissions | Environmental Science & Technology. Retrieved from https://pubs.acs.org/doi/abs/10.1021/acs.est.0c00343

Li, X., Hu, X.-M., Cai, C., Jia, Q., Zhang, Y., Liu, J., et al. (2020). Terrestrial CO2 Fluxes, Concentrations, Sources and Budget in Northeast China: Observational and Modeling Studies. *Journal of Geophysical Research: Atmospheres*, *125*(6), e2019JD031686. https://doi.org/10.1029/2019JD031686

Lin, J. C., Gerbig, C., Wofsy, S. C., Andrews, A. E., Daube, B. C., Davis, K. J., & Grainger, C. A. (2003). A near-field tool for simulating the upstream influence of atmospheric observations: The Stochastic Time-Inverted Lagrangian Transport (STILT) model. *Journal of Geophysical Research: Atmospheres*, *108*(D16). https://doi.org/10.1029/2002JD003161

Lokupitiya, E., Denning, S., Paustian, K., Baker, I., Schaefer, K., Verma, S., et al. (2009). Incorporation of crop phenology in Simple Biosphere Model (SiBcrop) to improve land-atmosphere carbon exchanges from croplands. *Biogeosciences*, *6*(6), 969–986. https://doi.org/10.5194/bg-6-969-2009

Lopez-Coto, I., Ghosh, S., Prasad, K., & Whetstone, J. (2017). Tower-based greenhouse gas measurement network design—The National Institute of Standards and Technology North East Corridor Testbed. *Advances in Atmospheric Sciences*, *34*(9), 1095–1105. https://doi.org/10.1007/s00376-017-6094-6

Luus, K. A., & Lin, J. C. (2015). The Polar Vegetation Photosynthesis and Respiration Model: a parsimonious, satellite-data-driven model of high-latitude CO_2 exchange. *Geoscientific Model Development*, *8*(8), 2655–2674. https://doi.org/10.5194/gmd-8-2655-2015

Mahadevan, P., Wofsy, S. C., Matross, D. M., Xiao, X., Dunn, A. L., Lin, J. C., et al. (2008). A satellite-based biosphere parameterization for net ecosystem CO2 exchange: Vegetation Photosynthesis and Respiration Model (VPRM). *Global Biogeochemical Cycles*, *22*(2). https://doi.org/10.1029/2006GB002735

Martin, C. R., Zeng, N., Karion, A., Mueller, K., Ghosh, S., Lopez-Coto, I., et al. (2019). Investigating sources of variability and error in simulations of carbon dioxide in an urban region. *Atmospheric Environment*, *199*, 55–69. https://doi.org/10.1016/j.atmosenv.2018.11.013

Meir, P., Metcalfe, D. b, Costa, A. c. l, & Fisher, R. a. (2008). The fate of assimilated carbon during drought: impacts on respiration in Amazon rainforests. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1498), 1849–1855. https://doi.org/10.1098/rstb.2007.0021

Miles, N. L., Richardson, S. J., Martins, D. K., Davis, K. J., Lauvaux, T., Haupt, B. J., & Miller, S. K. (2018). ACT-America: L2 In Situ CO2, CO, and CH4 Concentrations from Towers, Eastern USA. *ORNL DAAC*. https://doi.org/10.3334/ORNLDAAC/1568

Miller, J. B., Lehman, S. J., Verhulst, K. R., Miller, C. E., Duren, R. M., Yadav, V., et al. (2020). Large and seasonally varying biospheric CO2 fluxes in the Los Angeles megacity revealed by atmospheric radiocarbon. *Proceedings of the National Academy of Sciences*, *117*(43), 26681–26687. https://doi.org/10.1073/pnas.2005253117

Mitchell, L., Lin, J. C., Hutyra, L. R., Sargent, M., Wofsy, S. C., Miles, N. L., et al. (2019). NACP: Urban Greenhouse Gases across the CO2 Urban Synthesis and Analysis Network. *ORNL DAAC*. https://doi.org/10.3334/ORNLDAAC/1743

Molchanov, A. G. (2009). Effect of moisture availability on photosynthetic productivity and autotrophic respiration of an oak stand. *Russian Journal of Plant Physiology*, *56*(6), 769. https://doi.org/10.1134/S1021443709060065

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE*, *50*(3), 885–900.

Mueller, K., Yadav, V., Lopez-Coto, I., Karion, A., Gourdji, S., Martin, C., & Whetstone, J. (2018). Siting Background Towers to Characterize Incoming Air for Urban Greenhouse Gas Estimation: A Case Study in the Washington, DC/Baltimore Area. *Journal of Geophysical Research: Atmospheres*, *123*(5), 2910–2926. https://doi.org/10.1002/2017JD027364

NCEI, NWS, NOAA, & US DOC. (2020). North American Mesoscale Forecast System (NAM) [12 km]. Retrieved February 7, 2021, from https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00630

Nicolini, G., Aubinet, M., Feigenwinter, C., Heinesch, B., Lindroth, A., Mamadou, O., et al. (2018). Impact of CO2 storage flux sampling uncertainty on net ecosystem exchange measured by eddy covariance. *Agricultural and Forest Meteorology*, *248*, 228–239. https://doi.org/10.1016/j.agrformet.2017.09.025

NOAA Earth System Research Laboratory, Global Monitoring Division. (2019). Cooperative Global Atmospheric Data Integration Project (2019): Multi-laboratory compilation of atmospheric carbon dioxide data for the period 1957-2018; obspack_co2_1_GLOBALVIEWplus_v5.0_2019_08_12. Retrieved from http://dx.doi.org/10.25925/20190812

Nowak, D. J., & Crane, D. E. (2002). Carbon storage and sequestration by urban trees in the USA. *Environmental Pollution*, *116*(3), 381–389. https://doi.org/10.1016/S0269-7491(01)00214-7

Parazoo, N. C., Frankenberg, C., Köhler, P., Joiner, J., Yoshida, Y., Magney, T., et al. (2019). Towards a Harmonized Long-Term Spaceborne Record of Far-Red Solar-Induced Fluorescence. *Journal of Geophysical Research: Biogeosciences*, *124*(8), 2518–2539. https://doi.org/10.1029/2019JG005289

Parton, W. J., Stewart, J. W. B., & Cole, C. V. (1988). Dynamics of C, N, P and S in Grassland Soils: A Model. *Biogeochemistry*, *5*(1), 109–131.

Peters, W., Krol, M. C., Werf, G. R. V. D., Houweling, S., Jones, C. D., Hughes, J., et al. (2010). Seven years of recent European net terrestrial carbon dioxide exchange constrained by atmospheric observations. *Global Change Biology*, *16*(4), 1317–1337. https://doi.org/10.1111/j.1365-2486.2009.02078.x

Pillai, D., Gerbig, C., Ahmadov, R., Rödenbeck, C., Kretschmer, R., Koch, T., et al. (2011). High-resolution simulations of atmospheric CO_2 over complex terrain – representing the Ochsenkopf mountain tall tower. *Atmospheric Chemistry and Physics*, *11*(15), 7445–7464. https://doi.org/10.5194/acp-11-7445-2011

Pitt, J. R., Lopez-Coto, I., Hajny, K. D. D., Mendoza Tomlin, J., Kaeser, R., Jayarathne, T., et al. (2020). New York City greenhouse gas emissions estimated with inverse modeling of aircraft measurements. Presented at the AGU Fall Meeting 2020, AGU.

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*(4), 811–841. https://doi.org/10.1029/93GB02725

Raciti, S. M., Hutyra, L. R., & Newell, J. D. (2014). Mapping carbon storage in urban trees with multi-source remote sensing data: relationships between biomass, land use, and demographics in Boston neighborhoods. *The Science of the Total Environment*, *500–501*, 72–83. https://doi.org/10.1016/j.scitotenv.2014.08.070

Raczka, B. M., Davis, K. J., Huntzinger, D., Neilson, R. P., Poulter, B., Richardson, A. D., et al. (2013). Evaluation of continental carbon cycle simulations with North American flux tower observations. *Ecological Monographs*, *83*(4), 531–556. https://doi.org/10.1890/12-0893.1

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*(4), 585–602. https://doi.org/10.1029/96GB01981

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., et al. (2005). On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global Change Biology*, *11*(9), 1424–1439. https://doi.org/10.1111/j.1365-2486.2005.001002.x

Richardson, A. D., Anderson, R. S., Arain, M. A., Barr, A. G., Bohrer, G., Chen, G., et al. (2012). Terrestrial biosphere models need better representation of vegetation phenology: results from the North American Carbon Program Site Synthesis. *Global Change Biology*, *18*(2), 566–584. https://doi.org/10.1111/j.1365-2486.2011.02562.x

Richardson, S. J., Miles, N. L., Davis, K. J., Lauvaux, T., Martins, D. K., Turnbull, J. C., et al. (2017). Tower measurement network of in-situ CO2, CH4, and CO in support of the Indianapolis FLUX (INFLUX) Experiment. *Elementa: Science of the Anthropocene*, *5*(59). https://doi.org/10.1525/elementa.140

Rödenbeck, C., Houweling, S., Gloor, M., & Heimann, M. (2003). CO_2 flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport. *Atmospheric Chemistry and Physics*, *3*(6), 1919–1964. https://doi.org/10.5194/acp-3-1919-2003

Sargent, M., Barrera, Y., Nehrkorn, T., Hutyra, L. R., Gately, C. K., Jones, T., et al. (2018). Anthropogenic and biogenic CO2 fluxes in the Boston urban region. *Proceedings of the National Academy of Sciences*, *115*(29), 7491–7496. https://doi.org/10.1073/pnas.1803715115

Schwalm, C. R., Williams, C. A., Schaefer, K., Anderson, R., Arain, M. A., Baker, I., 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: Biogeosciences*, *115*(G3). https://doi.org/10.1029/2009JG001229

Sellers, P. J., Mintz, Y., Sud, Y. C., & Dalcher, A. (1986). A Simple Biosphere Model (SIB) for Use within General Circulation Models. *Journal of Atmospheric Sciences*, *43*(6), 505–531. https://doi.org/10.1175/1520-0469(1986)043<0505:ASBMFU>2.0.CO;2

Sellers, P. J., Randall, D. A., Collatz, G. J., Berry, J. A., Field, C. B., Dazlich, D. A., et al. (1996). A Revised Land Surface Parameterization (SiB2) for Atmospheric GCMS. Part I: Model Formulation. *Journal of Climate*, *9*(4), 676–705. https://doi.org/10.1175/1520-0442(1996)009<0676:ARLSPF>2.0.CO;2

Shiga, Y. P., Michalak, A. M., Gourdji, S. M., Mueller, K. L., & Yadav, V. (2014). Detecting fossil fuel emissions patterns from subcontinental regions using North American in situ CO2 measurements. *Geophysical Research Letters*, *41*(12), 4381–4388. https://doi.org/10.1002/2014GL059684

Shiga, Y. P., Tadić, J. M., Qiu, X., Yadav, V., Andrews, A. E., Berry, J. A., & Michalak, A. M. (2018). Atmospheric CO2 Observations Reveal Strong Correlation Between Regional Net Biospheric Carbon Uptake and Solar-Induced Chlorophyll Fluorescence. *Geophysical Research Letters*, *45*(2), 1122–1132. https://doi.org/10.1002/2017GL076630

Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., Wang, W., et al. (2008). *A Description of the Advanced Research WRF Version 3* [Application/pdf] (p. 1002 KB). UCAR/NCAR. https://doi.org/10.5065/D68S4MVH

Turner, A. J., & Jacob, D. J. (2015). Balancing aggregation and smoothing errors in inverse models. *Atmospheric Chemistry and Physics*, *15*(12), 7039–7048. https://doi.org/10.5194/acp-15-7039-2015

Turner, Alexander J., Köhler, P., Magney, T. S., Frankenberg, C., Fung, I., & Cohen, R. C. (2020). A double peak in the seasonality of California's photosynthesis as observed from space. *Biogeosciences*, *17*(2), 405–422. https://doi.org/10.5194/bg-17-405-2020

Wang, X., Dannenberg, M. P., Yan, D., Jones, M. O., Kimball, J. S., Moore, D. J. P., et al. (2020). Globally Consistent Patterns of Asynchrony in Vegetation Phenology Derived From Optical, Microwave, and Fluorescence Satellite Data. *Journal of Geophysical Research: Biogeosciences*, *125*(7), e2020JG005732. https://doi.org/10.1029/2020JG005732

Wang, Y., Deutscher, N. M., Palm, M., Warneke, T., Notholt, J., Baker, I., et al. (2016). Towards understanding the variability in biospheric CO_2 fluxes: using FTIR spectrometry and a chemical transport model to investigate the sources and sinks of carbonyl sulfide and its link to CO_2. *Atmospheric Chemistry and Physics*, *16*(4), 2123–2138. https://doi.org/10.5194/acp-16-2123-2016

Xiao, J., Davis, K. J., Urban, N. M., Keller, K., & Saliendra, N. Z. (2011). Upscaling carbon fluxes from towers to the regional scale: Influence of parameter variability and land cover representation on regional flux estimates. *Journal of Geophysical Research: Biogeosciences*, *116*(G3). https://doi.org/10.1029/2010JG001568

Xiao, X., Hollinger, D., Aber, J., Goltz, M., Davidson, E. A., Zhang, Q., & Iii, B. M. (2004). Satellite-based modeling of gross primary production in an evergreen needleleaf forest. Retrieved from https://www.nrs.fs.fed.us/pubs/6798

Yadav, V., Michalak, A. M., Ray, J., & Shiga, Y. P. (2016). A statistical approach for isolating fossil fuel emissions in atmospheric inverse problems. *Journal of Geophysical Research: Atmospheres*, *121*(20), 12,490-12,504. https://doi.org/10.1002/2016JD025642

Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S. M., et al. (2018). A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. *ISPRS Journal of Photogrammetry and Remote Sensing*, *146*, 108–123. https://doi.org/10.1016/j.isprsjprs.2018.09.006

Yi, C., Davis, K. J., Bakwin, P. S., Berger, B. W., & Marr, L. C. (2000). Influence of advection on measurements of the net ecosystem-atmosphere exchange of CO2 from a very tall tower. *Journal of Geophysical Research: Atmospheres*, *105*(D8), 9991–9999. https://doi.org/10.1029/2000JD900080

Zeng, N., Zhao, F., Collatz, G. J., Kalnay, E., Salawitch, R. J., West, T. O., & Guanter, L. (2014). Agricultural Green Revolution as a driver of increasing atmospheric CO 2 seasonal amplitude. *Nature*, *515*(7527), 394–397. https://doi.org/10.1038/nature13893

Zhou, Y., Williams, C. A., Lauvaux, T., Davis, K. J., Feng, S., Baker, I., et al. (2020). A Multiyear Gridded Data Ensemble of Surface Biogenic Carbon Fluxes for North America: Evaluation and Analysis of Results. *Journal of Geophysical Research: Biogeosciences*, *125*(2), e2019JG005314. https://doi.org/10.1029/2019JG005314

Zhou, Y., Williams, C.A., Lauvaux, T., Feng, S., Baker, I.T., Wei, Y., et al. (2020). Atmospheric Carbon and Transport - America (ACT-America)ACT-America: Gridded Ensembles of Surface Biogenic Carbon Fluxes, 2003-2019 (Version 1.1) [NetCDF], 0 MB. https://doi.org/10.3334/ORNLDAAC/1675

Tables & Figures

Table 1: Comparison of features across the three versions of VPRM.

	$VPRM_{ann}$	**$VPRM_{seas}$**	**$VPRM_{new}$**
Respiration model	original model: linear function of temperature	original model: linear function of temperature	expanded model including EVI, non-linear temperature and interactions with water stress
Parameter seasonality	no	yes (winter, spring, summer, fall)	no
Optimization technique	GPP and R_e parameters optimized simultaneously	GPP and R_e parameters optimized simultaneously	R_e parameters optimized with night-time data; GPP parameters optimized separately after subtracting predicted R_e from daytime NEE observations

Table 2: median across towers of the absolute monthly mean bias of simulated – observed biologic CO_2 enhancements for each biospheric model and month, using both WRF-STILT, NAMS-STILT and the mean of WRF-STILT and NAMS-STILT convolutions. Model/month combinations with a median absolute error less than 1 µmol/mol are shaded in light yellow, from 1 µmol/mol to 1.5 µmol/mol in dark yellow and > 1.5 µmol/mol in peach. The model(s) with the smallest median absolute bias (within 0.1 umol/mol) for each month is (are) highlighted in bold. The last row shows the value across all towers in the full year (where the number of towers varies by month).

	WRF-STILT					**NAMS-STILT**					**Mean WRF-STILT & NAMS-STILT**				
	$VPRM_{ann}$	$VPRM_{seas}$	$VPRM_{new}$	CASA	SiB4	$VPRM_{ann}$	$VPRM_{seas}$	$VPRM_{new}$	CASA	SiB4	$VPRM_{ann}$	$VPRM_{seas}$	$VPRM_{new}$	CASA	SiB4
201611	**0.54**	**0.46**	0.68	1.03	1.02	0.72	0.68	1.03	**0.54**	0.85	**0.59**	**0.64**	0.89	0.88	0.85
201612	1.72	2.39	1.82	**0.88**	**0.93**	2.04	2.66	2.21	1.29	**1.02**	1.87	2.63	2.07	1.15	**0.96**
201701	1.20	1.75	1.43	**1.03**	1.33	1.29	2.01	1.50	**1.15**	1.84	1.24	1.78	1.49	**0.99**	1.66
201702	0.67	**0.54**	**0.51**	1.55	1.35	**0.41**	0.59	0.71	1.62	1.27	**0.38**	0.56	0.65	1.54	1.31
201703	**0.46**	0.58	**0.43**	**0.53**	0.83	**0.32**	0.56	**0.35**	0.65	0.89	**0.42**	0.50	**0.33**	0.59	0.87
201704	**0.31**	0.57	0.54	0.63	0.82	0.72	**0.59**	**0.51**	0.64	1.23	**0.59**	**0.52**	**0.54**	**0.55**	1.14
201705	0.65	0.60	**0.41**	1.67	1.00	0.76	0.76	**0.44**	1.46	1.00	0.63	0.74	**0.37**	1.63	0.89
201706	**0.85**	1.56	1.05	1.46	1.41	2.18	1.01	1.02	**0.59**	0.87	1.43	0.96	0.98	1.02	**0.82**
201707	2.03	**1.01**	**1.10**	**1.05**	1.77	3.34	2.37	**1.07**	**1.08**	3.08	2.59	1.75	0.99	**0.76**	2.35
201708	1.59	**0.78**	0.88	0.95	**0.68**	2.29	0.90	1.49	**0.66**	**0.74**	1.58	0.92	1.26	**0.70**	0.90
201709	**0.82**	1.07	1.87	**0.87**	2.15	**0.67**	1.29	1.95	**0.72**	2.32	**0.86**	1.22	1.95	**0.84**	2.25
201710	**0.49**	**0.48**	0.72	1.46	0.96	0.64	0.75	**0.50**	1.37	1.20	**0.45**	0.61	0.61	1.33	1.11
Full year	***0.83***	***0.82***	***0.89***	*1.09*	*1.16*	***1.02***	***1.02***	***0.99***	***0.96***	*1.17*	**0.93**	**0.93**	**0.88**	**0.94**	1.14

Figure 1: Map of dominant land cover in domain at 0.02° in eastern USA and Canada, with a rectangle around the flux simulation domain (left panel). Also shown are the flux towers included in the historical parameter optimization (triangles), with the towers used for model evaluation in 2016/ 2017 explicitly labeled. (Labeled towers with star symbols are included in the 2017 evaluation, but not the historical parameter optimization.) Deciduous broadleaf forest and cropland pixels used for spatial aggregation are shown in the panel on the right, with these pixels selected as containing > 50 % coverage at 0.1° for VPRM and CASA, and > 25 % for SiB4 at 0.5° (Table S1; Figure S6).

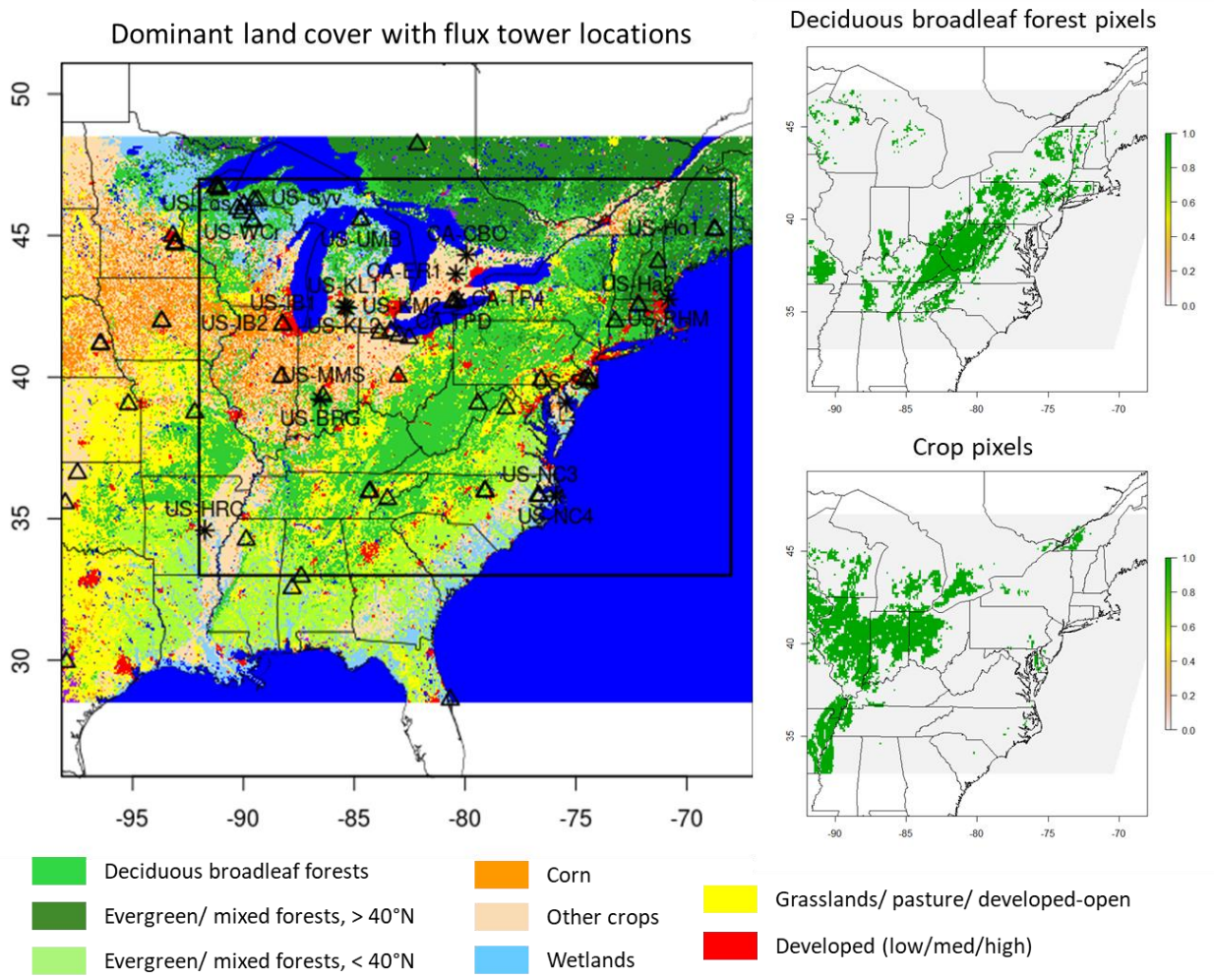


Figure 2: a) map of CO_2 observational towers and mean July 2017 afternoon atmospheric footprints (averaged across WRF-STILT and NAMS-STILT and summed across towers). The inner nests for the WRF simulation are shown in dark pink. (SNJ and SMT have no CO_2 observations in July 2017; therefore, their footprints are not included in the map.) b) and c) mean spatially integrated footprints in July 2017 as a function of hours back from receptor time for two towers: UNY (45 m inlet height) and MNC (213 m inlet height). Time series are averaged across all days in the month for each afternoon receptor hour. Receptor hours starting at 12 – 4 pm EST are shown with a thicker line width, although the expanded definition of "afternoon" in this study (as described in the supplemental material) allows for more hours with well-mixed conditions during summer months (shown with green shading). Other hours back in time are shaded to indicate day (yellow) or night (blue).

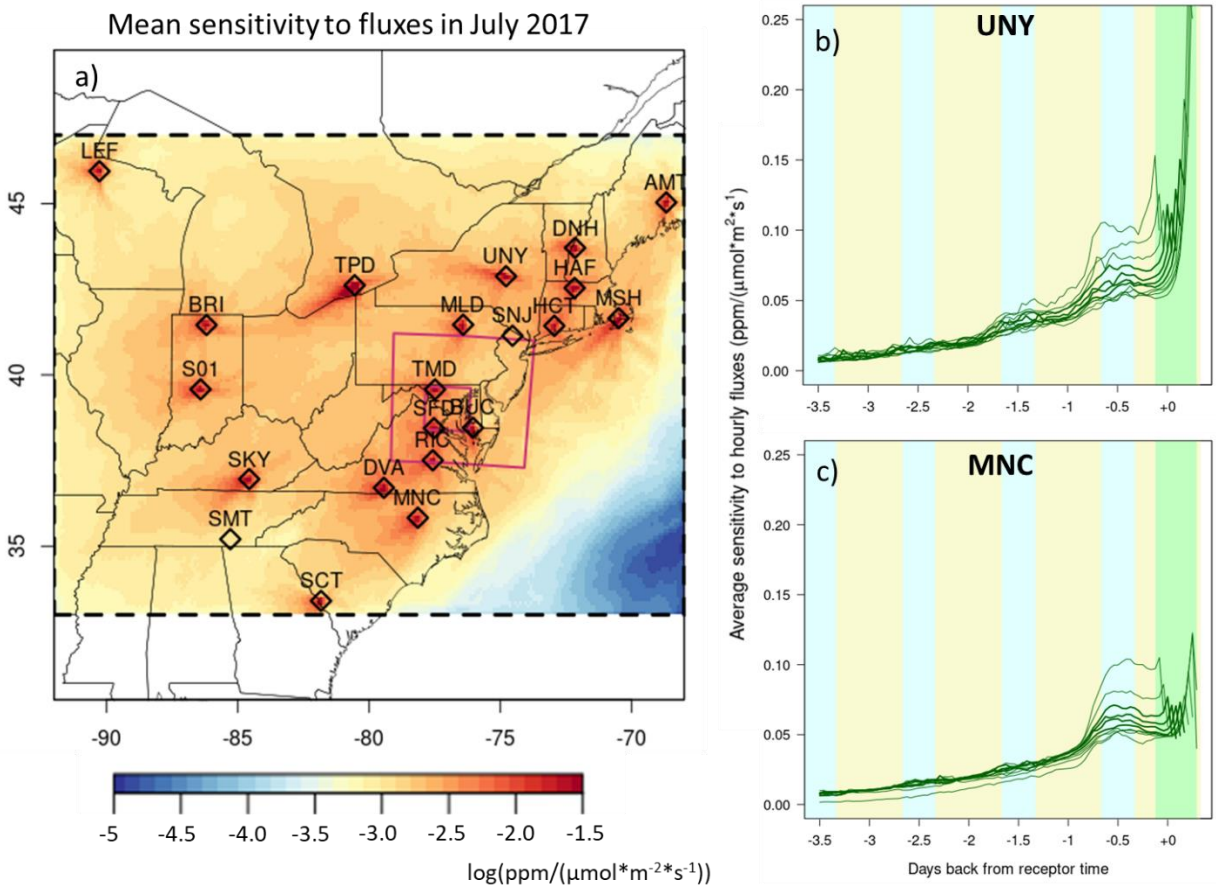


Figure 3: Atmospheric CO_2 observations across towers compared in four different ways: (a) total afternoon average CO_2 mole fraction for each tower across full year (in grey), with their average in red and background contribution from CT19 in blue, b) afternoon average biologic enhancements (using CT19 background) for each tower across full year (in grey), with difference in background conditions shown in orange, c) mean hourly diurnal cycle of biological enhancements with CT19 background conditions for each tower in January (grey) and the mean FF contribution across towers in gold, and d) the same as (c) but in July. For (c) and (d), afternoon hours have a thicker line width, with "afternoon" defined as described in the text. Red line indicates the spatial mean across towers in (a) and (b), and green for (c) and (d). Convolutions and background conditions are averaged across WRF-STILT and NAMS-STILT transport.

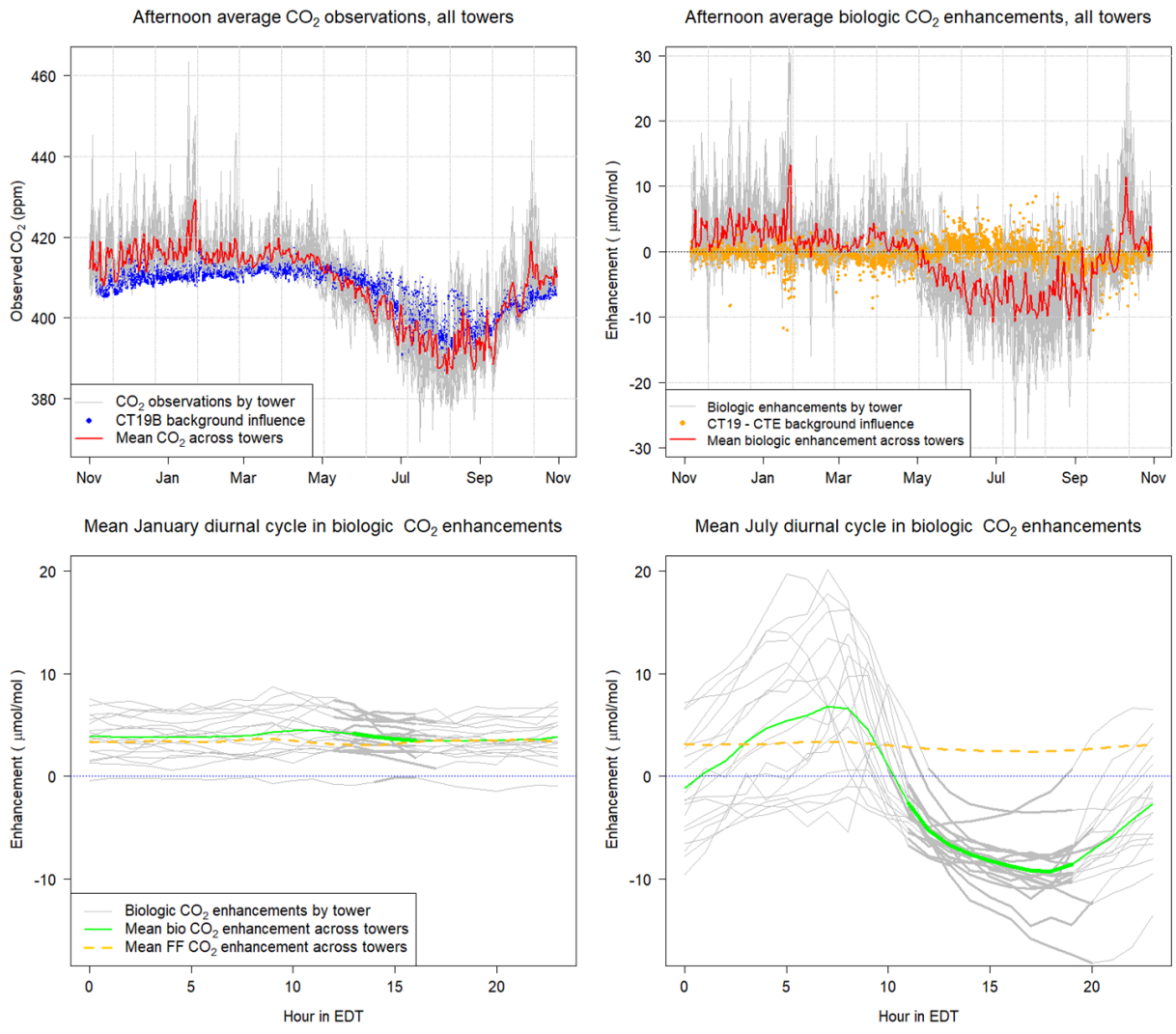


Figure 4: Scatter plots of observed air temperature vs. night-time average NEE for historical flux tower data used in the VPRM parameter optimization. The $VPRM_{ann}$ model fit is shown in yellow, $VPRM_{seas}$ in light green, and $VPRM_{new}$ in dark green, with four lines for $VPRM_{seas}$ corresponding to each season. Also shown are results from a linear regression model fit to just night-time NEE data (purple) for comparison. Results are shown for four PFT's (representing ~65% of total land cover in domain): deciduous broadleaf forests (27% + 5% urban), evergreen needleleaf/mixed forests >40°N (12%), maize crops (8%) and soybean/ other crops (13%). The NSC values (equivalent to the adjusted r^2 for $VPRM_{new}$ and the linear regression model) are also shown to assess relative performance for each model and PFT.

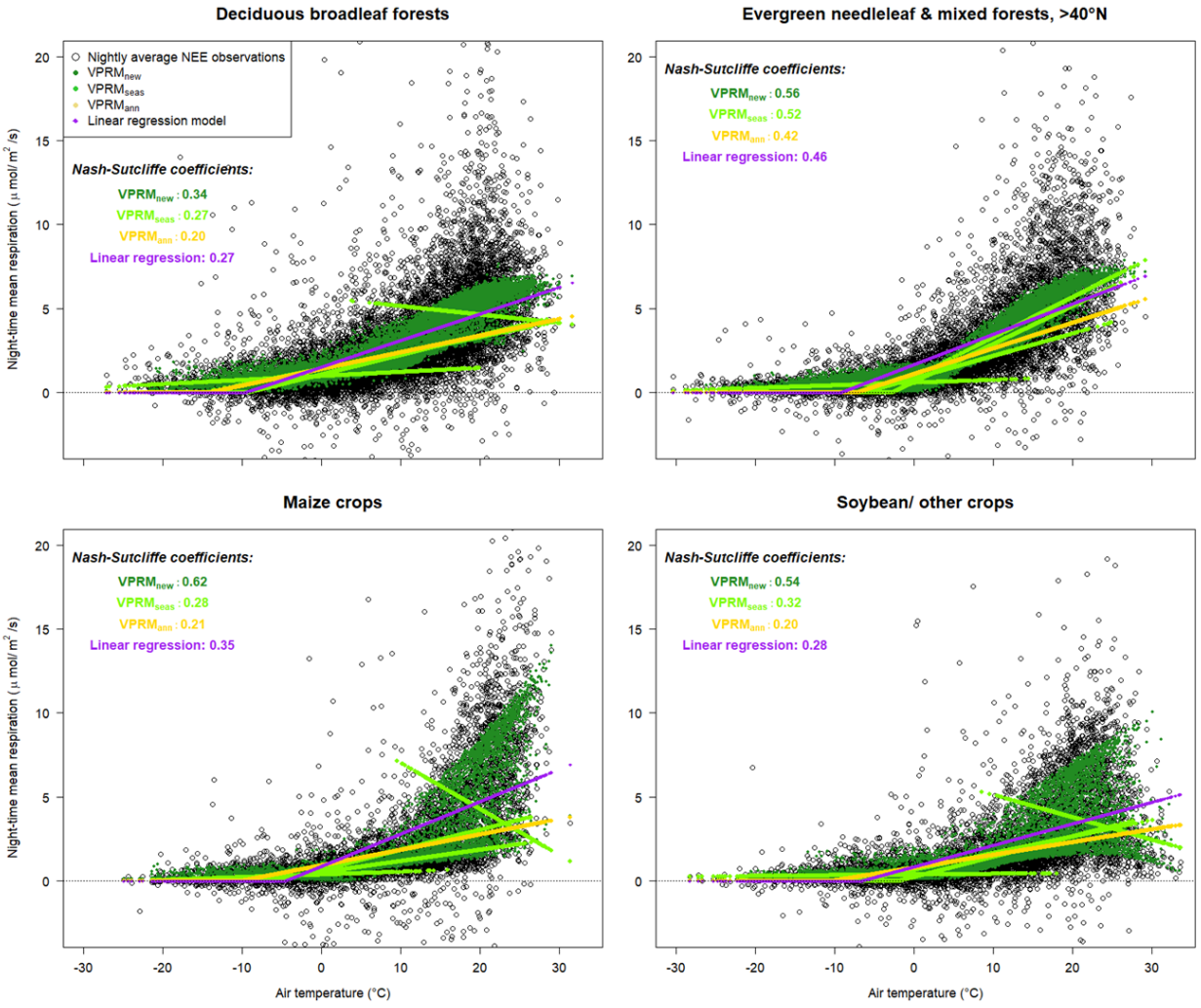


Figure 5: Mean 24-hour gridded GPP, ecosystem respiration (R_e) and NEE at 0.1° for SiB4, CASA and $VPRM_{new}$ in winter months (December/ January/ February).

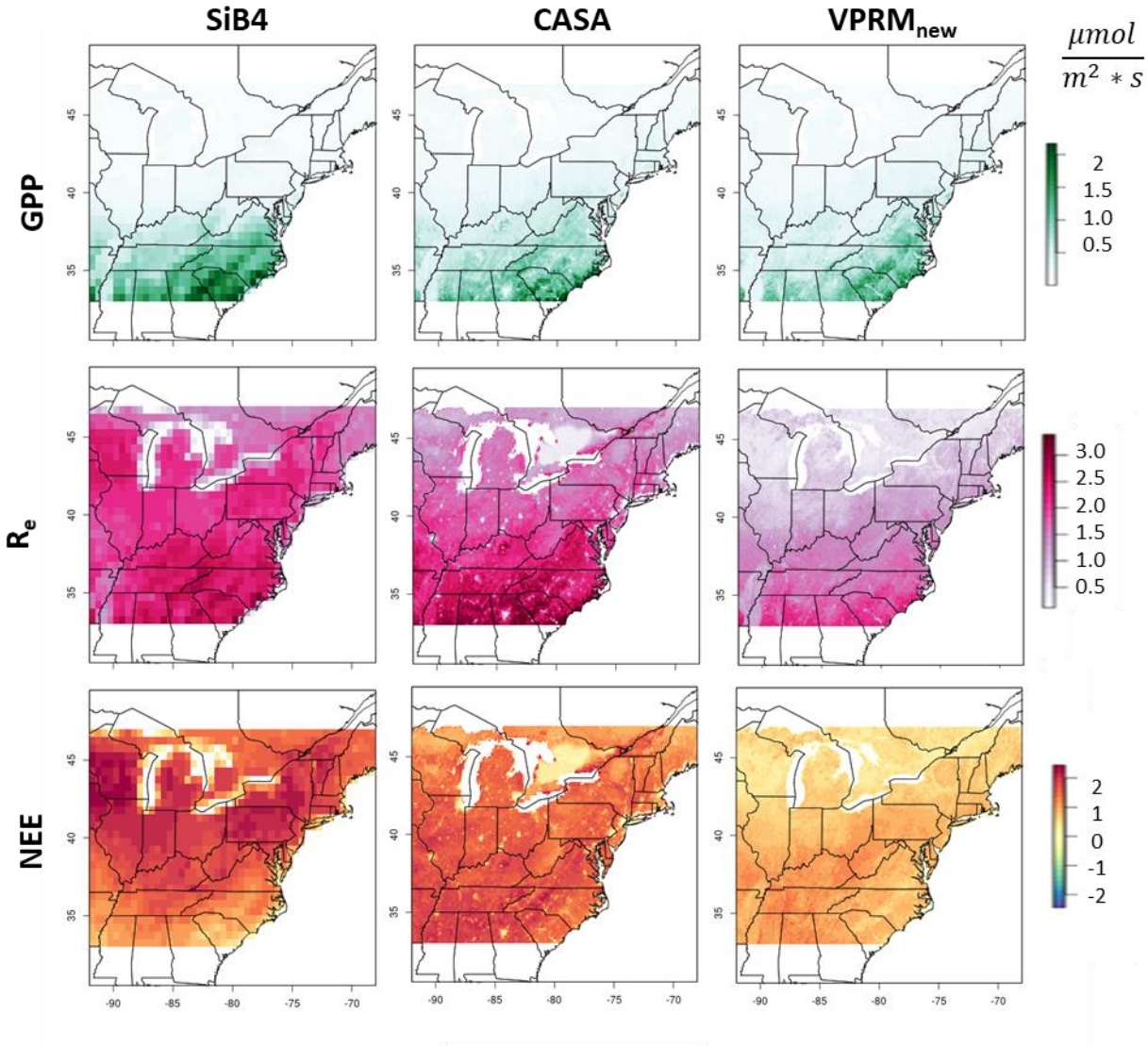


Figure 6: Mean 24-hour gridded GPP, ecosystem respiration (R_e) and NEE at 0.1° for SiB4, CASA and $VPRM_{new}$ in spring months (March/ April/ May).

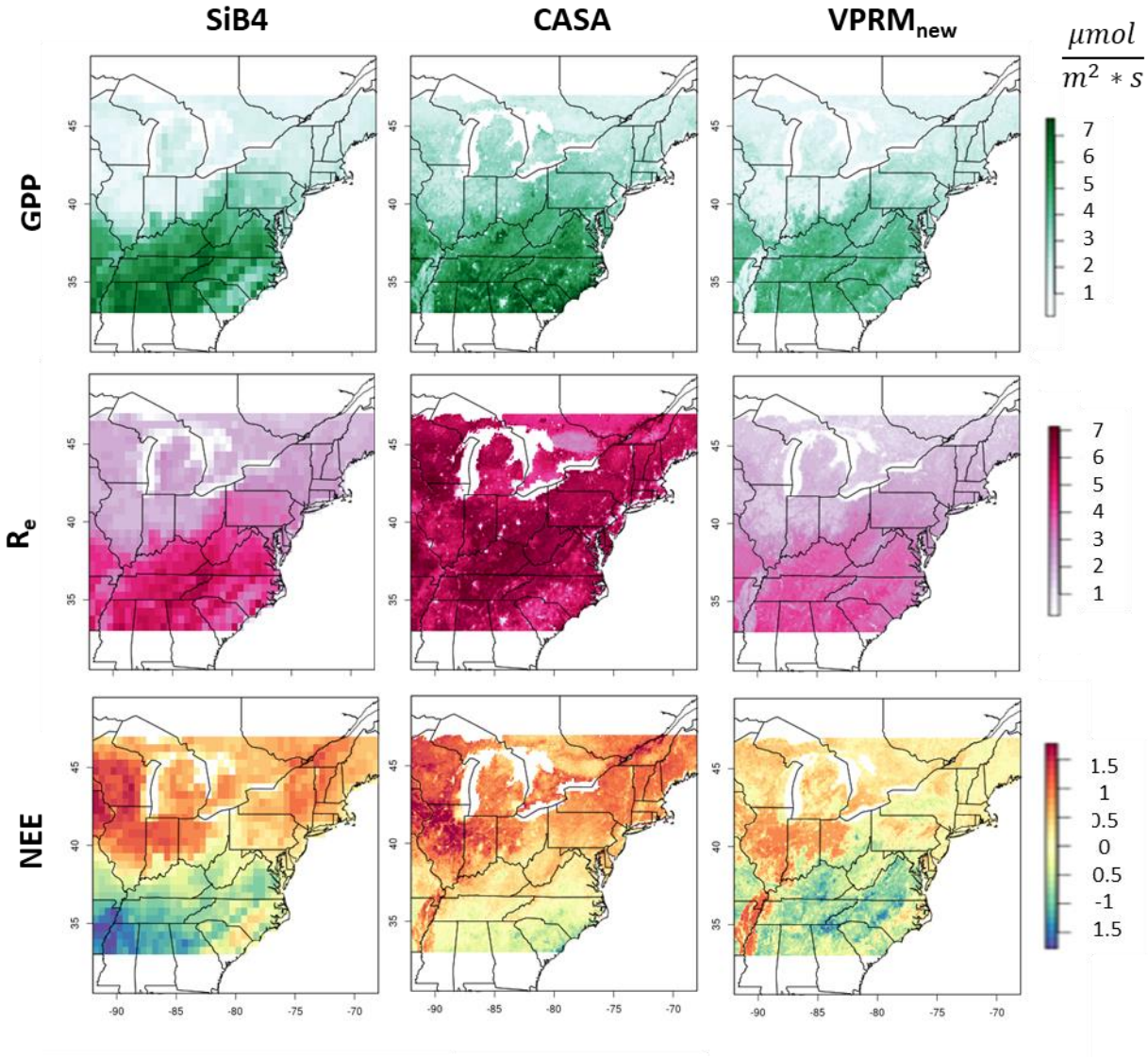


Figure 7: Mean 24-hour gridded GPP, ecosystem respiration (R_e) and NEE at 0.1° for SiB4, CASA and $VPRM_{new}$ in summer months (June/ July/ August).

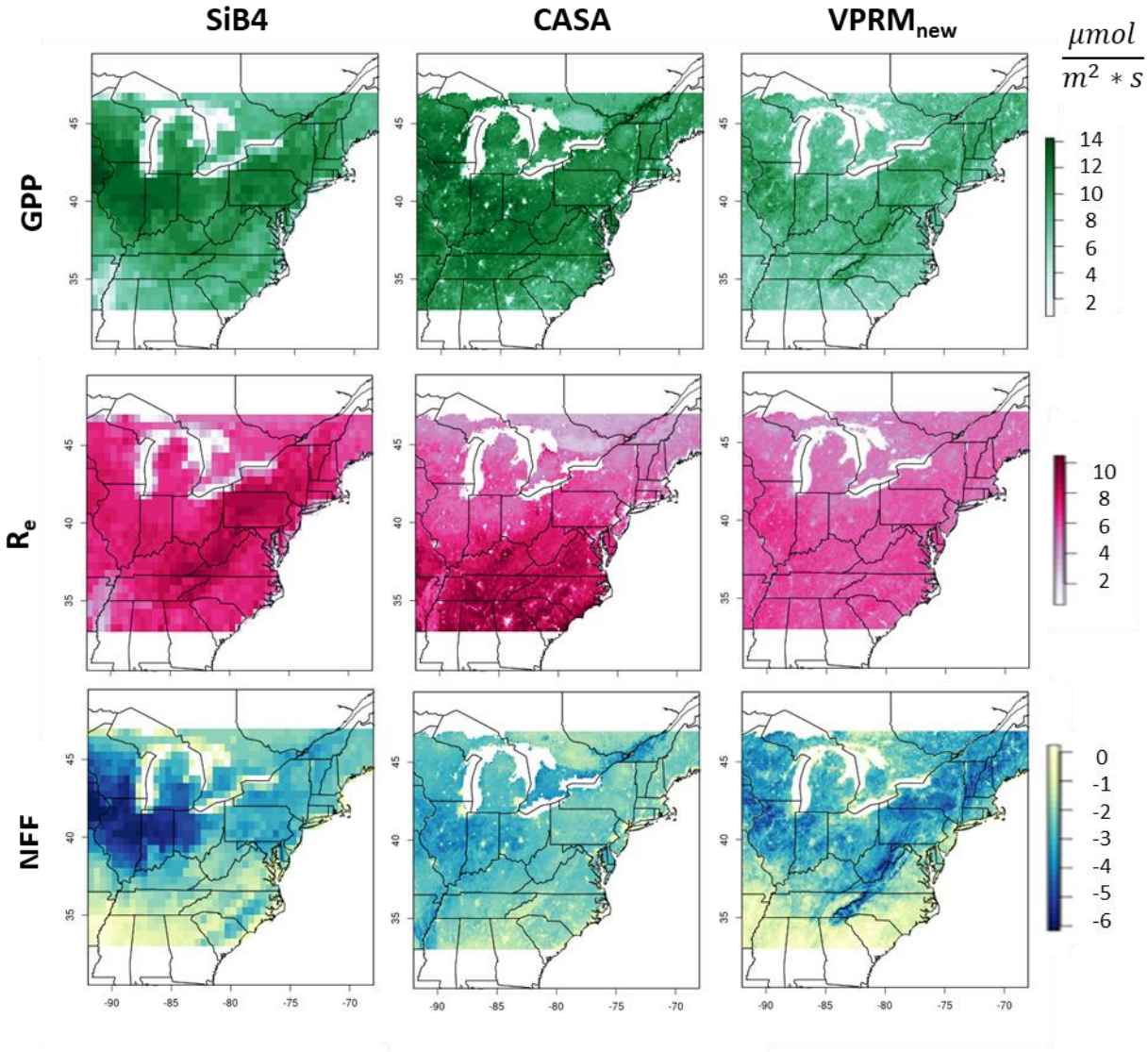


Figure 8: Mean 24-hour gridded GPP, ecosystem respiration (R_e) and NEE at 0.1° for SiB4, CASA and $VPRM_{new}$ in fall months (September/ October/ November).

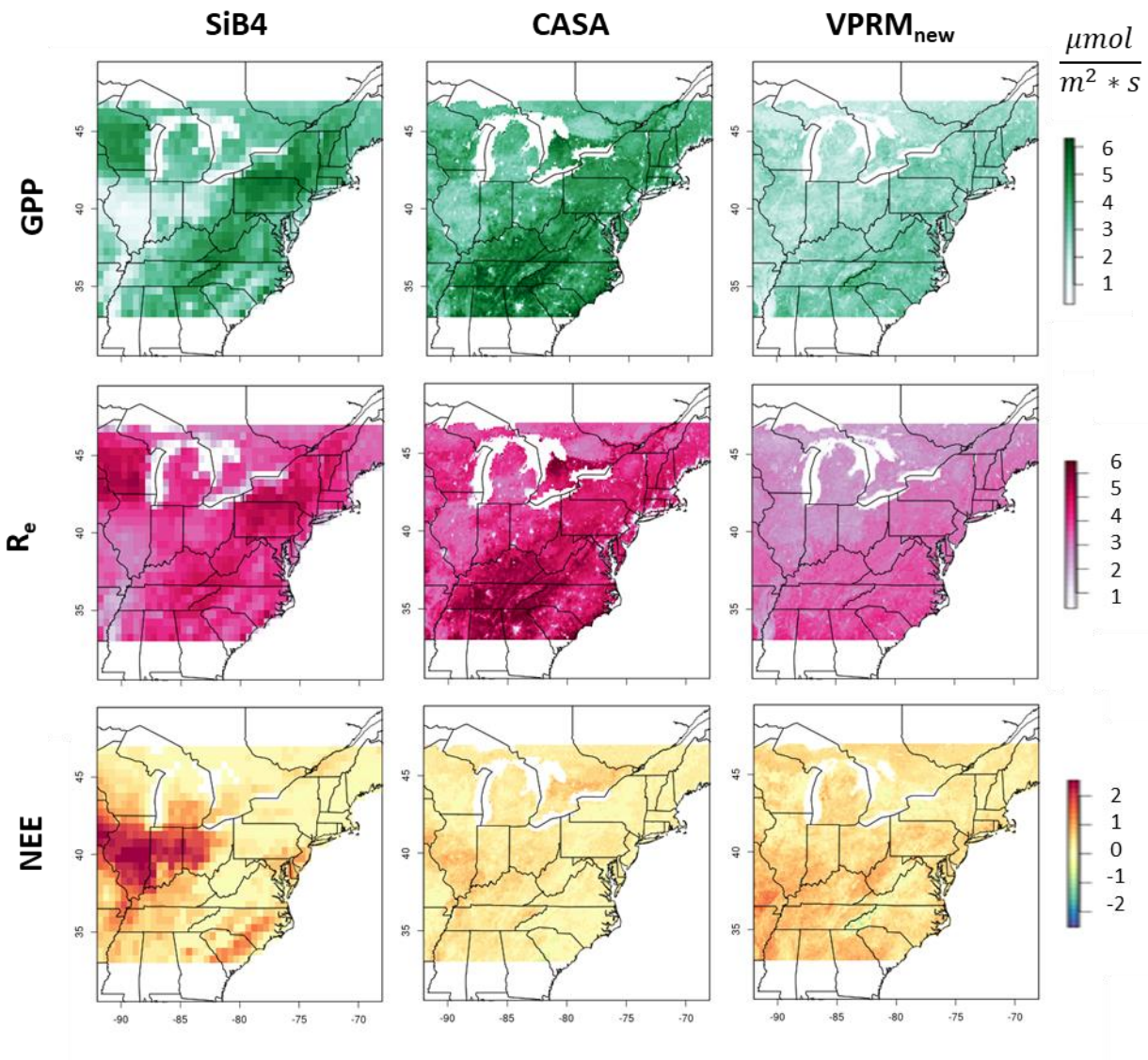


Figure 9: Seasonal cycle of weekly mean GPP, R_e & NEE fluxes, spatially aggregated across pixels with predominantly deciduous broadleaf forests (DBF, top row) and croplands (bottom row), as indicated in Figure 1. Annual means are shown with dashed lines.

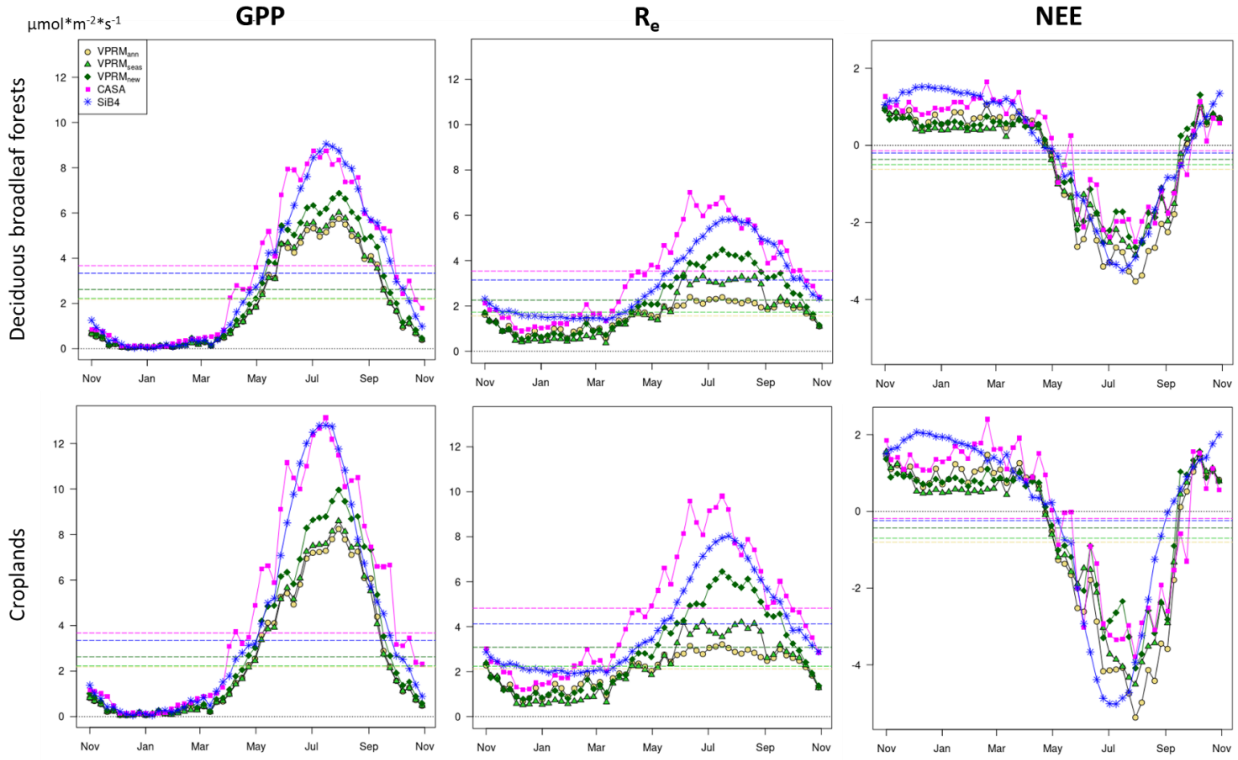


Figure 10: Comparison of $VPRM_{new}$, $VPRM_{seas}$, $VPRM_{ann}$, CASA and SiB4 to observed NEE at 22 flux towers in 2016/ 2017. Model output is extracted at flux tower locations at the 0.02° scale for VPRM, 500 m (or 5 km) for CASA and for the flux tower PFT in the SiB4 0.5° pixel. a) boxplot across towers of monthly mean biases (model – observations) for each model during nighttime hours. b) same as a), but for daytime hours. c) mean July NEE diurnal cycle comparing observations to models at the US-IB1 tower. d) same as c), but for the US-UMB tower. The 22 flux towers included in a) and b) are shown in Figure 1.

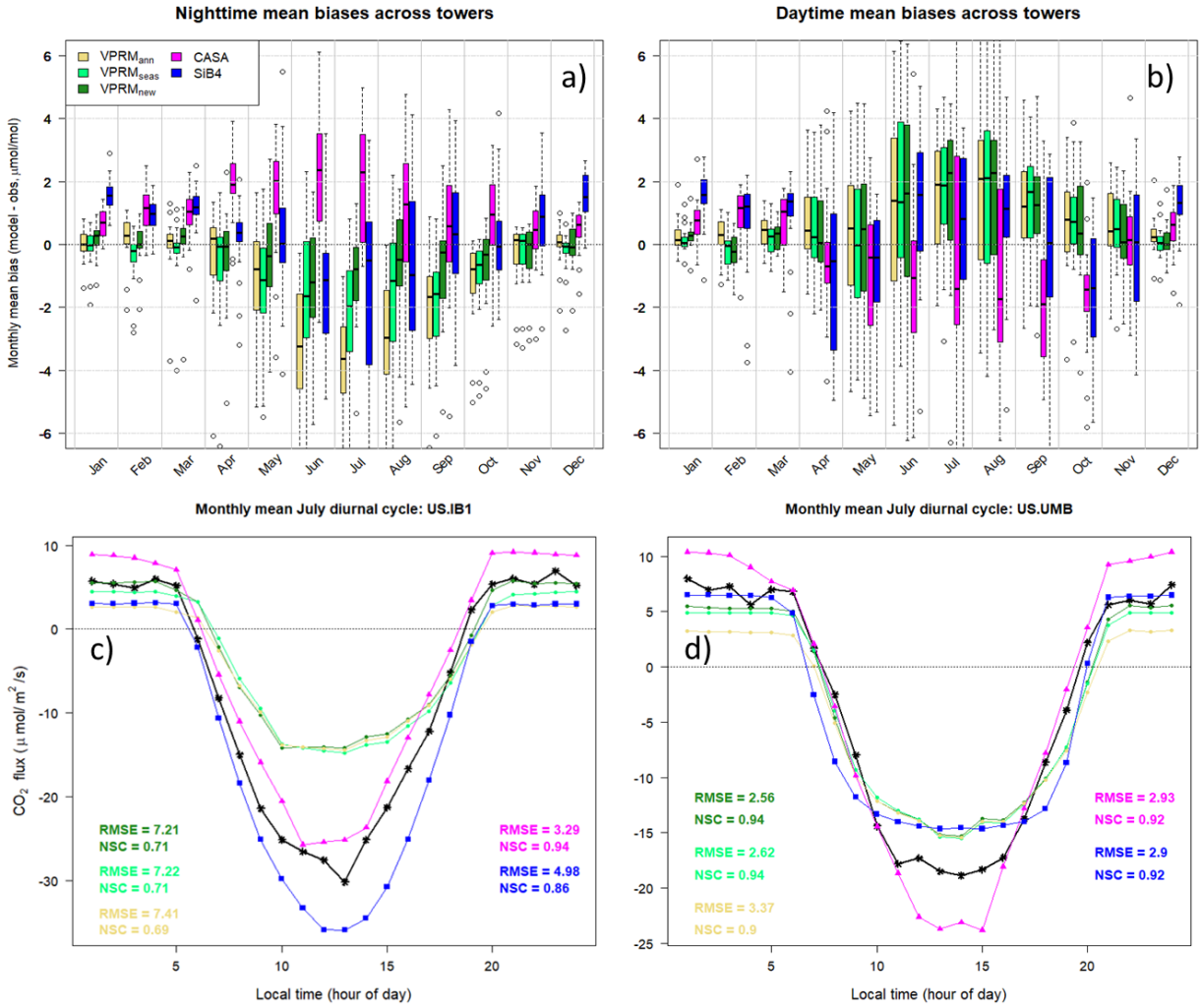
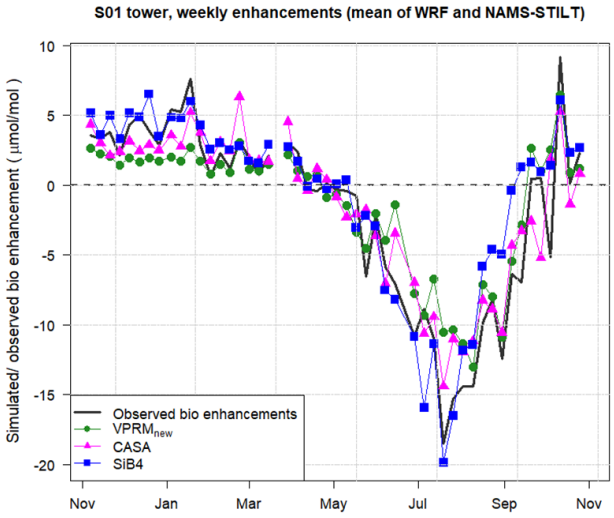
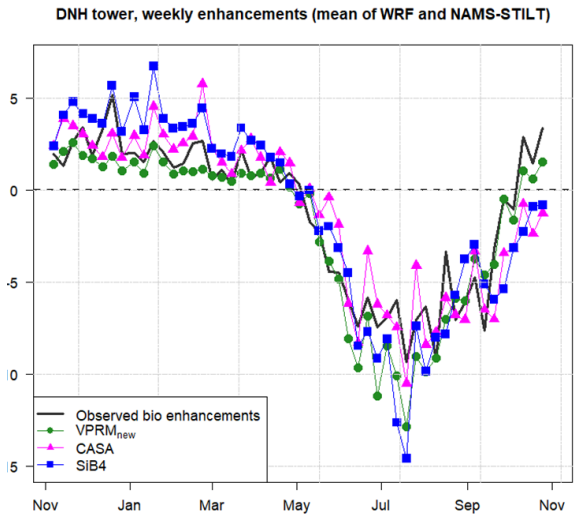


Figure 11: Weekly mean observed and simulated biological enhancements for $VPRM_{new}$, CASA and SiB4 at two [CO_2] towers: S01 in Indiana (44 % crops, 30 % forested within the footprint; left) and DNH in New Hampshire (63 % forested, 9 % crops within the footprint; right). The tables below show NSC and R^2_a metrics comparing weekly mean enhancements from May to October and November to April, with the best performing biospheric model highlighted in red and bold. Enhancements are determined using average convolutions with WRF-STILT and NAMS-STILT transport, and with "optimal" monthly background conditions and Vulcan 3.0 + FFDAS fossil fuel emissions. The same figures using WRF-STILT and NAMS-STILT transport alone are shown in Figures S13 and S14.

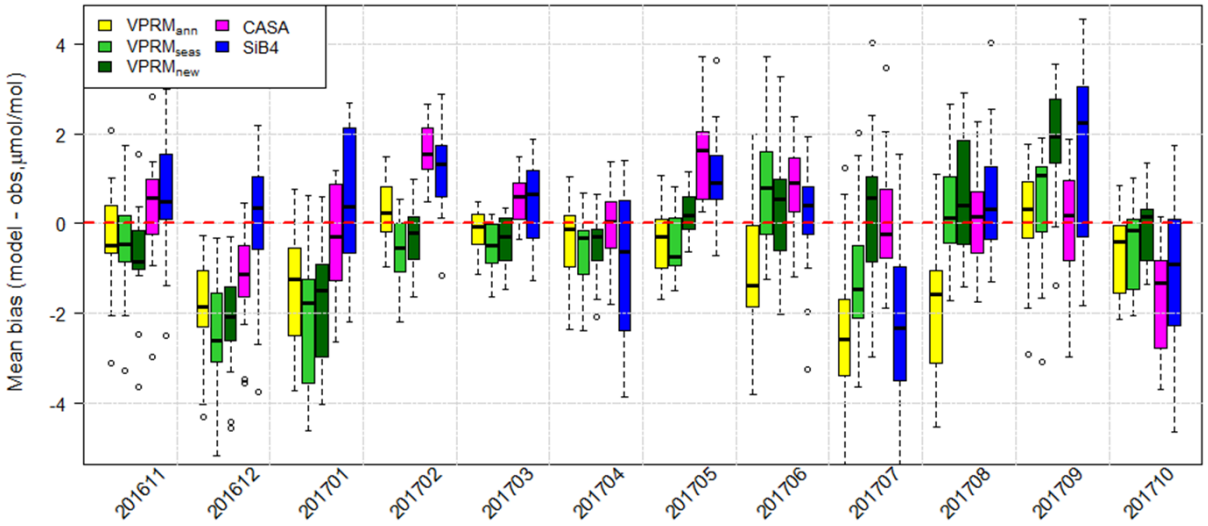




	NSC (May-Oct)	R^2_a (May-Oct)	NSC (Nov-Apr)	R^2_a (Nov-Apr)
$VPRM_{new}$	0.73	**0.82**	0.04	0.53
CASA	**0.76**	0.78	0.41	0.40
SiB4	0.66	0.73	**0.68**	**0.74**

	NSC (May-Oct)	R^2_a (May-Oct)	NSC (Nov-Apr)	R^2_a (Nov-Apr)
$VPRM_{new}$	**0.66**	**0.80**	-0.02	0.30
CASA	0.55	0.55	-0.56	0.13
SiB4	0.28	0.50	-1.71	**0.47**

Figure 12: Boxplots across towers of monthly mean biases (simulated - observed biospheric CO_2 enhancements) from November 2016 to October 2017 for each biospheric model, using mean of WRF-STILT and NAMS-STILT convolutions, "optimal" background conditions and Vulcan 3.0 + FFDAS fossil fuel emissions. The table indicates the mean absolute error across towers for winter (Dec/Jan/Feb), spring (Mar/Apr/May), summer (Jun/Jul/Aug) and fall (Sep/Oct/Nov) months for each biospheric model, calculated using the monthly mean bias for each tower. Numbers in bold in the table indicate the least biased biospheric model(s) in each season (including models within 0.1 of the minimum). The same figures using WRF-STILT and NAMS-STILT transport alone are shown in Figure S15.



MAE across towers	$VPRM_{ann}$	$VPRM_{seas}$	$VPRM_{new}$	CASA	SiB4
Dec-Feb	1.47	1.94	1.70	**1.11**	1.25
Mar-May	**0.57**	0.71	**0.61**	0.97	1.19
Jun-Aug	1.73	1.20	**1.03**	1.19	1.64
Sep-Nov	1.25	**1.00**	1.30	1.20	1.67

Figure 13: Monthly Nash-Sutcliffe coefficients (left panel) and adjusted R^2 (right panel) comparing convolutions from each biospheric model to observed hourly biologic enhancements across all towers. Convolutions using WRF-STILT and NAMS-STILT transport are averaged, and Vulcan3.0 (+FFDAS in Canada) fossil fuel emissions and "optimal" monthly background conditions are used for all comparisons. The same plots using WRF-STILT or NAMS-STILT transport alone are shown in the supplemental material in Figure S17.