Using atmospheric observations to evaluate the spatiotemporal variability of CO_2 fluxes simulated by terrestrial biospheric models

Yuanyuan Fang[1], Anna M. Michalak[1], Yoichi P. Shiga[1,2], Vineet Yadav[1]

[1] Department of Global Ecology, Carnegie Institution for Science, Stanford, CA, USA

[2] Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA

Correspondence to: Yuanyuan Fang (yyfang@stanford.edu)

Abstract

Terrestrial biospheric models (TBMs) are used to extrapolate local observations and process-level understanding of land-atmosphere carbon exchange to larger regions, and serve as a predictive tool for examining carbon-climate interactions. Understanding the performance of TBMs is thus crucial to the carbon cycle and climate science. In this study, we present and assess an approach for evaluating the spatiotemporal patterns, rather than aggregated magnitudes, of net ecosystem exchange (NEE) simulated by TBMs using atmospheric CO_2 measurements. The approach is based on statistical model selection implemented within a high-resolution atmospheric inverse model. Using synthetic data experiments, we find that current atmospheric observations are sensitive to the underlying spatiotemporal flux variability at sub-biome scales for a large portion of North America, and that atmospheric observations can therefore be used to evaluate simulated spatiotemporal flux patterns as well as to differentiate among multiple competing TBMs. Experiments using real atmospheric observations and four prototypical TBMs further confirm the applicability of the method, and demonstrate that the performance of TBMs in simulating the spatiotemporal patterns of NEE varies substantially across seasons, with best performance during the growing season and more limited skill during transition seasons. This result is consistent with previous work showing that the ability of TBMs to model flux magnitudes is also seasonally-dependent. Overall, the proposed approach provides a new avenue for evaluating TBM performance based on sub-biome scale flux patterns, presenting an opportunity for assessing and informing model development using atmospheric observations.

1 Introduction

A key question in carbon cycle science is how terrestrial carbon sinks will evolve within the context of a rapidly changing climate. Such projections of future carbon-climate interactions largely depend on the accuracy of current terrestrial biospheric models (TBMs), the main tool used to simulate the processes controlling the biospheric carbon cycle. Thus, understanding and evaluating the performance of current TBMs is an essential step toward improving the state of carbon cycle research.

TBM predictions of carbon flux can be directly evaluated against eddy covariance tower measurements at various time scales ranging from hourly to interannual (Baker et al., 2003; Balzarolo et al., 2014; Keenan et al., 2012; Raczka et al., 2013; Richardson et al., 2012; Sasai et al., 2005; Schaefer et al., 2012; Schwalm et al., 2010), but the information provided by flux towers is only representative of small spatial scales (~1km^2) relative to the scales of interest for global analyses. On the other end of the spectrum, TBM predictions aggregated to large spatial and/or temporal scales (*e.g.*, continental/monthly to global/annual) are routinely intercompared with flux estimates obtained from inverse-modeling based on observed atmospheric CO_2 mixing ratios (Canadell et al., 2011; Gourdji et al., 2012; Hayes et al., 2012; McGuire et al., 2012; Turner et al., 2011), but such large-scale comparisons make it difficult to provide directly usable information regarding the processes driving carbon exchange. In addition, differences among TBMs exist across a full range of spatiotemporal scales, including inter-annual variability, the timing of phenology, and the spatiotemporal distribution of biospheric carbon fluxes within regions (Gourdji et al., 2012; Huntzinger et al., 2012; Keenan et al., 2012; Raczka et al., 2013; Richardson et al., 2012; Schaefer et al., 2012; Schwalm et al., 2010). These differences reflect

52 the fact that processes controlling carbon-climate feedbacks are manifested differently across
53 TBMs.

54 Assessing the spatial and/or temporal variability of carbon fluxes as a method for evaluating
55 TBMs, therefore, offers the potential to examine the environmental processes driving carbon
56 exchange, and hence provides an alternative path forward in the assessment of TBM predictions.
57 For example, evaluating the timing of modeled phenology can highlight issues associated with a
58 model's representation of Light Use Efficiency (LUE), temperature response, and GPP response
59 under various conditions (Richardson et al., 2012; Schwalm et al., 2010). Examining the
60 interannual variability of TBM output can identify problems with the representation of
61 interannual variability in spring phenology, soil thaw, snowpack melt and lagged response to
62 extreme climatic events (Keenan et al., 2012).

63 The majority of previous studies examining carbon flux variability are still based on spatially
64 and/or temporally aggregated carbon fluxes, however. An evaluation of flux variability, or flux
65 patterns, at the fine native spatiotemporal scales of TBM simulations would make it possible to
66 more directly target the fine-scale spatiotemporal patterns of carbon fluxes that have been shown
67 to directly relate to environmental/climatic factors, such as precipitation, radiation and nighttime
68 temperature (Beer et al., 2010; Mueller et al., 2010; Yadav et al., 2010). Such evaluations could
69 therefore inform model improvements at the process level.

70 Observations of atmospheric CO_2 can potentially be used to assess such fine-scale
71 spatiotemporal flux patterns. On one hand, atmospheric CO_2 observations are sensitive to fine-
72 scale NEE spatial and temporal variability (Huntzinger et al., 2011). On the other hand,
73 variations in atmospheric CO_2 measurements are routinely used in inverse modeling frameworks
74 to infer upwind sources and sinks of CO_2, and recent studies suggest that atmospheric

observations contain information about flux patterns at spatial and temporal resolutions comparable to those of TBMs run for regional to continental domains (Broquet et al., 2013; Göckede et al., 2010; Gourdji et al., 2010; Gourdji et al., 2012). Despite the uncertainties existing in regional inversions due to uncertainties in atmospheric transport, fossil fuel emissions, fire disturbance, and boundary conditions, these studies do point to the possibility of evaluating the spatiotemporal patterns of fluxes from biospheric models through the use of inverse models.

With this goal in mind, what is needed is an atmospheric-inversion-based method that can use variations in atmospheric CO_2 to assess the spatiotemporal patterns of surface carbon fluxes simulated by TBMs. The purpose of this paper is to present, evaluate, and demonstrate the application of such an approach, applied here to the evaluation of the $1°\times1°$ and 3-hourly spatiotemporal variability of Net Ecosystem Exchange (NEE) simulated by TBMs using atmospheric CO_2 measurements. This fine scale variability is evaluated here within each month and biome over North America, thus providing a way to evaluate the seasonal and biome-specific differences in model performance. The distinguishing feature of the proposed approach is that it targets the evaluation of flux patterns at fine scales, rather than flux magnitudes at aggregated scales, thereby providing a closer link to process-based understanding of TBM performance. The approach is first evaluated with a series of synthetic data experiments where the underlying flux patterns affecting the atmospheric CO_2 signals are known. The application of this approach is further tested and demonstrated using actual atmospheric measurements and a prototypical small set of extensively studied TBM simulations from the North American Carbon Program (NACP) Regional Interim Synthesis (RIS) effort (Huntzinger et al., 2012).

2 Data description

2.1 Atmospheric CO_2 measurements

We use continuous, high-precision atmospheric CO_2 concentration measurements from 35 towers for the year 2008 to evaluate the simulated NEE spatiotemporal variability over North American land. Figure 1 shows the location of these towers along with the geographic coverage of seven North American biomes as modified from Olson et al. (2001). A majority of towers are located in Temperate Broadleaf and Mixed Forests, Temperature Grasslands, Savannas and Shrublands, Temperature Coniferous Forests and Boreal Forests and Taiga, while very few towers are located in the other biomes (Tundra, Desserts and Xeric, and Tropical and Subtropical biomes). This distribution of towers is expected to affect the sensitivity of atmospheric CO_2 data to NEE within those biomes. The year 2008 is used as it includes the expansion of continuous measurement locations from the Mid-Continent Intensive (MCI) project (Miles et al., 2012; Ogle et al., 2006). Atmospheric CO_2 measurements are processed and averaged to 3-hourly intervals as described in Gourdji et al. (2012). Data from all hours of the day are used for tall towers with a height over 300m while afternoon data are used for most short towers (lower than 100m) and nighttime data are used for sites with complex topography (e.g. Niwot Ridge - NWR), as detailed in Shiga et al. (2014). We further remove data that are strongly influenced by only a few 1°×1° grid cells, in order to exclude data that are likely subject to systematic transport model errors (Göckede et al., 2010; Gourdji et al., 2012; Peters et al., 2007). The total number of resulting observations is $n = 28{,}717$.

To remove the effect of boundary conditions, we pre-subtract the GLOBALVIEW-CO2 boundary condition (GLOBALVIEW-CO2, 2010) from atmospheric measurements as in Gourdji et al. (2012). We further remove the impact of fossil fuel emissions by pre-subtracting

concentrations modeled based on the VULCAN-ODIAC fossil fuel emissions inventory (Shiga et al., 2014).

2.2 Sensitivity footprints from atmospheric transport model

The sensitivity of the available atmospheric observations to underlying CO_2 fluxes (in units of ppmv/(μmol $m^{-2}s^{-1}$)) is quantified as described in Gourdji et al. (2012). In brief, footprints are derived from the Stochastic Time-Inverted Lagrangian Transport (STILT) model (Lin et al., 2003), driven by meteorological fields from the Weather Research and Forecast (WRF) model (Skamarock and Klemp, 2008). The STILT transport model has been used and examined extensively at regional and continental scales (Chatterjee et al., 2012; Gourdji et al., 2010; Gourdji et al., 2012; Huntzinger et al., 2011; Kort et al., 2008; McKain et al., 2012). Footprints are also used to generate synthetic observational time series based on TBM flux simulations.

2.3 Terrestrial Biospheric Models (TBMs)

We use simulations from four TBMs to evaluate the proposed approach, namely CASA-GFED (van der Werf et al., 2006), SiB3 (Baker et al., 2008), ORCHIDEE (Krinner et al., 2005) and VEGAS2 (Zeng et al., 2005), using the runs submitted to the NACP RIS activity. These four models were selected for analysis because of the availability of 3-hourly NEE flux output. While CASA-GFED and VEGAS2 have a coarser native temporal resolution, their NEE fluxes have been downscaled to a 3-hourly resolution as described in Huntzinger et al. (2011). Our evaluation is based on the overall NEE simulated by each TBM, although model definitions of NEE differ: CASA-GFED includes fire disturbance while other models do not; ORCHIDEE exclude crop harvest while others do not. A comparison and summary of these simulations can be found in Table S1 in the supplementary material. Further details on the NACP RIS simulations can be found in Huntzinger et al. (2012).

3 Regression framework linking atmospheric CO_2 to NEE

The overall goal of the proposed approach is to evaluate the spatiotemporal variability of NEE as simulated by various TBMs using atmospheric CO_2 measurements. Such an approach must be based on an inverse model that can infer NEE from atmospheric CO_2 measurements. It must also include a statistical model selection component to evaluate the degree to which NEE patterns predicted by TBMs are useful in explaining the observed atmospheric CO_2 variability. Rather than quantifying the magnitude of NEE, the primary goal here is to evaluate the spatiotemporal NEE patterns (at a 1°×1° and 3-hourly resolution) within specific biomes of North America and for specific months. The approach presented here builds on the geostatistical inverse modeling (GIM) framework (Gourdji et al., 2010; Gourdji et al., 2012; Michalak et al., 2004), but is presented here in the form of a regression analysis to simplify the presentation and emphasize the introduction of model selection aspect of the proposed approach.

To this end, we first formulate a multi-linear regression framework that relates atmospheric observations to NEE spatiotemporal variability. Statistical model selection is then applied to determine whether, when, and where the spatiotemporal variability of simulated NEE is consistent with that evident from variability in atmospheric CO_2. Here, the NEE spatiotemporal variability is defined at a 1°×1° spatial and 3-hourly temporal resolution, and the TBMs are evaluated within specific biome-month combinations. Figure 2 shows the distribution of NEE in one specific biome-month combination (*i.e.*, Boreal Forests and Taiga in July) as an example.

To link atmospheric measurement to surface fluxes we first define the observed atmospheric CO_2 concentrations, with the influence of boundary conditions and fossil fuel emissions pre-subtracted, as:

$$\mathbf{z} = \mathbf{Hs} + \boldsymbol{\varepsilon} \quad (1)$$

where $\mathbf{z}$ is an $n \times 1$ vector of atmospheric CO_2 observations, $\mathbf{s}$ is an $m \times 1$ vector of the underlying NEE fluxes at 1°×1° and 3-hourly resolution, $\mathbf{H}$ $(n \times m)$ are the sensitivity footprints, namely a Jacobian matrix representing the sensitivity of each observation to each underlying flux (i.e., $\frac{\partial \mathbf{z}_i}{\partial \mathbf{s}_j}$) as quantified using an atmospheric transport model (see Section 2.2), and $\boldsymbol{\varepsilon}$ $(n \times 1)$ is the model-data mismatch term that represents any discrepancies between observed (**z)** and modeled (**Hs**) CO_2 mixing ratios. The model-data mismatch term encompasses the influence of errors in the boundary conditions, errors in the fossil fuel inventory, representation errors, aggregation errors, transport model errors, and measurement errors. These errors are assumed to have zero mean and be uncorrelated across measurements, with their variances represented by a diagonal covariance matrix $\mathbf{R}$ $(n \times n)$. The dimensions of the matrices and vectors are based on the total number of observations, $n = 28{,}717$, and the total number of fluxes at a 1°×1° (2635 such grid cells within the domain used here) and 3-hourly resolution $(366 \times 8 = 2928)$ such periods within the span of the one-year inversion), $m = 2635 \times 2928 = 7{,}715{,}280$.

The spatiotemporal NEE distribution of $\mathbf{s}$ is represented as a linear model of NEE as predicted by various TBMs within specific biome-month combinations:

$$\mathbf{s} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} \quad (2)$$

where $\mathbf{X}$ is a $m \times p$ matrix with each column representing NEE 1°×1° 3-hourly spatiotemporal variability within a specific biome-month combination from a specific TBM, such that a given column is populated by the modeled NEE from a given TBM for a given biome-month for those rows (i.e. elements of $\mathbf{s}$) corresponding to that specific biome-month combination, while the

remainder of the column is filled with zeros. These individual columns of **X** are thus predictor variables for the dependent variable **s**. With 7 biomes (Figure 1) and 12 months, there are a total of 84 possible predictor variables for each TBM (i.e., $p \leq 84$ for one TBM). The $p \times 1$ vector $\boldsymbol{\beta}$ represents the drift coefficient describing the relationship between **X** and **s,** and $\mathbf{X}\boldsymbol{\beta}$ together thus represents a statistical model of the trend of NEE. The $m \times 1$ vector $\boldsymbol{\xi}$ represents the portion of the variability of **s** that cannot be explained by the predictor variables in **X**, and these deviations are modeled as having a mean of zero and a covariance matrix $\mathbf{Q}$ $(m \times m)$ that represents how the flux deviations from the model of the trend (i.e., $\mathbf{s} - \mathbf{X}\boldsymbol{\beta}$) are correlated in time and space.

Combining these two equations, we represent the atmospheric observations **z** in terms of the NEE predictor variables **X**:

$$\mathbf{z} = \mathbf{H}\mathbf{X}\boldsymbol{\beta} + \mathbf{H}\boldsymbol{\xi} + \boldsymbol{\varepsilon} \tag{3}$$

where **z** is seen to have a spatiotemporally variable mean $\mathbf{H}\mathbf{X}\boldsymbol{\beta}$ and, assuming independence between $\boldsymbol{\xi}$ and $\boldsymbol{\varepsilon}$, a residual covariance of:

$$\boldsymbol{\Sigma} = \mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R} \tag{4}$$

where T is the matrix transpose operation. From a statistical standpoint, our goal then becomes to select a subset of TBM biome-month combinations that capture a substantial portion of the CO_2 variability observed in **z**. This constitutes a classical statistical model selection problem, in which we examine which predictor variables (candidate columns in **X**) are useful in explaining the atmospheric CO_2 measurements (**z**).

A widely applied approach for statistical model selection is the Bayesian Information Criterion (BIC) (Schwarz, 1978). BIC takes into account both the goodness of fit, i.e., the residual sum of squares (RSS), and the numbers of auxiliary variables (k) in each candidate model, and can be

used to compare non-nested candidate models. BIC has also been adapted for use with spatiotemporally autocorrelated residuals (Hoeting et al., 2006; Mueller et al., 2010) and within the context of atmospheric inversions where atmospheric observations are used to inform underlying surface fluxes (Gourdji et al., 2012), making it ideal for the application presented here. The standard expression for BIC is:

$$BIC = \underbrace{\ln|\mathbf{\Sigma}| + RSS}_{\log likelihood} + \underbrace{k ln(n)}_{\substack{penalty \\ term}} \tag{5}$$

where RSS represents the residual sums of squares of a given candidate model $\mathbf{X}_c$, $\mathbf{\Sigma}$ is the $n \times n$ covariance matrix of the residuals (Eq. 4), $|\,|$ denotes the matrix determinant, and k is the number of parameters in a particular candidate model. For the specific application presented here (Eq. 1-4) and factoring out the unknown drift coefficients, $\boldsymbol{\beta}$ and RSS become as in Gourdji et al. (2012):

$$\boldsymbol{\beta} = \left((\mathbf{H}\mathbf{X}_c)^T \mathbf{\Sigma}^{-1} (\mathbf{H}\mathbf{X}_c)\right)^{-1} (\mathbf{H}\mathbf{X}_c)^T \mathbf{\Sigma}^{-1} \boldsymbol{z} \tag{6}$$

$$RSS = \left[\mathbf{z}^T \left(\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}(\mathbf{H}\mathbf{X}_c)\left((\mathbf{H}\mathbf{X}_c)^T \mathbf{\Sigma}^{-1} (\mathbf{H}\mathbf{X}_c)\right)^{-1} (\mathbf{H}\mathbf{X}_c)^T \mathbf{\Sigma}^{-1}\right) \mathbf{z}\right] \tag{7}$$

The specific covariance parameters needed to define **Q** and **R**, which are needed to define **Σ,** vary between experiments and are obtained as described in the supplementary materials.

Model selection built on this framework aims to identify the "best" model of the trend based on a tradeoff between model size and the model's power in explaining the variations in observed atmospheric CO_2. Here, the "best" model is specially defined as one with the minimum BIC value, providing an optimal balance between model complexity and model fit. To identify this model, BIC is compared across all possible combinations of predictor variables (i.e. 84 NEE biome-months per TBM). Due to the large number of candidate predictor variables considered

here, we implement the branch-and-bound algorithm of Yadav et al. (2013) to improve computational efficiency.

The final selected subset of TBM biome-months represents those biomes and months within which a given TBM exhibits spatiotemporal variability that explains a substantial portion of the variability observed in the observations **z** (see Eq. 3). For a given TBM biome-month distribution to be "selected" as part of the "best" model of the trend, therefore, (1) the available atmospheric observations must be sensitive to the spatiotemporal variability of fluxes within that biome-month (as represented through **H**), i.e., the information contained in atmospheric data sufficiently constrains the spatiotemporal variability within that biome-month, and (2) the variability within a particular biome-month as represented by a particular TBM must explain a sufficient portion of the variability in the atmospheric observations to offset the penalty term in Eq. (5), i.e. the reduction in RSS must outweigh the penalty term. On the contrary, if a given TBM biome-month distribution is "not selected" then, either (1) or (2) as given above is not satisfied, i.e., either that atmospheric observations are not sensitive to the NEE variability within that biome-month, or that the NEE variability as represented in the model is inconsistent with atmospheric observations. In other words, selecting or not selecting a TBM biome-month combination directly reflects on the performance of the TBM in that biome and month, as long as we have fulfilled the requirement in (1) above. If the condition in (1) is not met, we are not able to use the model selection results to examine model performance, due to the insufficient coverage of the network. We henceforth refer to the TBM biome-month combinations included in the final selected subset as the "selected" combinations or elements, or alternately as the TBM biome-month combinations "identified" using the atmospheric data.

4 Synthetic data and real data experiments

In this Section, we design a series of Synthetic Data (SD) experiments (Figure 3), in which the underlying fluxes are prescribed, to assess the sensitivity of atmospheric CO_2 measurements to NEE flux spatiotemporal patterns within all biome-month combinations, and identify when and where results from the proposed approach reliably reflect model performance in simulating NEE spatiotemporal variability. We further introduce two Real Data (RD) experiments as a proof-of-concept demonstration of our approach. In those RD experiments, we use actual atmospheric CO_2 measurements to evaluate the spatiotemporal variability of NEE as simulated by four prototypical TBMs (Section 2.3).

In the SD experiments, synthetic atmospheric observations ($\mathbf{z}$) are generated as described in Eq. (1) using fluxes ($\mathbf{s}_{\mathrm{TBM}}$) that include NEE as simulated by one of the TBMs and, in some cases, spatiotemporally-correlated flux residuals ($\boldsymbol{\xi}$) and model-data mismatch errors ($\boldsymbol{\varepsilon}$), *i.e.*, $\mathbf{z} = \mathbf{H}(\mathbf{s}_{\mathrm{TBM}} + \boldsymbol{\xi}) + \boldsymbol{\varepsilon}$. The superset of candidate ancillary variables (Figure 3, $\mathbf{X}$) includes NEE from one or more TBMs. TBMs included in $\mathbf{s}_{\mathrm{TBM}}$ and $\mathbf{X}$ are denoted as the "truth" and the "candidate (s)" respectively henceforth.

The first SD case study, SD-one-ØØ (Figure 3), is designed to investigate whether, when, and where the information contained in current atmospheric data enables the identification of the correct candidate TBM for a case where it is the only TBM considered in the model selection, where this TBM fully represents the variability in the synthetic atmospheric observations ($\boldsymbol{\xi}$=0), and where no model-data mismatch errors are included in the simulation ($\boldsymbol{\varepsilon}$=0). Given that in this case the candidate TBM explains all of the variability in the synthetic atmospheric observations, it should always be selected if the atmospheric data are sufficiently sensitive to NEE across all

biome-months; hence, biome-months for which the TBM is not selected are ones to which the atmospheric CO_2 observations are not sufficiently sensitive to offset the penalty term in Eq. (5).

The second and third SD case studies, SD-one-Ø**ε** and SD-one-**ξε** (Figure 3), are analogous to SD-one-ØØ, but include model-data mismatch errors (**ε** ≠0, denoted by **ε**) and/or spatially correlated flux residuals (**ξ** ≠0, denoted by **ξ**). These case studies are designed to test the degree to which current atmospheric observations can inform the spatiotemporal variability of NEE in cases with realistic model-data mismatch errors and where the candidate TBM only represents a portion of the true underlying NEE variability. In these case studies, noise (**ε)** is added to observations, generated as a random vector of independent normally-distributed values with variances corresponding to the diagonal elements of **R**, which are inferred from the RD-all-**ξε** experiment (described below), and a mean of 0. In addition for SD-one-**ξε**, the flux signal from the TBMs is augmented with additional spatially-correlated fluxes (**ξ)** generated as a random vector of normally distributed values with a covariance structure equal to that inferred from the RD-all-**ξε** experiment (described below). The details of the model-data mismatch errors and flux residuals are summarized in the supplementary materials.

The final SD case study, SD-all-**ξε**, builds on SD-one-**ξε** (Figure 3), but is designed to test whether the correct TBM can be identified when all four TBMs are used as candidate variables. This case study therefore explores whether current atmospheric observations can be used to differentiate among candidate TBMs. No constraints are placed on the model selection, such that more than one TBM can be selected for the same biome-month, but only the dominant TBM (i.e. the one with the largest **β**, Eq. 6) is discussed in analyzing this case.

Finally, two RD case studies, RD-one-**ξε** and RD-all-**ξε**, are defined analogously to SD-one-**ξε** and SD-all-**ξε**, to further test the applicability of our approach by examining the actual performance of the four prototypical TBMs based on available atmospheric observations. The observations (**z**) here are the actual atmospheric measurements, which by definition encompass model-data mismatch errors, and the flux residuals are also inherently present as no TBM is expected to perfectly reflect the true underlying fluxes. In each RD-one-**ξε** experiment, one of the four prototypical TBMs is used as the candidate TBM in order to assess individual TBM performance. In RD-all-**ξε**, all four TBMs are included, analogously to SD-all-**ξε**, to identify the TBM (if any) that best represents the spatiotemporal variability of NEE within a given biome-month, based on the information provided by the atmospheric measurements.

5 Sensitivity of atmospheric observations to NEE spatiotemporal variability and evaluation of the proposed approach

The SD-one-ØØ experiment examines the sensitivity of atmospheric observations to underlying flux variability and evaluates the proposed approach under idealized conditions where the true flux field is perfectly represented by the candidate TBM model, and where no model-data mismatch errors are included in the synthetic atmospheric observations.

Results indicate that the candidate TBM is selected for over 90% of all biome-months (Figure 4, top row), demonstrating that atmospheric observations are sensitive to NEE spatiotemporal variability, and that the proposed approach leverages this sensitivity to correctly identify the TBM model as being representative of the flux variability within the vast majority of biomes and months. The only notable exception is for the Tundra biome for which, other than during the height of the growing season, the atmospheric data do not provide a sufficient constraint on the flux variability, due to the poor data coverage and the weak biospheric signal. Because this biome is expected to play an important role in future global carbon cycle and climate (Belshe et

al., 2013; Ping et al., 2008; Schuur et al., 2009; Tarnocai et al., 2009) and large uncertainties remain in quantifying its role in carbon cycling (McGuire et al., 2012), this result highlights the need for strategic placement of additional CO_2 monitoring stations in the vicinity of this biome to constrain its carbon flux distribution.

The SD-one-Ø**ε** and SD-one-**ξε** case studies examine the degree to which the presence of model-data mismatch errors and additional flux variability not represented by the candidate TBM limit the information content of available observations, and the ability of the proposed approach to identify the consistency between the true underlying NEE patterns and those simulated by TBMs.

Results of SD-one-Ø**ε** show that including realistic model-data mismatch errors decreases the information content of atmospheric observations to the point where a TBM that in reality represents the full spatiotemporal flux variability is not selected for many months and TBMs within the Tropical and Subtropical biome, as well as the Desert and Xeric Shrublands biome, in addition to the Tundra biome that was not well constrained even under idealized conditions (Figure 4, middle row). The identification of a TBM as correctly representing the flux patterns also becomes more challenging during winter and spring within the Boreal Forests and Taiga biome, and the Temperate Coniferous Forests biome (Figure 4, middle row), especially when VEGAS2 is used as the true flux distribution. This result is related to the fact that the magnitude and the spatiotemporal variability of NEE simulated by VEGAS2 within those biome-months are much smaller than for other TBMs. For example, the standard deviation of NEE simulated by VEGAS2 is less than a half of that of other TBMs. Overall, the inclusion of realistic model-data mismatch, combined with the coverage of the monitoring network, make the identification of TBMs that represent the spatiotemporal variability of fluxes within biomes unreliable for three of

the seven biomes considered here, namely the Tundra, Tropical and Subtropical, and Desert and Xeric Shrublands biomes. Subsequent analyses therefore focus on the remaining four better-constrained biomes, namely the (i) Boreal Forests and Taiga, (ii) Temperate Coniferous Forests, (iii) Temperate Grasslands, Savannas, and Shrublands, and (iv) Temperate Broadleaf and Mixed Forests biomes.

SD-one-$\boldsymbol{\xi\varepsilon}$ is the most realistic single-TBM synthetic data experiment, as it includes not only model-data mismatch errors, but also variability in the spatiotemporal flux distribution that is not represented by the candidate TBM. Results for the better-constrained biomes indicate that the ability to identify a model as correctly representing a portion of the true flux variability deteriorates in the winter months for the Boreal Forests and Taiga, but remains largely unchanged in the other biomes (Figure 4, bottom row). For the winter in the Boreal Forests and Taiga biome, the TBM is only identified when the fluxes are based on SiB3, likely because this TBM has a stronger flux signal in this biome during the winter relative to the other TBMs, thereby overcoming the confounding impacts of model-data mismatch errors and additional flux variability unexplained by the TBM.

Results of SD-one-$\boldsymbol{\xi\varepsilon}$ indicate that under realistic conditions, the proposed approach is able to correctly identify a TBM that represents a portion of the true underlying flux variability within four of the seven biomes considered here, given the monitoring network used here. The magnitude of the model data mismatch used here was derived from the real-data experiments (RD-one-$\boldsymbol{\xi\varepsilon}$), and includes the impact of errors in the transport model, boundary conditions, fossil fuel emissions, and fire emissions, as well as measurement and aggregation errors. Therefore, results suggest that conclusions over the four considered biomes are robust in spite of the influences of those uncertainties. We acknowledge that the errors applied do not fully address

the complexity of uncertainties in the real world, as we assume that errors are independent and follow a Gaussian distribution. However, the results presented here, together with evidence from the literature (*e.g.*, Gourdji et al., 2012; Pillai et al., 2012), support the ability to infer flux patterns despite the many sources of uncertainty in regional inversions.

The final SD case, SD-all-**ξε**, is designed to explore whether atmospheric observations can be used to differentiate among several competing TBMs to identify the TBM that best represents the underlying flux variability. Results indicate that across the majority of the examined biomes, months, and TBMs, the proposed approach combined with the available atmospheric data are able to discriminate among models for a similar fraction of TBM-biome-month combination (Figure 5) as when only the "correct" TBM was offered as a candidate model (SD-one-**ξε**, Figure 4, bottom row). One noticeable difference, however, occurs during the growing season in the Boreal Forests and Taiga when VEGAS2 or CASA-GFED is used to represent a substantial portion of the true flux variability. In these cases, the other of these two models is often identified in the model selection procedure. This is not surprising, because these two models yield fluxes that are highly spatiotemporally correlated to one another (Figure 6), and because biospheric signals simulated by VEGAS2 are particularly weak (Huntzinger et al., 2011). Overall, therefore, for the four better-constrained biomes, the information content of the atmospheric data is sufficient to identify a TBM that represents a substantial portion of the true underlying variability using the proposed approach, even when multiple competing TBMs are available. In other words, atmospheric observations can be used to differentiate among competing TBMs. The exception, not surprisingly, is when the competing TBMs have fluxes that are highly correlated ($R>0.8$), which, for the four TBMs examined here, occurs most often over the Boreal Forests and Taiga and Temperate Coniferous Forests biomes (where biospheric

signals are relative weak and atmospheric data are less sensitive), for the VEGAS2 and CASA-GFED as well as SiB3 and ORCHIDEE model pairs (Figure 6).

6 Demonstration of the proposed approach using atmospheric observations

The results presented in Section 5 confirm that, given the coverage of atmospheric data available in 2008, the proposed approach is able to identify TBMs representing a substantial portion of the underlying NEE spatiotemporal variability over four better-constrained biomes of North America throughout most of the year. In this Section, by focusing on the RD experiment results, we demonstrate the application of the proposed approach using "real" data, by evaluating four prototypical TBMs participating in the NACP RIS.

6.1 Performance of TBMs in simulating the spatiotemporal variability of NEE

The RD-one-**ξε** case study includes four experiments, each evaluating one prototypical TBM. As a general indication of individual TBM performance across biomes and months, we sum the number of candidate TBMs selected across the four RD-one-**ξε** cases (Figure 7). We find that the capability of TBMs to simulate the NEE spatiotemporal variability varies strongly across biomes and seasons. TBMs are most frequently identified over the Temperate Broadleaf and Mixed Forests biome (7 out of 12 months with at least one TBM identified), and least frequently identified over the Boreal Forests and Taiga biome (3 out of 12 months with at least one TBM identified). Across seasons, TBMs are most frequently identified during the growing season (May-Sept, 15 out of 20 biome-months with at least one TBM identified). TBMs are least frequently identified during transition seasons (Mar-Apr and Oct-Nov, with 2 out of 16 biome-months with at least one TBM identified), likely reflecting known challenges of TBMs in representing the seasonal cycle of phenology (Richardson et al., 2012; Schaefer et al., 2012; Schwalm et al., 2010). Specifically, during Oct-Nov, none of the TBMs is identified as

representing the flux spatiotemporal variability in any of the biomes, in agreement with the finding in Gourdji et al. (2012) that carbon fluxes simulated by over 70% of the NACP TBMs are outside the 95% confidence intervals of atmospheric inversion estimates in October.

Of all 48 biome-months examined, none of the four TBMs are identified as substantially representing the spatiotemporal variability within 27 biome-months, and only one TBM is identified in 5 additional biome-months (Figure 7). Multiple TBMs are identified as representing a portion of the spatiotemporal variability within the remaining 16 biome-months (Figure 7). Interestingly, SiB3 and ORCHIDEE are selected in almost all of these 16 biome-months, suggesting that they both have the potential to explain a substantial portion of the observed variability in atmospheric CO_2. This is consistent with the similarity in NEE spatiotemporal series between SiB3 and ORCHIDEE shown in Figure 6.

The RD-all-**ξε** case study identifies the TBM that best represents the underlying flux variability (Figure 8). Out of 27 biome-months for which no individual TBM was selected in the RD-one-**ξε** experiments, 5 biome-months lead to models being selected when more than one model can be used in combination, with the dominant TBM being ORCHIDEE over the Temperate Coniferous Forests biome in Apr and May and the Temperate Broadleaf and Mixed Forests in Feb, SiB3 over the Boreal Forests and Taiga in Aug, and VEGAS2 over the Temperate Grasslands, Savannas and Shrublands in Dec.

Overall, SiB3 and ORCHIDEE are selected as the dominant TBM in explaining the flux variability as observed through the atmospheric CO_2 measurements more often than VEGAS2 and CASA-GFED (Figure 8). SiB3 appears most representative of flux patterns over boreal biomes, whereas ORCHIDEE is most representative over temperate biomes. Although SiB3

appears to be selected most often (13 biome-months), followed by ORCHIDEE (10 biome-months), none of the TBMs is consistently better than the others across all biomes and seasons.

6.2 Evaluation of the TBMs and the proposed approach within the context of earlier studies

To further evaluate the performance of, and value added provided by, the proposed approach, we assess the RD-one-$\boldsymbol{\xi\varepsilon}$ results within the context of the existing literature to determine whether (1) results are consistent with the literature wherever they are comparable, and (2) the proposed approach can provide insights that go beyond those provided by other model evaluation strategies.

Many of our findings are consistent with early work analyzing the examined TBMs within the framework of the NACP RIS. For example, we find distinctive seasonal differences in TBM performance in simulating NEE (Figures 7 and 8), consistent with the previously noted model misrepresentation of phenology seasonality based on site-level measurements (Richardson et al., 2012; Schaefer et al., 2012; Schwalm et al., 2010). In addition, we find that models perform better for Temperate Broadleaf and Mixed Forests, and that SiB3 appears to be more consistent with observations than other models, both of which are consistent with existing literature evaluating NACP RIS models (Raczka et al., 2013; Schwalm et al., 2010). The consistency between our results and existing literature further supports the performance of the proposed approach. It also implies that, although the approach proposed here is subject to many of the same uncertainties in fossil fuel emissions, fire disturbance, boundary conditions and transport models that affect all regional inversions, the main conclusions regarding TBM performance for the four major biomes examined here are quite robust.

449 The proposed approach also provides the opportunity to draw conclusions that go beyond the
450 current literature. We present two examples here.

451 First, results indicate that model capability in simulating the spatiotemporal variability (i.e.
452 patterns) of NEE varies strongly with seasons, with greater skill during the growing season than
453 during the transition seasons. In other words, even within specific biomes and months, the
454 variability of NEE is better represented during the growing season. This seasonal variability in
455 model performance may be due to seasonal differences in the dominant environmental drivers
456 controlling the spatiotemporal variability of NEE. For example, Mueller et al. (2010) found that
457 the environmental drivers controlling NEE at a hardwood forest vary across seasons, with
458 radiation, nighttime temperature and vegetative radiation indices (*i.e*., fPAR) dominating during
459 the growing, non-growing and leaf-out seasons, respectively. We hypothesize that the seasonal
460 differences in model performance is likely related to the models' ability to represent the
461 seasonally-varying influence of such environmental drivers. Because the NEE spatiotemporal
462 variability is directly related to environmental processes and drivers (Beer et al., 2010; Mueller et
463 al., 2010; Yadav et al., 2012; Gourdji et al., 2012), the proposed approach provides a close link
464 between model performance and environmental processes.

465 Second, we find that SiB3 and ORCHIDEE are identified more often as representing the
466 spatiotemporal flux variability than VEGAS2 and CASA-GFED. Given that the simulated NEE
467 spatiotemporal variability is more similar between SiB3 and ORCHIDEE, and between VEGAS2
468 and CASA-GFED, relative to across these two model pairs (Figure 6), this finding suggests that
469 aspects of the model internal structure common within the pairs likely contribute to similarities
470 in simulated flux patterns and associated performance. Such features include: 1) SiB3 and
471 ORCHIDEE use Enzyme Kinetic (EK) models while CASA-GFED2 and VEGAS use Light Use

Efficiency (LUE) models to formulate their photosynthesis processes; 2) the native model time step of SiB3 and ORCHIDEE is shorter than a day while that of CASA-GFED and VEGAS2 varies from daily to monthly; and 3) SiB3 and ORCHIDEE have substantially more plant functional types (PFTs) than CASA-GFED and VEGAS2. Although it is not possible to draw definite conclusions about the links between model structure and model performance in simulating flux patterns based on the small number of TBMs examined here and the lack of a uniform simulation protocol, a future application of this approach to a larger ensemble of models following a uniform protocol would make it possible to explore these connection in more detail.

7 Concluding remarks

In this paper, we present, evaluate and demonstrate a statistical approach based on GIM and the Bayesian Information Criterion to evaluate the spatiotemporal variability of net ecosystem exchange (NEE) as simulated by TBMs against atmospheric CO_2 concentration measurements from 35 towers in North America in 2008. We demonstrate the applicability of this approach by evaluating 4 prototypical TBMs participating in the North American Carbon Program Regional Interim Synthesis (NACP RIS).

We first design a series of synthetic data experiments in which the underlying fluxes are prescribed, to test the proposed approach and examine whether, when, and where atmospheric measurements are sensitive to, and hence can constrain, the spatiotemporal variability simulated by different TBMs. We find that due to the poor data coverage and weaker biospheric signals, current atmospheric observations cannot be used to reliably assess the flux spatiotemporal variability in the Tundra, Desert and Xeric Shrublands, and Tropical and Subtropical biomes. The remaining four biomes (i.e., Temperate Broadleaf and Mixed Forests, Temperate Grasslands, Savannas and Shrublands, Boreal Forests and Taiga, and Temperate Coniferous

Forest), however, are found to be well constrained by atmospheric data. Over these four biomes, the synthetic data experiments suggest that the proposed model selection approach, combined with the available atmospheric data, are able to identify the TBMs that represent a substantial portion of the underlying flux variability, as well as differentiate among multiple competing TBMs.

We further test and demonstrate the application of the approach by evaluating the performance of four prototypical TBMs that have been extensively assessed in literature using actual atmospheric observations. We find that conclusions about model performance are consistent with existing literature for cases where results are comparable, further supporting the applicability of our approach. Those results include that 1) TBMs represent fluxes best during the growing season (May-September) and least consistently with atmospheric observations during the transition seasons, especially in October and November; and that 2) TBMs appear to perform best over the Temperate Broadleaf and Mixed Forests biome. The experiments performed here also lead to new conclusions about the examined TBMs. For example, results show that SiB3 and ORCHIDEE appear to represent the flux variability within individual biomes and months better relative to CASA-GFED and VEGAS2. In addition, this approach has the potential to link model performance with environmental processes, making it possible to test the hypothesis that seasonal differences in TBM performance reflect models' ability to represent the seasonal variability in the dominant environmental controls on fluxes.

The comparison conducted here only included four TBMs, and was intended primarily as a demonstration of the proposed approach. Furthermore, these four TBMs were not run using a uniform experimental protocol (Huntzinger et al., 2012), precluding any conclusive results about linkages between model performance and model structure. Applying the approach presented

here to a larger ensemble of models, ideally following a uniform simulation protocol, therefore represents a logical next step.

Acknowledgments

The authors thank the biospheric modelers participating the NACP Regional Interim Synthesis, and specifically Ning Zeng, Ian Baker, Nicolas Viovy, and James Randerson who provided the model results used in the analysis presented here. We thank Deborah Huntzinger for downscaling the CASA-GFED and VEGAS2 fluxes to 3-hourly temporal resolution. Atmospheric and Environmental Research (AER), and in particular Thomas Nehrkorn, John Henderson, and Janusz Eluszkiewicz, performed the WRF-STILT simulations and provided the sensitivity footprints. We acknowledge various data providers for the continuous in-situ CO_2 measurements, as well as Sharon Gourdji and Kim Mueller for their earlier efforts in collecting and processing those datasets. This work is funded by the National Aeronautics and Space Administration (NASA) under Grant No. NNX12AB90G and No. NNX12AM97G.

Reference

Baker, I., Denning, A. S., Hanan, N., Prihodko, L., Uliasz, M., Vidale, P.-L., Davis, K., and Bakwin, P.: Simulated and observed fluxes of sensible and latent heat and CO2 at the WLEF-TV tower using SiB2.5, Global Change Biology, 9, 1262-1277, 2003.

Baker, I. T., Prihodko, L., Denning, A. S., Goulden, M., Miller, S., and Da Rocha, H. R.: Seasonal drought stress in the Amazon: Reconciling models and observations, Journal of Geophysical Research: Biogeosciences, doi: 10.1029/2007JG000644, 2008..

Balzarolo, M., Boussetta, S., Balsamo, G., Beljaars, A., Maignan, F., Calvet, J. C., Lafont, S., Barbu, A., Poulter, B., Chevallier, F., Szczypta, C., and Papale, D.: Evaluating the potential of large-scale simulations to predict carbon fluxes of terrestrial ecosystems over a European Eddy Covariance network, Biogeosciences, 11, 2661-2678, 2014.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M. A., Baldocchi, D., Bonan, G. B., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K. W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, F. I., and Papale, D.: Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate, Science, 329, 834-838, 2010.

549 Belshe, E. F., Schuur, E. a. G., and Bolker, B. M.: Tundra ecosystems observed to be CO2
550 sources due to differential amplification of the carbon cycle, Ecology Letters, 16, 1307-
551 1315, 2013.
552 Broquet, G., Chevallier, F., Bréon, F. M., Kadygrov, N., Alemanno, M., Apadula, F., Hammer,
553 S., Haszpra, L., Meinhardt, F., Morguí, J. A., Necki, J., Piacentino, S., Ramonet, M.,
554 Schmidt, M., Thompson, R. L., Vermeulen, A. T., Yver, C., and Ciais, P.: Regional
555 inversion of CO2 ecosystem fluxes from atmospheric measurements: reliability of the
556 uncertainty estimates, Atmos. Chem. Phys., 13, 9039-9056, 2013.
557 Canadell, J. G., Ciais, P., Gurney, K., Le Quéré, C., Piao, S., Raupach, M. R., and Sabine, C. L.:
558 An International Effort to Quantify Regional Carbon Fluxes, Eos, Transactions American
559 Geophysical Union, 92, 81-82, 2011.
560 Chatterjee, A., Michalak, A. M., Anderson, J. L., Mueller, K. L., and Yadav, V.: Toward reliable
561 ensemble Kalman filter estimates of CO2 fluxes, Journal of Geophysical Research:
562 Atmospheres, 117, D22306, 2012.
563 Globalview-Co2: Cooperative Atmospheric Data Integration Project – Carbon Dioxide (2010),
564 CD-ROM, NOAA ESRL, Boulder, Colorado, also available at: ftp.cmdl.noaa.gov, Path:
565 ccg/co2/GLOBALVIEW]. NOAA Global Monitoring Division: Boulder, Colorado,
566 U.S.A., 2010.
567 Göckede, M., Michalak, A. M., Vickers, D., Turner, D. P., and Law, B. E.: Atmospheric inverse
568 modeling to constrain regional-scale CO2 budgets at high spatial and temporal resolution,
569 Journal of Geophysical Research: Atmospheres, 115, D15113, 2010.
570 Gourdji, S. M., Hirsch, A. I., Mueller, K. L., Yadav, V., Andrews, A. E., and Michalak, A. M.:
571 Regional-scale geostatistical inverse modeling of North American CO2 fluxes: a
572 synthetic data study, Atmos. Chem. Phys., 10, 6151-6167, 2010.
573 Gourdji, S. M., Mueller, K. L., Yadav, V., Huntzinger, D. N., Andrews, A. E., Trudeau, M.,
574 Petron, G., Nehrkorn, T., Eluszkiewicz, J., Henderson, J., Wen, D., Lin, J., Fischer, M.,
575 Sweeney, C., and Michalak, A. M.: North American CO2 exchange: inter-comparison of
576 modeled estimates with results from a fine-scale atmospheric inversion, Biogeosciences,
577 9, 457-475, 2012.
578 Hayes, D. J., Turner, D. P., Stinson, G., Mcguire, A. D., Wei, Y., West, T. O., Heath, L. S., De
579 Jong, B., Mcconkey, B. G., Birdsey, R. A., Kurz, W. A., Jacobson, A. R., Huntzinger, D.
580 N., Pan, Y., Post, W. M., and Cook, R. B.: Reconciling estimates of the contemporary
581 North American carbon balance among terrestrial biosphere models, atmospheric
582 inversions, and a new approach for estimating net ecosystem exchange from inventory-
583 based data, Global Change Biology, 18, 1282-1299, 2012.
584 Hoeting, J. A., Davis, R. A., Merton, A. A., and Thompson, S. E.: Model Selection For
585 Geostatistical Models, Ecological Applications, 16, 87-98, 2006.
586 Huntzinger, D. N., Gourdji, S. M., Mueller, K. L., and Michalak, A. M.: The utility of
587 continuous atmospheric measurements for identifying biospheric CO2 flux variability,
588 Journal of Geophysical Research: Atmospheres, 116, D06110, 2011.
589 Huntzinger, D. N., Post, W. M., Wei, Y., Michalak, A. M., West, T. O., Jacobson, A. R., Baker,
590 I. T., Chen, J. M., Davis, K. J., Hayes, D. J., Hoffman, F. M., Jain, A. K., Liu, S.,
591 Mcguire, A. D., Neilson, R. P., Potter, C., Poulter, B., Price, D., Raczka, B. M., Tian, H.
592 Q., Thornton, P., Tomelleri, E., Viovy, N., Xiao, J., Yuan, W., Zeng, N., Zhao, M., and
593 Cook, R.: North American Carbon Program (NACP) regional interim synthesis:
594 Terrestrial biospheric model intercomparison, Ecological Modelling, 232, 144-157, 2012.

595 Keenan, T. F., Baker, I., Barr, A., Ciais, P., Davis, K., Dietze, M., Dragoni, D., Gough, C. M.,
596 Grant, R., Hollinger, D., Hufkens, K., Poulter, B., Mccaughey, H., Raczka, B., Ryu, Y.,
597 Schaefer, K., Tian, H., Verbeeck, H., Zhao, M., and Richardson, A. D.: Terrestrial
598 biosphere model performance for inter-annual variability of land-atmosphere CO2
599 exchange, Global Change Biology, 18, 1971-1987, 2012.
600 Kort, E. A., Eluszkiewicz, J., Stephens, B. B., Miller, J. B., Gerbig, C., Nehrkorn, T., Daube, B.
601 C., Kaplan, J. O., Houweling, S., and Wofsy, S. C.: Emissions of CH4 and N2O over the
602 United States and Canada based on a receptor-oriented modeling framework and
603 COBRA-NA atmospheric observations, Geophysical Research Letters, 35, L18808, 2008.
604 Krinner, G., Viovy, N., De Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais,
605 P., Sitch, S., and Prentice, I. C.: A dynamic global vegetation model for studies of the
606 coupled atmosphere-biosphere system, Global Biogeochemical Cycles, 19, GB1015,
607 2005.
608 Lin, J. C., Gerbig, C., Wofsy, S. C., Andrews, A. E., Daube, B. C., Davis, K. J., and Grainger, C.
609 A.: A near-field tool for simulating the upstream influence of atmospheric observations:
610 The Stochastic Time-Inverted Lagrangian Transport (STILT) model, Journal of
611 Geophysical Research: Atmospheres, 108, 4493, 2003.
612 Mcguire, A. D., Christensen, T. R., Hayes, D., Heroult, A., Euskirchen, E., Kimball, J. S.,
613 Koven, C., Lafleur, P., Miller, P. A., Oechel, W., Peylin, P., Williams, M., and Yi, Y.: An
614 assessment of the carbon balance of Arctic tundra: comparisons among observations,
615 process models, and atmospheric inversions, Biogeosciences, 9, 3185-3204, 2012.
616 Mckain, K., Wofsy, S. C., Nehrkorn, T., Eluszkiewicz, J., Ehleringer, J. R., and Stephens, B. B.:
617 Assessment of ground-based atmospheric observations for verification of greenhouse gas
618 emissions from an urban region, Proceedings of the National Academy of Sciences, 109,
619 8423-8428, 2012.
620 Michalak, A. M., Bruhwiler, L., and Tans, P. P.: A geostatistical approach to surface flux
621 estimation of atmospheric trace gases, Journal of Geophysical Research: Atmospheres,
622 109, D14109, 2004.
623 Miles, N. L., Richardson, S. J., Davis, K. J., Lauvaux, T., Andrews, A. E., West, T. O., Bandaru,
624 V., and Crosson, E. R.: Large amplitude spatial and temporal gradients in atmospheric
625 boundary layer CO2mole fractions detected with a tower-based network in the U.S. upper
626 Midwest, Journal of Geophysical Research: Biogeosciences, 117, G01019, 2012.
627 Mueller, K. L., Yadav, V., Curtis, P. S., Vogel, C., and Michalak, A. M.: Attributing the
628 variability of eddy-covariance CO2 flux measurements across temporal scales using
629 geostatistical regression for a mixed northern hardwood forest, Global Biogeochemical
630 Cycles, 24, 2010.
631 Ogle, S., Davis, K., Andrews, A., West, T., Cook, R., Parkin, R., Morisette, J., Verma, S., and
632 Wofsy, S.: Science Plan: Mid-Continent Intensive Campaign. Greenbelt, Md., 2006.
633 Peters, W., Jacobson, A. R., Sweeney, C., Andrews, A. E., Conway, T. J., Masarie, K., Miller, J.
634 B., Bruhwiler, L. M. P., Pétron, G., Hirsch, A. I., Worthy, D. E. J., Van Der Werf, G. R.,
635 Randerson, J. T., Wennberg, P. O., Krol, M. C., and Tans, P. P.: An atmospheric
636 perspective on North American carbon dioxide exchange: CarbonTracker, Proceedings of
637 the National Academy of Sciences, 104, 18925-18930, 2007.
638 Pillai, D., Gerbig, C., Kretschmer, R., Beck, V., Karstens, U., Neininger, B., and Heimann, M.:
639 Comparing Lagrangian and Eulerian models for CO2 transport – a step towards Bayesian
640 inverse modeling using WRF/STILT-VPRM, Atmos. Chem. Phys., 12, 8979-8991, 2012.

641 Ping, C.-L., Michaelson, G. J., Jorgenson, M. T., Kimble, J. M., Epstein, H., Romanovsky, V. E.,
642 and Walker, D. A.: High stocks of soil organic carbon in the North American Arctic
643 region, Nature Geosci, 1, 615-619, 2008.
644 Raczka, B. M., Davis, K. J., Huntzinger, D. N., Neilson, R., Poulter, B., Richardson, A., Xiao, J.,
645 Baker, I., Ciais, P., Keenan, T. F., Law, B., Post, W. M., Ricciuto, D., Schaefer, K., Tian,
646 H., Tomelleri, E., Verbeeck, H., and Viovy, N.: Evaluation of continental carbon cycle
647 simulations with North American flux tower observations, Ecological Monographs, 83,
648 531-556, 2013.
649 Richardson, A. D., Anderson, R. S., Arain, M. A., Barr, A. G., Bohrer, G., Chen, G., Chen, J. M.,
650 Ciais, P., Davis, K. J., Desai, A. R., Dietze, M. C., Dragoni, D., Garrity, S. R., Gough, C.
651 M., Grant, R., Hollinger, D. Y., Margolis, H. A., Mccaughey, H., Migliavacca, M.,
652 Monson, R. K., Munger, J. W., Poulter, B., Raczka, B. M., Ricciuto, D. M., Sahoo, A. K.,
653 Schaefer, K., Tian, H., Vargas, R., Verbeeck, H., Xiao, J., and Xue, Y.: Terrestrial
654 biosphere models need better representation of vegetation phenology: results from the
655 North American Carbon Program Site Synthesis, Global Change Biology, 18, 566-584,
656 2012.
657 Sasai, T., Ichii, K., Yamaguchi, Y., and Nemani, R.: Simulating terrestrial carbon fluxes using
658 the new biosphere model "biosphere model integrating eco-physiological and
659 mechanistic approaches using satellite data" (BEAMS), Journal of Geophysical Research:
660 Biogeosciences, 110, G02014, 2005.
661 Schaefer, K., Schwalm, C. R., Williams, C., Arain, M. A., Barr, A., Chen, J. M., Davis, K. J.,
662 Dimitrov, D., Hilton, T. W., Hollinger, D. Y., Humphreys, E., Poulter, B., Raczka, B. M.,
663 Richardson, A. D., Sahoo, A., Thornton, P., Vargas, R., Verbeeck, H., Anderson, R.,
664 Baker, I., Black, T. A., Bolstad, P., Chen, J., Curtis, P. S., Desai, A. R., Dietze, M.,
665 Dragoni, D., Gough, C., Grant, R. F., Gu, L., Jain, A., Kucharik, C., Law, B., Liu, S.,
666 Lokipitiya, E., Margolis, H. A., Matamala, R., Mccaughey, J. H., Monson, R., Munger, J.
667 W., Oechel, W., Peng, C., Price, D. T., Ricciuto, D., Riley, W. J., Roulet, N., Tian, H.,
668 Tonitto, C., Torn, M., Weng, E., and Zhou, X.: A model-data comparison of gross
669 primary productivity: Results from the North American Carbon Program site synthesis,
670 Journal of Geophysical Research: Biogeosciences, 117, G03010, 2012.
671 Schuur, E. a. G., Vogel, J. G., Crummer, K. G., Lee, H., Sickman, J. O., and Osterkamp, T. E.:
672 The effect of permafrost thaw on old carbon release and net carbon exchange from
673 tundra, Nature, 459, 556-559, 2009.
674 Schwalm, C. R., Williams, C. A., Schaefer, K., Anderson, R., Arain, M. A., Baker, I., Barr, A.,
675 Black, T. A., Chen, G., Chen, J. M., Ciais, P., Davis, K. J., Desai, A., Dietze, M.,
676 Dragoni, D., Fischer, M. L., Flanagan, L. B., Grant, R., Gu, L., Hollinger, D., Izaurralde,
677 R. C., Kucharik, C., Lafleur, P., Law, B. E., Li, L., Li, Z., Liu, S., Lokupitiya, E., Luo,
678 Y., Ma, S., Margolis, H., Matamala, R., Mccaughey, H., Monson, R. K., Oechel, W. C.,
679 Peng, C., Poulter, B., Price, D. T., Riciutto, D. M., Riley, W., Sahoo, A. K., Sprintsin, M.,
680 Sun, J., Tian, H., Tonitto, C., Verbeeck, H., and Verma, S. B.: A model-data
681 intercomparison of CO2 exchange across North America: Results from the North
682 American Carbon Program site synthesis, Journal of Geophysical Research:
683 Biogeosciences, 115, 566-584, 2010.
684 Shiga, Y. P., Michalak, A. M., Gourdji, S. M., Mueller, K. L., and Yadav, V.: Detecting fossil
685 fuel emissions patterns from subcontinental regions using North American in situ CO2
686 measurements, Geophysical Research Letters, 41, 2014GL059684, 2014.

Skamarock, W. C. and Klemp, J. B.: A time-split nonhydrostatic atmospheric model for weather research and forecasting applications, J. Comput. Phys., 227, 3465-3485, 2008.

Tarnocai, C., Canadell, J. G., Schuur, E. a. G., Kuhry, P., Mazhitova, G., and Zimov, S.: Soil organic carbon pools in the northern circumpolar permafrost region, Global Biogeochemical Cycles, 23, GB2023, 2009.

Turner, D. P., Göckede, M., Law, B. E., Ritts, W. D., Cohen, W. B., Yang, Z., Hudiburg, T., Kennedy, R., and Duane, M.: Multiple constraint analysis of regional land–surface carbon flux, Tellus B, 63, 207-221, 2011.

Van Der Werf, G. R., Randerson, J. T., Giglio, L., Collatz, G. J., Kasibhatla, P. S., and Arellano Jr, A. F.: Interannual variability in global biomass burning emissions from 1997 to 2004, Atmos. Chem. Phys., 6, 3423-3441, 2006.

Yadav, V., Mueller, K. L., Dragoni, D., and Michalak, A. M.: A geostatistical synthesis study of factors affecting gross primary productivity in various ecosystems of North America, Biogeosciences, 7, 2655-2671, 2010.

Yadav, V., Mueller, K. L., and Michalak, A. M.: A backward elimination discrete optimization algorithm for model selection in spatio-temporal regression models, Environmental Modelling & Software, 42, 88-98, 2013.

Zeng, N., Mariotti, A., and Wetzel, P.: Terrestrial mechanisms of interannual CO2 variability, Global Biogeochemical Cycles, 19, GB1016, 2005.

Figures

Figure 1. North American biomes, modified from Olson (2001), as defined for the case studies; green triangles indicate the locations of atmospheric CO_2 measurement towers used in the analysis.

Figure 2. Illustration of the 1°×1° and 3-hourly spatiotemporal variability of NEE simulated by CASA-GFED for Boreal Forests and Taiga in July. A vector including these 1°×1° and 3-hourly fluxes corresponds to one ancillary variable (i.e. one column) in **X**)

Figure 3. Illustration of Synthetic Data (SD) case studies as described in Section 4.

Figure 4. Average numbers of months within each season for which the candidate TBM is selected for the SD-one-ØØ, SD-one-Ø**ε** and SD-one-**ξε** case studies (Figure 3). Grey shading in SD-one-**ξε** represents biomes that were determined not to be well constrained by available atmospheric data. DJF: December, January, February; MAM: March, April, May; JJA: June, July, August; SON: September, October, November. The criteria for grey areas includes: 1) no models are selected in one season; or 2) the overall model selection is less than 50% in a year.

Figure 5. Average numbers of months within each season for which the candidate TBM is selected for the SD-all-**ξε** case study (Figure 3). Grey shading represents biomes that were determined not to be well constrained by available atmospheric data. DJF: December, January, February; MAM: March, April, May; JJA: June, July, August; SON: September, October, November.

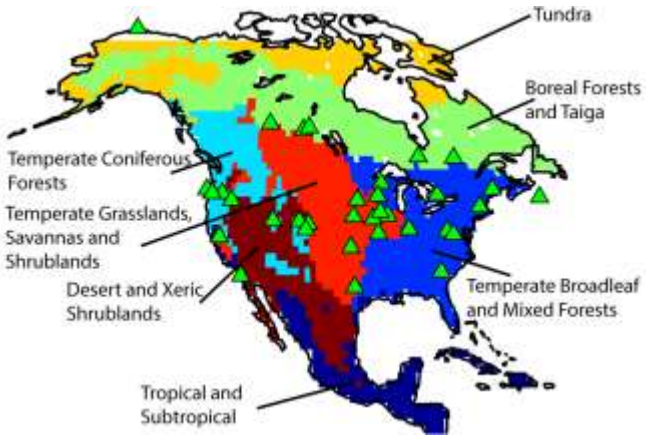
Figure 6. The correlation coefficient of NEE spatiotemporal series as simulated by different TBMs throughout 2008 for the four biomes better constrained by available atmospheric observations. TGSS: Temperate Grasslands, Savannas, Shrublands; Bore: Boreal Forests and Taiga; TCoF: Temperate Coniferous Forests; TBMF: Temperate Boradleaf and Mixed Forests.

Figure 7. Number of TBMs that are selected for each biome-month in the RD-one-**ξε** cases study. Grey shading represents biomes that were determined not to be well constrained by available atmospheric data.

Figure 8. The TBM that explains the most variability in atmospheric measurements for a given biome-month, as identified by the RD-all-**ξε** experiment. Grey shading represents biomes that were determined not to be well constrained by available atmospheric data.

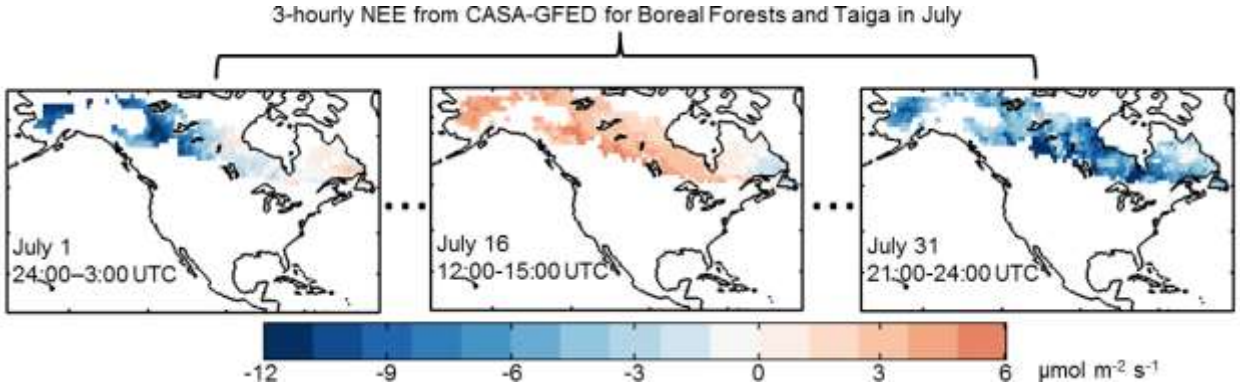
738 Figures

739 Figure 1



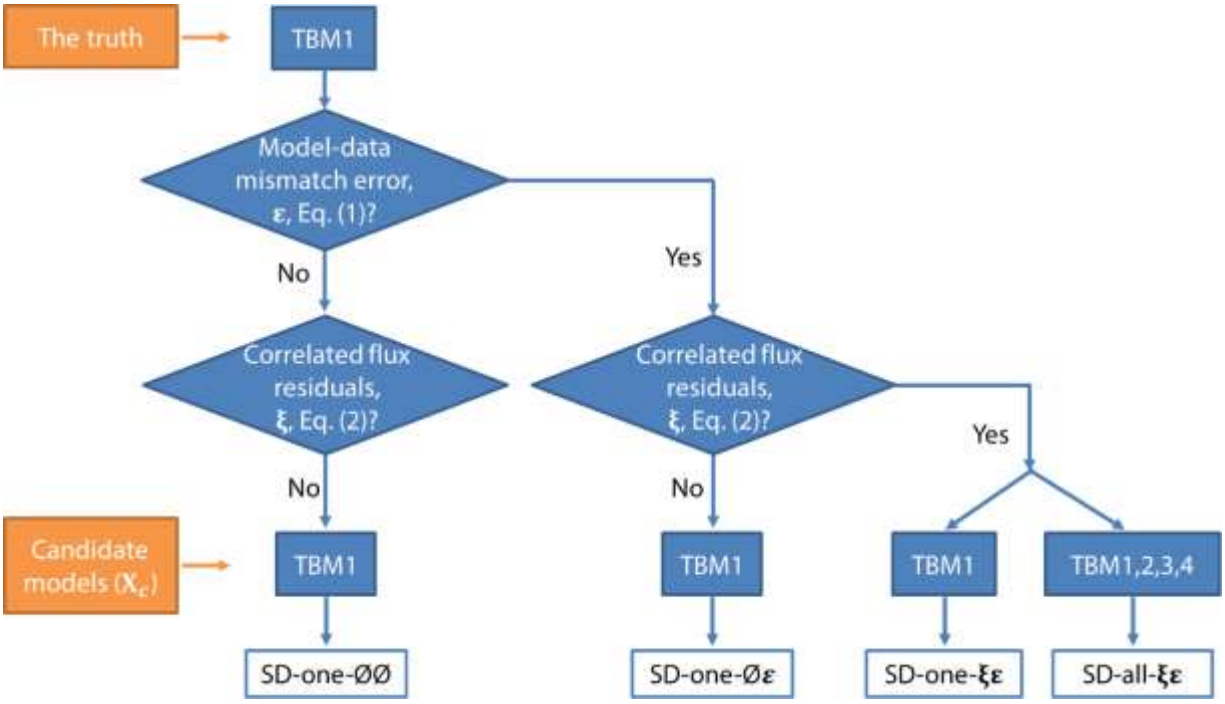
740

741 Figure 2



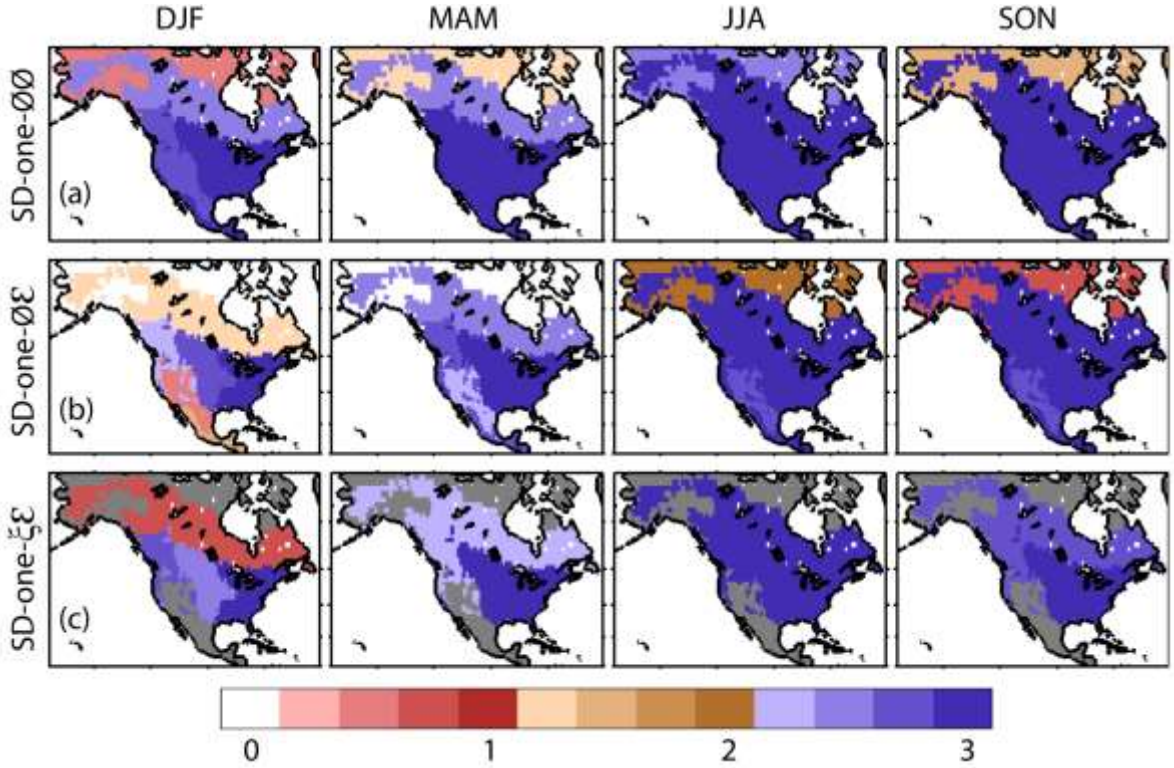
742

743 Figure 3



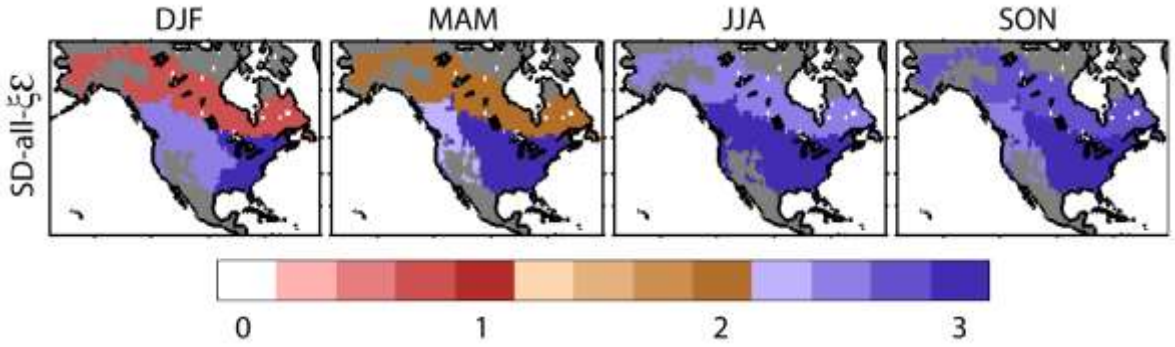
744

745 Figure 4



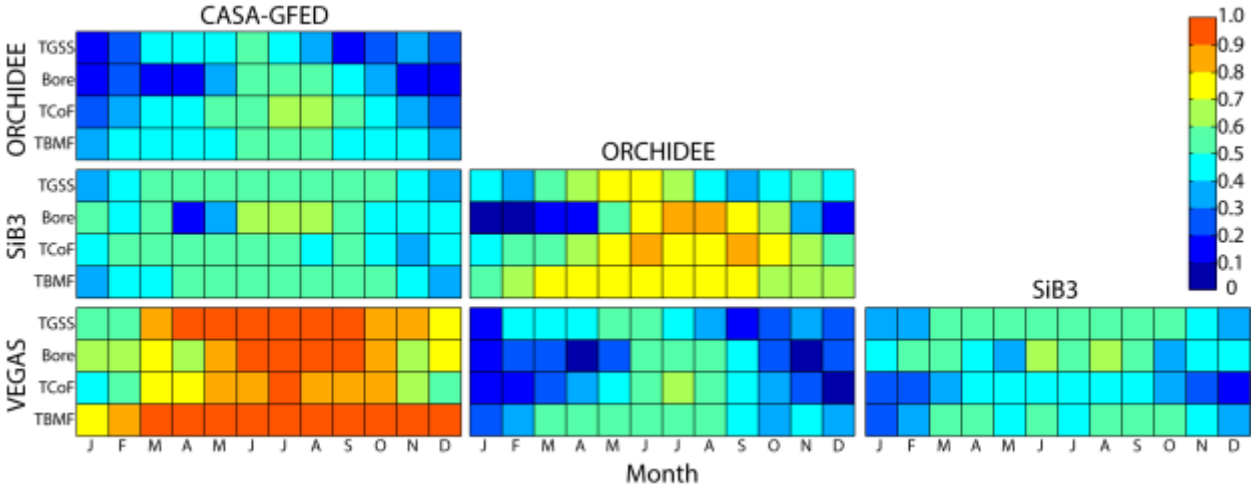
746

747 Figure 5



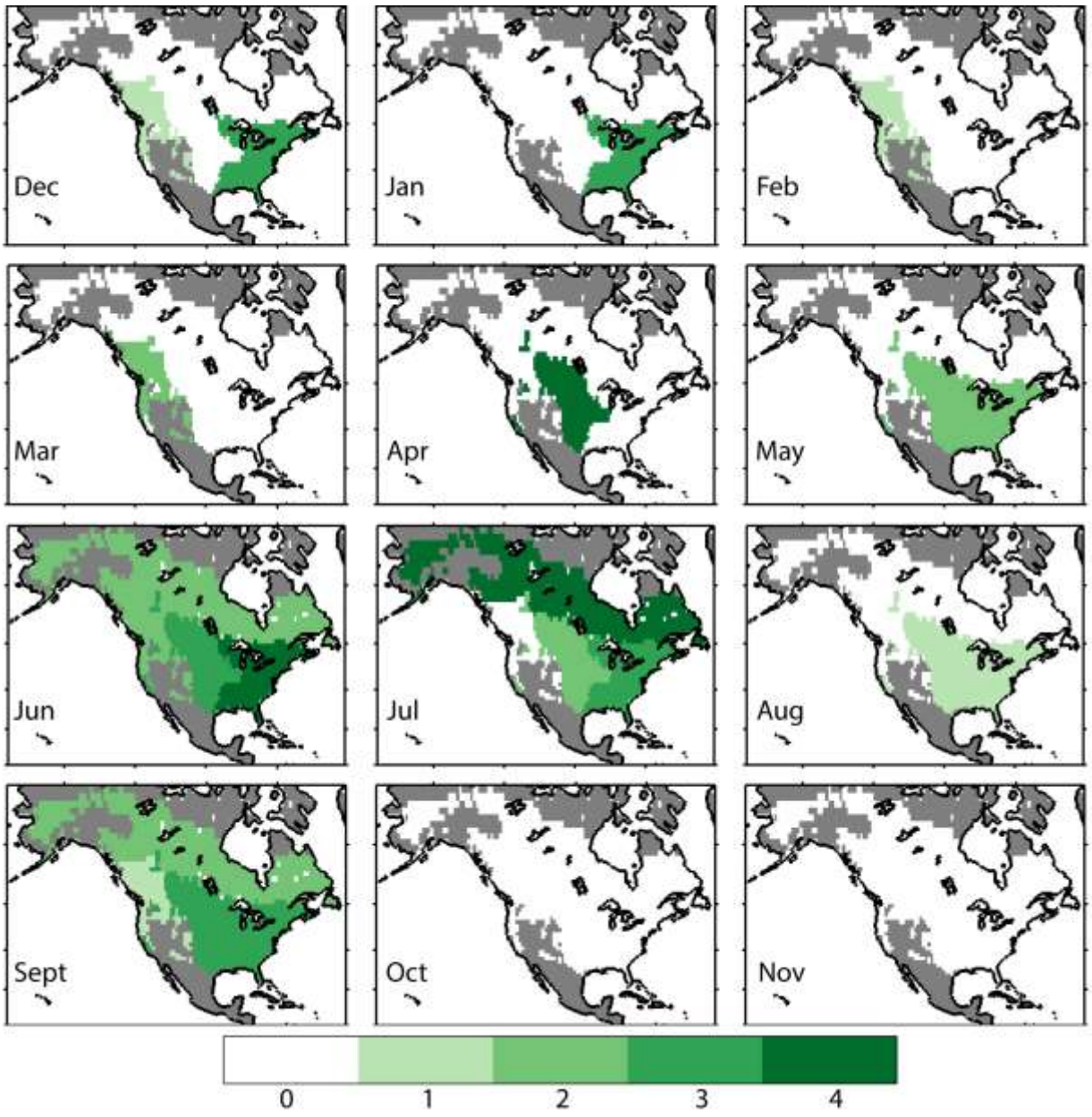
748

749 Figure 6



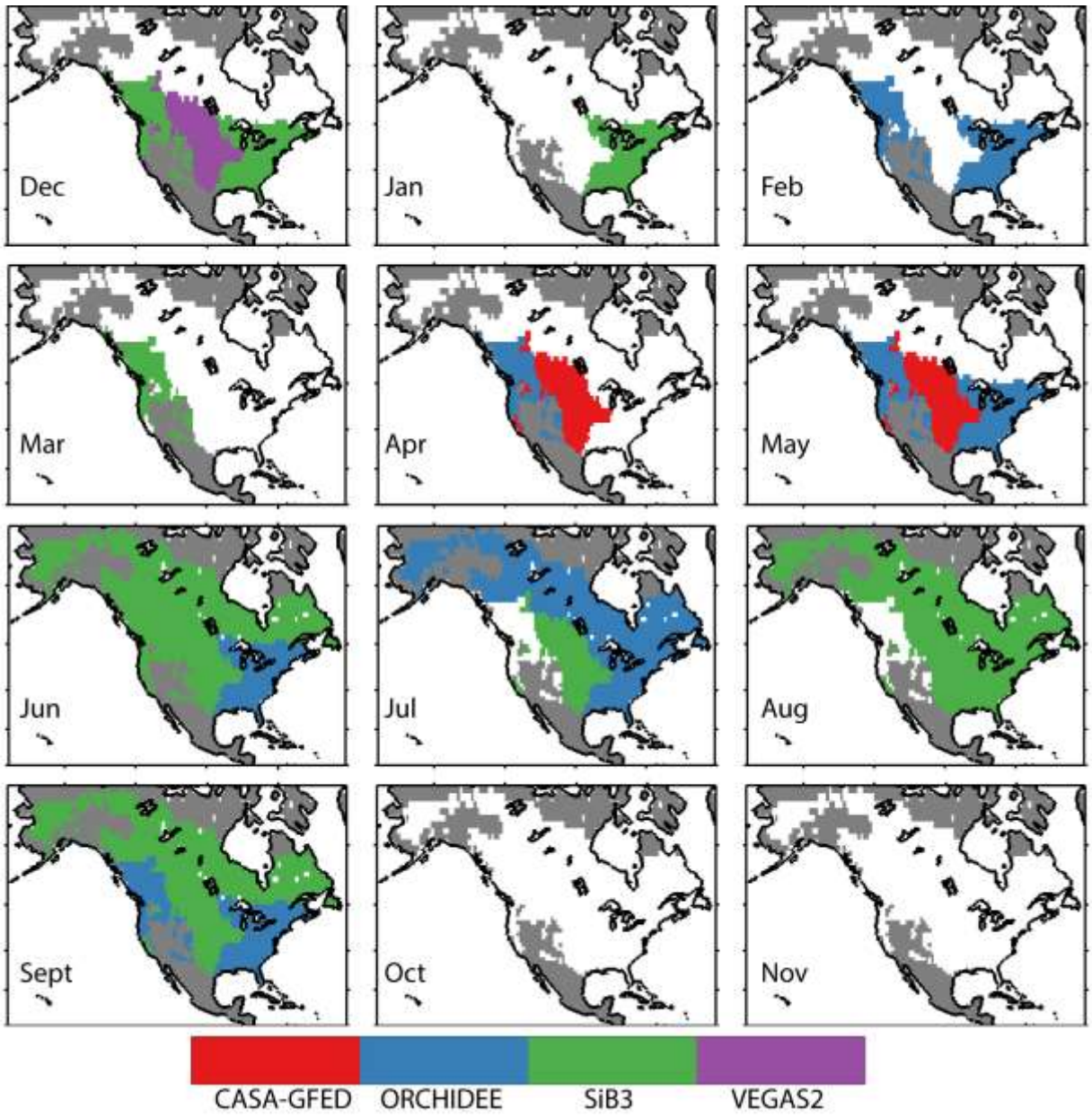
750

751 Figure 7



752

753 Figure 8



754