Dear Editor:

In order to avoid a lengthy and repetitive response letter, please find our point-by-point responses in the authors' comments we uploaded to BGD discussion forum.

We have, however, included a manuscript with changes tracked to show all relevant changes we made to improve the manuscript.

We thank you for your time and look forward to hear from you.

Yuanyuan on behalf of all co-authors

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 ~~propose a statistical model selection~~ 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. ~~We~~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 ~~the~~ North ~~American continent~~America, and that atmospheric observations can therefore be used to evaluate simulated spatiotemporal flux patterns~~, rather than focusing solely on flux magnitudes at aggregated scales. Results show that the proposed approach can be used to assess whether a TBM represents a substantial portion of the underlying flux variability~~ as well as to differentiate among multiple competing TBMs. ~~When applying the proposed approach to~~ Experiments using real atmospheric observations and four prototypical TBMs~~, we find~~ 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 ~~seasonal difference in the~~ result is consistent with previous work showing that the ability of TBMs to ~~represent the spatiotemporal flux variability may reflect the models' capability to represent the~~model flux magnitudes is also seasonally-~~varying influence of environmental drivers of flux. While none of the TBMs consistently outperforms the others, differences among the examined models are at least partially~~

~~attributable to their internal structures.~~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 ~~the~~ 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., 2013; Keenan et al., 2012; Raczka et al., 2013; Richardson et al., 2012; Sasai et al., 2005; Schaefer et al., 2012; Schwalm et al., 2010)~~(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 ~~simulations~~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 ~~estimates~~ 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)~~(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

Formatted: Font color: Text 1

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)~~(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 the fact that processes controlling carbon-climate feedbacks are manifested differently across TBMs.

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

The majority of previous studies examining carbon flux variability are still based on spatially and/or temporally aggregated carbon fluxes, however. An evaluation of flux variability, or flux patterns, at the fine native spatiotemporal scales of TBM simulations would make it possible to more directly target the fine-scale spatiotemporal patterns of carbon fluxes that have been shown to directly relate to environmental/climatic factors, such as precipitation, radiation and nighttime temperature ~~(Beer et al., 2010; Mueller et al., 2010; Yadav et al., 2010)~~(Beer et al., 2010;

Mueller et al., 2010; Yadav et al., 2010). Such evaluations could therefore inform model improvements at the process level.

Observations of atmospheric CO_2 can potentially be used to assess such fine-scale spatiotemporal flux patterns. ~~Variations~~On one hand, atmospheric CO_2 observations are sensitive to fine-scale NEE spatial and temporal variability (Huntzinger et al., 2011). On the other hand, variations in atmospheric CO_2 measurements are routinely used in inverse modeling frameworks 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~~ to global~~ domains ~~(Broquet et al., 2013; Göckede et al., 2010; Gourdji et al., 2010; Gourdji et al., 2012)~~(Broquet et al., 2013; Göckede et al., 2010; Gourdji et al., 2010; Gourdji et al., 2012). ~~Those~~Despite the uncertainties existing in regional inversions due to uncertainties in atmospheric transport, fossil fuel emissions, fire disturbance, and boundary conditions, these studies ~~demonstrate~~do point to the possibility of ~~comparing~~evaluating the spatiotemporal patterns of fluxes from biospheric models ~~to those~~through the use of ~~high-resolution~~ inverse models.

~~In~~With this ~~paper, we propose a statistical model selection approach for using~~goal in mind, what is needed is an atmospheric-inversion-based method that can use variations in atmospheric CO_2 ~~measurements~~ to ~~evaluate~~assess the spatiotemporal ~~variability of~~ 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°×1° and 3-hourly spatiotemporal variability of Net Ecosystem Exchange (NEE) ~~at relatively fine scales (1°×1° and 3-hourly resolution), in order to target scales at which the link between environmental drivers and~~ simulated ~~fluxes can inform TBM improvements.~~ by TBMs using atmospheric CO_2

measurements. This fine scale variability is evaluated here ~~across seasons (monthly)~~within each month and ~~biomes~~biome over North ~~American~~America, thus providing ~~an approach for evaluating~~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 ~~potentially~~ providing a closer link to process-based understanding of TBM performance. The approach is first evaluated ~~using~~with a series of synthetic data experiments~~, followed by an~~ where the underlying flux patterns affecting the atmospheric CO_2 signals are known. The application ~~to~~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)~~(Huntzinger et al., 2012).

~~The remainder of the paper is organized as follows. We describe the data used in the case studies in Section 2. The proposed statistical model selection approach is introduced in Section 3. The experimental case studies are listed in Section 4. In Section 5, we evaluate the feasibility of the proposed approach within the context of the information content of available atmospheric observations using synthetic data experiments. In Section 6, we present the prototypical application to evaluate four TBMs participating in the NACP RIS activities. Final conclusions are presented in Section 7.~~

~~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 (Shiga et al.) to evaluate the simulated NEE spatiotemporal variability over North American land. 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 as in Gourdji et al. (2012) and are sub-selected as in Shiga et al. (submitted). To remove the effect of boundary conditions, we pre-subtract the GLOBALVIEW-CO2 boundary condition from atmospheric measurements as in Gourdji et al. (2012). This earlier study suggested that GLOBALVIEW-CO2 gives more realistic estimate of CO_2 boundary conditions for North America relative to boundary conditions taken from CarbonTracker. 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., submitted).~~

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 ~~(Section 2.1)~~ 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)~~(Lin et al., 2003), driven by meteorological fields from the Weather Research and Forecast (WRF) model ~~(Skamarock and Klemp, 2008)~~(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., 2011b; Kort et al., 2008; McKain et al., 2012)~~(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 ~~can~~are also~~ be~~ used to generate synthetic observational time series based on TBM flux simulations.

2.3 Terrestrial Biospheric Models (TBMs)

We ~~evaluate~~use simulations from four TBMs to evaluate the proposed approach, namely CASA-GFED ~~(van der Werf et al., 2006)~~(van der Werf et al., 2006)~~, SiB3 (Baker et al., 2008), ORCHIDEE ~~, SiB3 (Baker et al., 2008), ORCHIDEE ~~(Krinner et al., 2005)~~(Krinner et al., 2005) and VEGAS2 ~~(Zeng et al., 2005)~~(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. (2011b).~~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)~~(2012).

~~3 Model selection based on Bayesian Information Criterion (BIC)~~

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 ~~do so~~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 ~~for~~within specific biome-month combinations. ~~The examined North American biomes are shown in~~ Figure ~~1~~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:

$$\mathbf{z} = \mathbf{H}\mathbf{s} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{z}$ is an $n \times 1$ vector of atmospheric CO_2 observations, $\mathbf{s}$ is an $m \times 1$ vector of the ~~“true”~~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 $(\mathbf{z})$ and modeled $(\mathbf{Hs})$ 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:

$$\mathbf{s} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} \tag{2}$$

where $\mathbf{X}$ is a $m \times p$ -matrix with each column representing NEE 1°×1° 3-hourly spatiotemporal variability ~~in~~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 $\mathbf{X}$ are thus predictor variables for the dependent variable $\mathbf{s}$. 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 $\mathbf{X}$ and $\mathbf{s}$, 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 $\mathbf{s}$ that cannot be explained by the predictor variables in $\mathbf{X}$, 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 $\mathbf{z}$ in terms of the NEE predictor variables $\mathbf{X}$:

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

where $\mathbf{z}$ 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 ~~optimally represent~~capture a substantial portion of the CO_2 variability ~~as~~ observed in $\mathbf{z}$. This constitutes a classical statistical model selection problem, in which we examine which predictor variables (candidate columns in $\mathbf{X}$) are useful in explaining the atmospheric CO_2 measurements ($\mathbf{z}$).

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)~~(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)~~(Gourdji et al., 2012), making it ideal for the application presented here. The standard expression for BIC is:

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

where RSS represents the residual sums of squares of a given candidate model $\mathbf{X}_c$, $\boldsymbol{\Sigma}$ 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)~~Gourdji et al. (2012):

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

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

The specific covariance parameters needed to define $\mathbf{Q}$ and $\mathbf{R}$, which are ~~themselves~~ needed to define $\boldsymbol{\Sigma}$, vary between experiments and are obtained as described in the supplementary materials.

~~The~~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 ~~that~~one with the minimum BIC value, ~~thus~~ 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)~~(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 ~~2~~3), in which the underlying fluxes are prescribed, to ~~test the proposed approach and~~ assess the ~~degree to which~~

~~current~~sensitivity of atmospheric ~~observations are sensitive to, and informative of, the~~CO_2 measurements to NEE flux spatiotemporal ~~variability of NEE.~~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 ~~the~~ four prototypical TBMs ~~described in~~ (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 ~~2~~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 ~~2~~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 ~~2~~3), are analogous to SD-one-ØØ, but include model-data mismatch errors ($\boldsymbol{\varepsilon} \neq 0$, denoted by **ε**) and/or spatially correlated flux residuals ($\boldsymbol{\xi} \neq 0$, 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. ~~The details of the model-data mismatch errors and flux residuals are described~~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 ~~2~~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 ~~examine~~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 ~~flux~~ 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 ~~3~~4, top row), demonstrating that atmospheric observations are sensitive to NEE spatiotemporal variability, and that the proposed ~~model selection~~ 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 ~~plays~~is expected to play an important role in ~~the~~future global carbon cycle and climate ~~(Belshe et al., 2013; Ping et al., 2008; Schuur et al., 2009; Tarnocai et al., 2009)~~(Belshe et al., 2013; Ping et al., 2008; Schuur et al., 2009; Tarnocai et al., 2009) and large uncertainties remain in quantifying its role ~~and evolution with time~~in carbon cycling ~~(McGuire et~~

~~al., 2012)~~(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 ~~a portion of~~additional flux variability not represented by the candidate TBM limit the information content of available observations, and the ability of the proposed approach to ~~observe the spatiotemporal flux variability under more realistic conditions~~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 ~~month-TBM combinations in~~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 ~~3~~4, middle row). The identification of a TBM as correctly representing the flux ~~variability~~patterns also becomes more challenging during winter and spring ~~over~~within the Boreal Forests and Taiga biome, and the Temperate Coniferous Forests biome (Figure ~~3~~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 ~~over~~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 ~~over~~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-**ξε** is ~~designed as~~ 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 ~~3~~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. ~~Overall, however, results~~

Results of SD-one-**ξε** indicate that~~,~~ under realistic conditions, the proposed approach ~~combined with the available atmospheric observations are~~ is able to correctly identify a TBM that ~~correctly~~ represents a portion of the true underlying flux variability ~~for much of the year over~~ 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- **ξε**), 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 ~~4~~5) as when only the "correct" TBM was offered as a candidate model (SD-one-**ξε**, Figure ~~3~~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 ~~5~~6), and because biospheric signals simulated by VEGAS2 are particularly weak ~~(Huntzinger et al., 2011b).~~(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 ~~5~~6).

~~6 Evaluation~~Demonstration of ~~NACP RIS simulations in representing NEE spatiotemporal variability~~

6 ~~Section 5 confirms that~~ the proposed ~~model selection~~ approach~~, combined with available CO_2 measurements,~~ 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 ~~4~~four better-constrained biomes of North America ~~largely~~ throughout ~~a~~most of the year. In this Section, by focusing on the RD experiment results, we ~~examine the performance of~~ 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~~, as assessed using atmospheric CO_2 measurements~~

The RD-one- **ξε** case study includes ~~4~~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 ~~6). Overall, we~~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)~~(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)~~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 ~~in~~within 27 biome-months, and only one TBM is identified in 5 additional biome-months (Figure ~~6~~7). Multiple TBMs are identified as representing a portion of the spatiotemporal variability ~~in~~within the remaining 16 biome-months (Figure ~~6~~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 ~~5~~6.

The RD-all- $\boldsymbol{\xi\varepsilon}$ case study identifies the TBM that best represents the underlying flux variability (Figure ~~7). Over the 16 biome-months for which more than one TBM was selected in the RD-one- $\boldsymbol{\xi\varepsilon}$ series of experiments, SiB3 is identified as the dominant model explaining the observed atmospheric CO_2 variability for 10 biome-months, ORCHIDEE is identified over 4 biome-months, while CASA-GFED2 is identified over the remaining 2 biome-months (Figure 7). For the 5 biome-months with only one TBM selected in the RD-one- $\boldsymbol{\xi\varepsilon}$ series of experiments, the selected TBM is consistent in the RD-all- $\boldsymbol{\xi\varepsilon}$ experiment, with 3 biome-months best represented~~

~~by SiB3 and 2 others by ORCHIDEE. Furthermore, out~~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 ~~7~~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 ~~Discussion~~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- **ξε** 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

~~We~~ within the framework of the NACP RIS. For example, we find distinctive seasonal differences in TBM performance in simulating ~~spatiotemporal flux variability (~~NEE (Figures ~~6 and~~ 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)~~(Richardson et al., 2012; Schaefer et al., 2012; Schwalm et al., 2010)~~.~~ ~~These model limitations~~ . 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.

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

First, results indicate that model capability in simulating the spatiotemporal variability (i.e. patterns) of NEE varies strongly with seasons, with greater skill during the growing season than during the transition seasons. In other words, even within specific biomes and months, the variability of NEE is better represented during the growing season. This seasonal variability in model performance may be due to seasonal differences in the dominant environmental drivers controlling the ~~NEE~~ spatiotemporal variability of NEE. For example, ~~Mueller et al. (2010)~~Mueller et al. (2010) found that ~~site-level NEE measurements~~the environmental drivers controlling NEE at ~~one location were best explained by~~a hardwood forest vary across seasons, with radiation, nighttime temperature and vegetative radiation indices (*i.e.*, fPAR) dominating during the growing, non-growing and leaf-out seasons, respectively. We hypothesize that the seasonal ~~difference~~differences in model performance is likely related to the models' ability to represent the seasonally-varying influence of such environmental drivers. ~~As our results are based on~~Because the NEE spatiotemporal variability~~, which has been showed to be~~ is directly related to environmental processes and drivers (Beer et al., 2010; Mueller et al., 2010; Yadav et al., 2012; Gourdji et al., 2012), ~~our work allows a potential further~~ the proposed approach provides a close link between model performance and environmental processes~~ to test this hypothesis.~~.

~~We also~~Second, we find that SiB3 and ORCHIDEE are identified more often as representing the spatiotemporal flux variability than VEGAS2 and CASA-GFED. ~~Overall, SiB3 and ORCHIDEE can both explain a substantial portion of the observed variability over almost all of the 16 biome-months for which multiple TBMs are selected in the RD-one-**ξε** experiments (Figure 6).~~ Given that the simulated NEE spatiotemporal variability is more similar between SiB3 and ORCHIDEE, and between VEGAS2 and CASA-GFED, relative to across these two model pairs (Figure ~~5~~6), this finding suggests that aspects of the model internal structure common within the pairs likely contribute to similarities in simulated ~~fluxes~~flux patterns and associated performance. ~~As shown in Table S1 in the supplementary material, those~~Such features include: 1) SiB3 and 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. ~~Schaefer et al. (2012) suggested that EK and LUE models can perform equally well in simulating fluxes, making this difference a less likely differentiating factor for performance between the two model pairs. Native model time step, on the other hand, has been shown by Schwalm et al. (2010) to be~~

~~an important factor affecting model performance. Finally, using fewer PFTs limits flux variability to larger scales, as indicated in Huntzinger et al. (2011a) who found that NEE simulated by SiB3 has greater variance and smaller correlation length scales than CASA-GFED. Model biases may arise from using uniform parameters within a limited number of PFTs that leads to flux variability at larger scales, while actual processes may vary strongly within each PFT and the actual fluxes vary at small scales (Schaefer et al., 2012; Schwalm et al., 2010). Therefore, we hypothesize that the fewer PFTs and daily-monthly time steps in CASA-GFED and VEGAS2 may explain their relatively poorer performance in simulating NEE spatiotemporal variability relative to SiB3 and ORCHIDEE~~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 ~~develop~~present, evaluate and demonstrate a statistical ~~model selection~~ approach ~~using~~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 apply~~We demonstrate the applicability of this ~~method to evaluate~~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.

~~When~~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 ~~available~~actual atmospheric observations~~, we~~. We find that ~~TBM~~conclusions about model performance ~~in simulating NEE spatiotemporal variability varies strongly across seasons and biomes.~~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 ~~fine-scale flux spatiotemporal variability~~fluxes best during the growing season (May-September) and least consistently with atmospheric observations during the transition seasons, especially in October and November~~. Regionally,~~; and that 2) TBMs appear to perform best over the Temperate Broadleaf and Mixed Forests biome~~, and least well over~~. The experiments performed here also lead to new conclusions about the ~~Boreal Forests and Taiga biome. None of the~~examined TBMs~~ evaluated is consistently better than the other TBMs across biomes and~~

~~seasons, although~~. For example, results show that SiB3 and ORCHIDEE appear to represent the flux variability ~~across more biome~~ within individual biomes and months better relative to CASA-GFED and VEGAS2.

~~The spatiotemporal variability of carbon fluxes can be related to~~In addition, this approach has the potential to link model ~~internal structure and~~performance with environmental processes ~~(Beer et al., 2010; Mueller et al., 2010; Yadav et al., 2012; Gourdji et al., 2012), and our work therefore highlights some potential linkages between model performances and structure/processes. We find distinctive~~, making it possible to test the hypothesis that seasonal differences in TBM performance~~, and hypothesize that these may~~ reflect models' ability to represent the seasonal variability in the dominant environmental controls on fluxes. ~~Future work will be conducted to explore the connection between environmental processes and model performance. In addition, we find that models with more PFTs and shorter native time steps may have an advantage in simulating fine scale flux patterns. It must be noted, however, that the comparison conducted here only included four TBMs, and that these TBMs were not run using a uniform experimental protocol (Huntzinger et al., 2012), therefore making the link between model performance and model structure preliminary at this stage. Repeating the analysis across a larger ensemble of models following a uniform protocol ensemble represents another logical next step.~~

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, ~~113, G00B01~~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~~ Discuss., 10, 11857-11897, 2013,~~ 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.

Belshe, E. F., Schuur, E. a. G., and Bolker, B. M.: Tundra ecosystems observed to be CO2 sources due to differential amplification of the carbon cycle, Ecology Letters, 16, 1307-1315, 2013.

Broquet, G., Chevallier, F., Bréon, F. M., Kadygrov, N., Alemanno, M., Apadula, F., Hammer, S., Haszpra, L., Meinhardt, F., Morguí, J. A., Necki, J., Piacentino, S., Ramonet, M., Schmidt, M., Thompson, R. L., Vermeulen, A. T., Yver, C., and Ciais, P.: Regional inversion of CO2 ecosystem fluxes from atmospheric measurements: reliability of the uncertainty estimates, Atmos. Chem. Phys., 13, 9039-9056, 2013.

Canadell, J. G., Ciais, P., Gurney, K., Le Quéré, C., Piao, S., Raupach, M. R., and Sabine, C. L.: An International Effort to Quantify Regional Carbon Fluxes, Eos, Transactions American Geophysical Union, 92, 81-82, 2011.

Chatterjee, A., Michalak, A. M., Anderson, J. L., Mueller, K. L., and Yadav, V.: Toward reliable ensemble Kalman filter estimates of CO2 fluxes, Journal of Geophysical Research: Atmospheres, 117, D22306, 2012.

Globalview-Co2: Cooperative Atmospheric Data Integration Project – Carbon Dioxide (2010), CD-ROM, NOAA ESRL, Boulder, Colorado, also available at: ftp.cmdl.noaa.gov, Path: ccg/co2/GLOBALVIEW]. NOAA Global Monitoring Division: Boulder, Colorado, U.S.A., 2010.

Göckede, M., Michalak, A. M., Vickers, D., Turner, D. P., and Law, B. E.: Atmospheric inverse modeling to constrain regional-scale CO2 budgets at high spatial and temporal resolution, Journal of Geophysical Research: Atmospheres, 115, D15113, 2010.

Gourdji, S. M., Hirsch, A. I., Mueller, K. L., Yadav, V., Andrews, A. E., and Michalak, A. M.: Regional-scale geostatistical inverse modeling of North American CO2 fluxes: a synthetic data study, Atmos. Chem. Phys., 10, 6151-6167, 2010.

Gourdji, S. M., Mueller, K. L., Yadav, V., Huntzinger, D. N., Andrews, A. E., Trudeau, M., Petron, G., Nehrkorn, T., Eluszkiewicz, J., Henderson, J., Wen, D., Lin, J., Fischer, M., Sweeney, C., and Michalak, A. M.: North American CO2 exchange: inter-comparison of modeled estimates with results from a fine-scale atmospheric inversion, Biogeosciences, 9, 457-475, 2012.

Hayes, D. J., Turner, D. P., Stinson, G., Mcguire, A. D., Wei, Y., West, T. O., Heath, L. S., De Jong, B., Mcconkey, B. G., Birdsey, R. A., Kurz, W. A., Jacobson, A. R., Huntzinger, D. N., Pan, Y., Post, W. M., and Cook, R. B.: Reconciling estimates of the contemporary North American carbon balance among terrestrial biosphere models, atmospheric inversions, and a new approach for estimating net ecosystem exchange from inventory-based data, Global Change Biology, 18, 1282-1299, 2012.

Hoeting, J. A., Davis, R. A., Merton, A. A., and Thompson, S. E.: Model Selection For Geostatistical Models, Ecological Applications, 16, 87-98, 2006.

Huntzinger, D. N., Gourdji, S. M., Mueller, K. L., and Michalak, A. M.: ~~A systematic approach for comparing modeled biospheric carbon fluxes across regional scales, Biogeosciences, 8, 1579-1593, 2011a.~~

~~Huntzinger, D. N., Gourdji, S. M., Mueller, K. L., and Michalak, A. M.:~~ The utility of continuous atmospheric measurements for identifying biospheric CO2 flux variability, Journal of Geophysical Research: Atmospheres, 116, D06110, ~~2011b~~2011.

Huntzinger, D. N., Post, W. M., Wei, Y., Michalak, A. M., West, T. O., Jacobson, A. R., Baker, I. T., Chen, J. M., Davis, K. J., Hayes, D. J., Hoffman, F. M., Jain, A. K., Liu, S., Mcguire, A. D., Neilson, R. P., Potter, C., Poulter, B., Price, D., Raczka, B. M., Tian, H. Q., Thornton, P., Tomelleri, E., Viovy, N., Xiao, J., Yuan, W., Zeng, N., Zhao, M., and Cook, R.: North American Carbon Program (NACP) regional interim synthesis: Terrestrial biospheric model intercomparison, Ecological Modelling, 232, 144-157, 2012.

Keenan, T. F., Baker, I., Barr, A., Ciais, P., Davis, K., Dietze, M., Dragoni, D., Gough, C. M., Grant, R., Hollinger, D., Hufkens, K., Poulter, B., Mccaughey, H., Raczka, B., Ryu, Y., Schaefer, K., Tian, H., Verbeeck, H., Zhao, M., and Richardson, A. D.: Terrestrial biosphere model performance for inter-annual variability of land-atmosphere CO2 exchange, Global Change Biology, 18, 1971-1987, 2012.

Kort, E. A., Eluszkiewicz, J., Stephens, B. B., Miller, J. B., Gerbig, C., Nehrkorn, T., Daube, B. C., Kaplan, J. O., Houweling, S., and Wofsy, S. C.: Emissions of CH4 and N2O over the United States and Canada based on a receptor-oriented modeling framework and COBRA-NA atmospheric observations, Geophysical Research Letters, 35, L18808, 2008.

Krinner, G., Viovy, N., De Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S., and Prentice, I. C.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, Global Biogeochemical Cycles, 19, GB1015, 2005.

Lin, J. C., Gerbig, C., Wofsy, S. C., Andrews, A. E., Daube, B. C., Davis, K. J., and Grainger, C. A.: 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, 4493, 2003.

Mcguire, A. D., Christensen, T. R., Hayes, D., Heroult, A., Euskirchen, E., Kimball, J. S., Koven, C., Lafleur, P., Miller, P. A., Oechel, W., Peylin, P., Williams, M., and Yi, Y.: An assessment of the carbon balance of Arctic tundra: comparisons among observations, process models, and atmospheric inversions, Biogeosciences, 9, 3185-3204, 2012.

Mckain, K., Wofsy, S. C., Nehrkorn, T., Eluszkiewicz, J., Ehleringer, J. R., and Stephens, B. B.: Assessment of ground-based atmospheric observations for verification of greenhouse gas emissions from an urban region, Proceedings of the National Academy of Sciences, 109, 8423-8428, 2012.

Michalak, A. M., Bruhwiler, L., and Tans, P. P.: A geostatistical approach to surface flux estimation of atmospheric trace gases, Journal of Geophysical Research: Atmospheres, 109, D14109, 2004.

Miles, N. L., Richardson, S. J., Davis, K. J., Lauvaux, T., Andrews, A. E., West, T. O., Bandaru, V., and Crosson, E. R.: Large amplitude spatial and temporal gradients in atmospheric boundary layer CO2mole fractions detected with a tower-based network in the U.S. upper Midwest, Journal of Geophysical Research: Biogeosciences, 117, G01019, 2012.

Mueller, K. L., Yadav, V., Curtis, P. S., Vogel, C., and Michalak, A. M.: Attributing the variability of eddy-covariance CO2 flux measurements across temporal scales using geostatistical regression for a mixed northern hardwood forest, Global Biogeochemical Cycles, 24, 2010.

Ogle, S., Davis, K., Andrews, A., West, T., Cook, R., Parkin, R., Morisette, J., Verma, S., and Wofsy, S.: Science Plan: Mid-Continent Intensive Campaign. Greenbelt, Md., 2006.

Peters, W., Jacobson, A. R., Sweeney, C., Andrews, A. E., Conway, T. J., Masarie, K., Miller, J. B., Bruhwiler, L. M. P., Pétron, G., Hirsch, A. I., Worthy, D. E. J., Van Der Werf, G. R., Randerson, J. T., Wennberg, P. O., Krol, M. C., and Tans, P. P.: An atmospheric perspective on North American carbon dioxide exchange: CarbonTracker, Proceedings of the National Academy of Sciences, 104, 18925-18930, 2007.

Pillai, D., Gerbig, C., Kretschmer, R., Beck, V., Karstens, U., Neininger, B., and Heimann, M.: Comparing Lagrangian and Eulerian models for CO2 transport – a step towards Bayesian inverse modeling using WRF/STILT-VPRM, Atmos. Chem. Phys., 12, 8979-8991, 2012.

Ping, C.-L., Michaelson, G. J., Jorgenson, M. T., Kimble, J. M., Epstein, H., Romanovsky, V. E., and Walker, D. A.: High stocks of soil organic carbon in the North American Arctic region, Nature Geosci, 1, 615-619, 2008.

Raczka, B. M., Davis, K. J., Huntzinger, D. N., Neilson, R., Poulter, B., Richardson, A., Xiao, J., Baker, I., Ciais, P., Keenan, T. F., Law, B., Post, W. M., Ricciuto, D., Schaefer, K., Tian, H., Tomelleri, E., Verbeeck, H., and Viovy, N.: Evaluation of continental carbon cycle simulations with North American flux tower observations, Ecological Monographs, 83, 531-556, 2013.

Richardson, A. D., Anderson, R. S., Arain, M. A., Barr, A. G., Bohrer, G., Chen, G., Chen, J. M., Ciais, P., Davis, K. J., Desai, A. R., Dietze, M. C., Dragoni, D., Garrity, S. R., Gough, C. M., Grant, R., Hollinger, D. Y., Margolis, H. A., Mccaughey, H., Migliavacca, M., Monson, R. K., Munger, J. W., Poulter, B., Raczka, B. M., Ricciuto, D. M., Sahoo, A. K., Schaefer, K., Tian, H., Vargas, R., Verbeeck, H., Xiao, J., and Xue, Y.: Terrestrial biosphere models need better representation of vegetation phenology: results from the North American Carbon Program Site Synthesis, Global Change Biology, 18, 566-584, 2012.

Sasai, T., Ichii, K., Yamaguchi, Y., and Nemani, R.: Simulating terrestrial carbon fluxes using the new biosphere model "biosphere model integrating eco-physiological and mechanistic approaches using satellite data" (BEAMS), Journal of Geophysical Research: Biogeosciences, 110, G02014, 2005.

Schaefer, K., Schwalm, C. R., Williams, C., Arain, M. A., Barr, A., Chen, J. M., Davis, K. J., Dimitrov, D., Hilton, T. W., Hollinger, D. Y., Humphreys, E., Poulter, B., Raczka, B. M., Richardson, A. D., Sahoo, A., Thornton, P., Vargas, R., Verbeeck, H., Anderson, R., Baker, I., Black, T. A., Bolstad, P., Chen, J., Curtis, P. S., Desai, A. R., Dietze, M., Dragoni, D., Gough, C., Grant, R. F., Gu, L., Jain, A., Kucharik, C., Law, B., Liu, S., Lokipitiya, E., Margolis, H. A., Matamala, R., Mccaughey, J. H., Monson, R., Munger, J. W., Oechel, W., Peng, C., Price, D. T., Ricciuto, D., Riley, W. J., Roulet, N., Tian, H., Tonitto, C., Torn, M., Weng, E., and Zhou, X.: A model-data comparison of gross primary productivity: Results from the North American Carbon Program site synthesis, Journal of Geophysical Research: Biogeosciences, 117, G03010, 2012.

Schuur, E. a. G., Vogel, J. G., Crummer, K. G., Lee, H., Sickman, J. O., and Osterkamp, T. E.: The effect of permafrost thaw on old carbon release and net carbon exchange from tundra, Nature, 459, 556-559, 2009.

Schwalm, C. R., Williams, C. A., Schaefer, K., Anderson, R., Arain, M. A., Baker, I., Barr, A., Black, T. A., Chen, G., Chen, J. M., Ciais, P., Davis, K. J., Desai, A., Dietze, M., Dragoni, D., Fischer, M. L., Flanagan, L. B., Grant, R., Gu, L., Hollinger, D., Izaurralde, R. C., Kucharik, C., Lafleur, P., Law, B. E., Li, L., Li, Z., Liu, S., Lokupitiya, E., Luo, Y., Ma, S., Margolis, H., Matamala, R., Mccaughey, H., Monson, R. K., Oechel, W. C., Peng, C., Poulter, B., Price, D. T., Riciutto, D. M., Riley, W., Sahoo, A. K., Sprintsin, M., Sun, J., Tian, H., Tonitto, C., Verbeeck, H., and Verma, S. B.: 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, 566-584, 2010.

Shiga, Y. P~~. Y., Anna, M.~~., Michalak, A. M., Gourdji, S. M., Mueller, K. L., and Yadav, V.: Detecting fossil fuel emissions patterns from ~~sub-continental~~subcontinental regions using

North American in~~-~~ situ CO_2 measurements ~~Geophys. Res. Lett., Submitted~~, 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 CO_2 variability, Global Biogeochemical Cycles, 19, GB1016, 2005.

Figures

Figure 1. North American biomes, modified from Olson (2001), as defined for the case studies; ~~stars~~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 ~~3~~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 ~~2~~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 ~~4~~5. Average numbers of months within each season for which the candidate TBM is selected for the SD-all-**ξε** case study (Figure ~~2~~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.

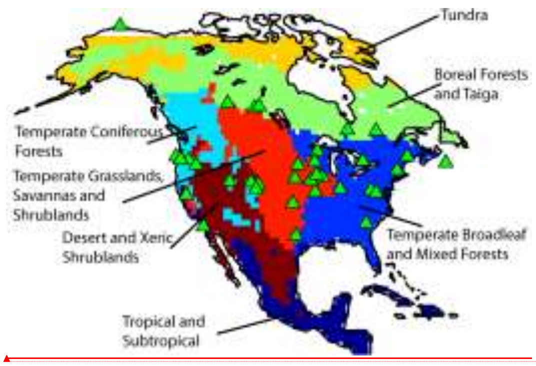
Figure ~~5~~6. The correlation coefficient of NEE spatiotemporal series as simulated by different TBMs throughout 2008 for the four biomes better constrained by available atmospheric observations. TGSS: Temperate Grasslands, Savannas, Shrublands; Bore: Boreal Forests and Taiga; TCoF: Temperate Coniferous Forests; TBMF: Temperate Boradleaf and Mixed Forests.

Figure ~~6~~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 ~~7~~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.

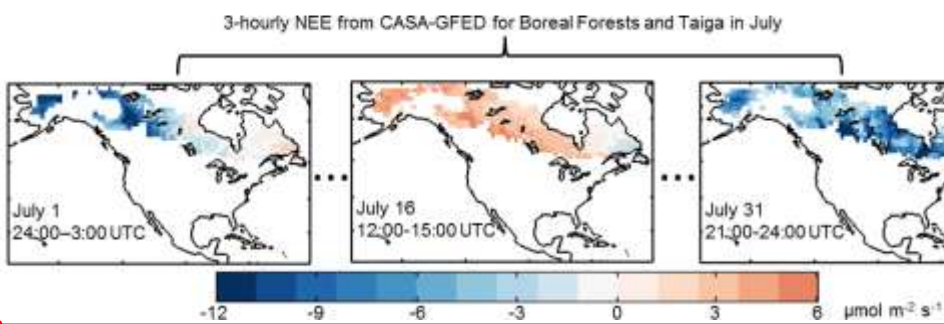
Figures

Figure 1



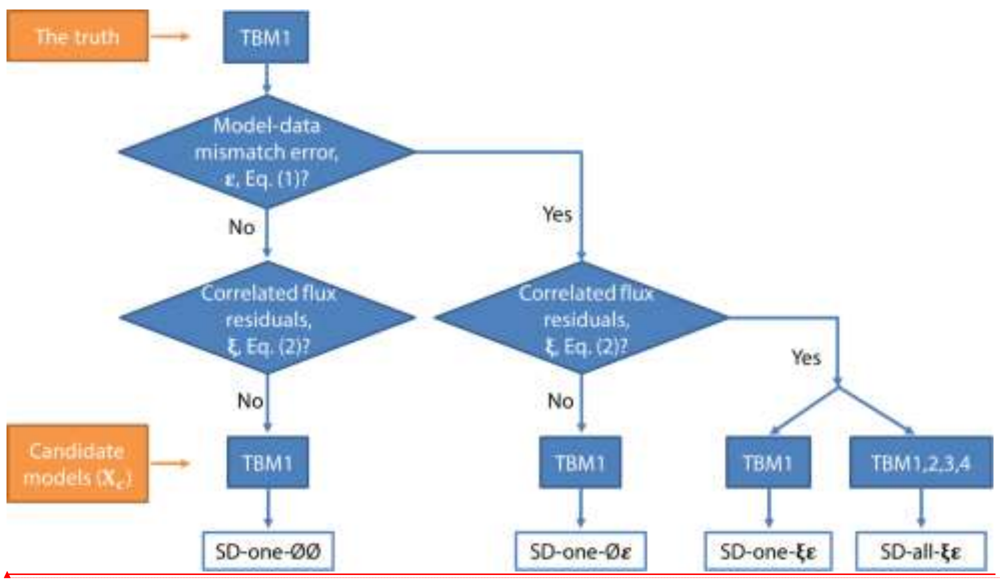
Formatted: Font: 12 pt

Figure 2



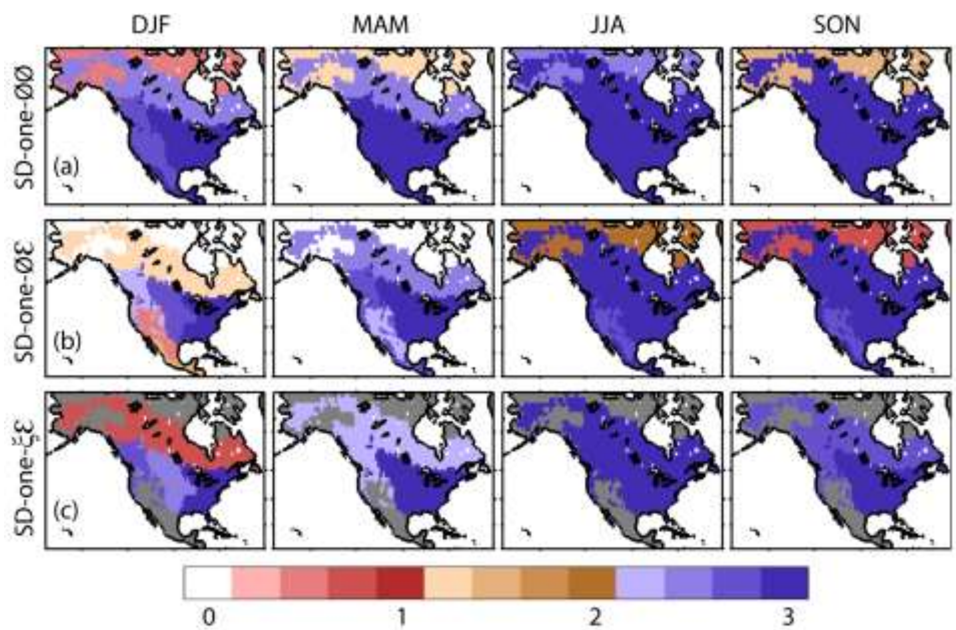
Formatted: Font: 12 pt

Figure 3



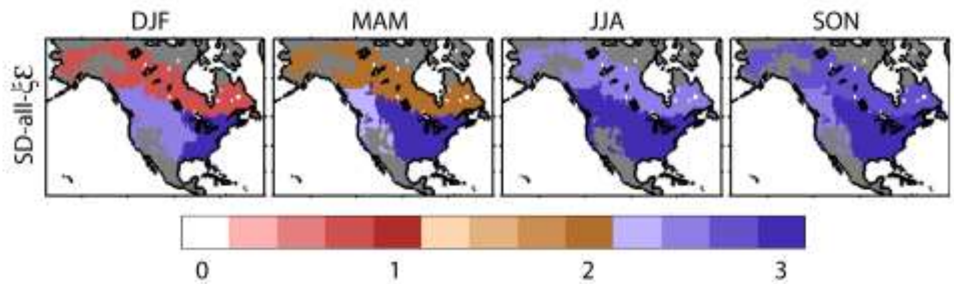
Formatted: Font: 12 pt

Figure 4



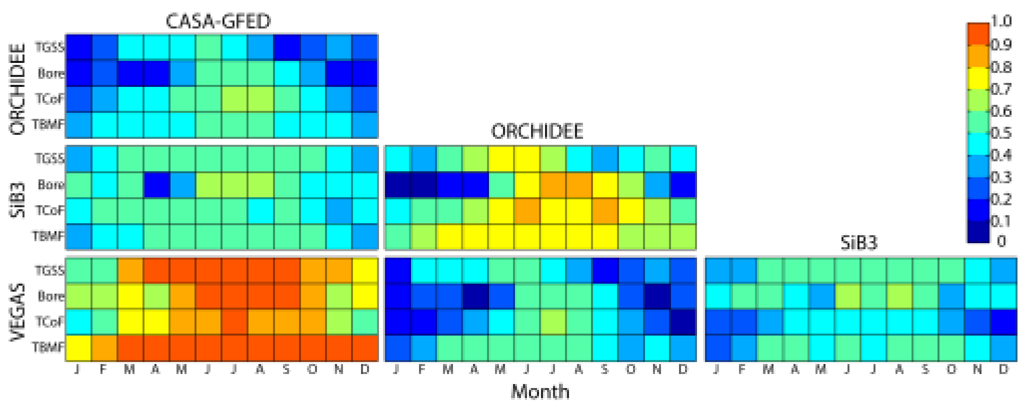
Formatted: Font: 12 pt

Figure 5



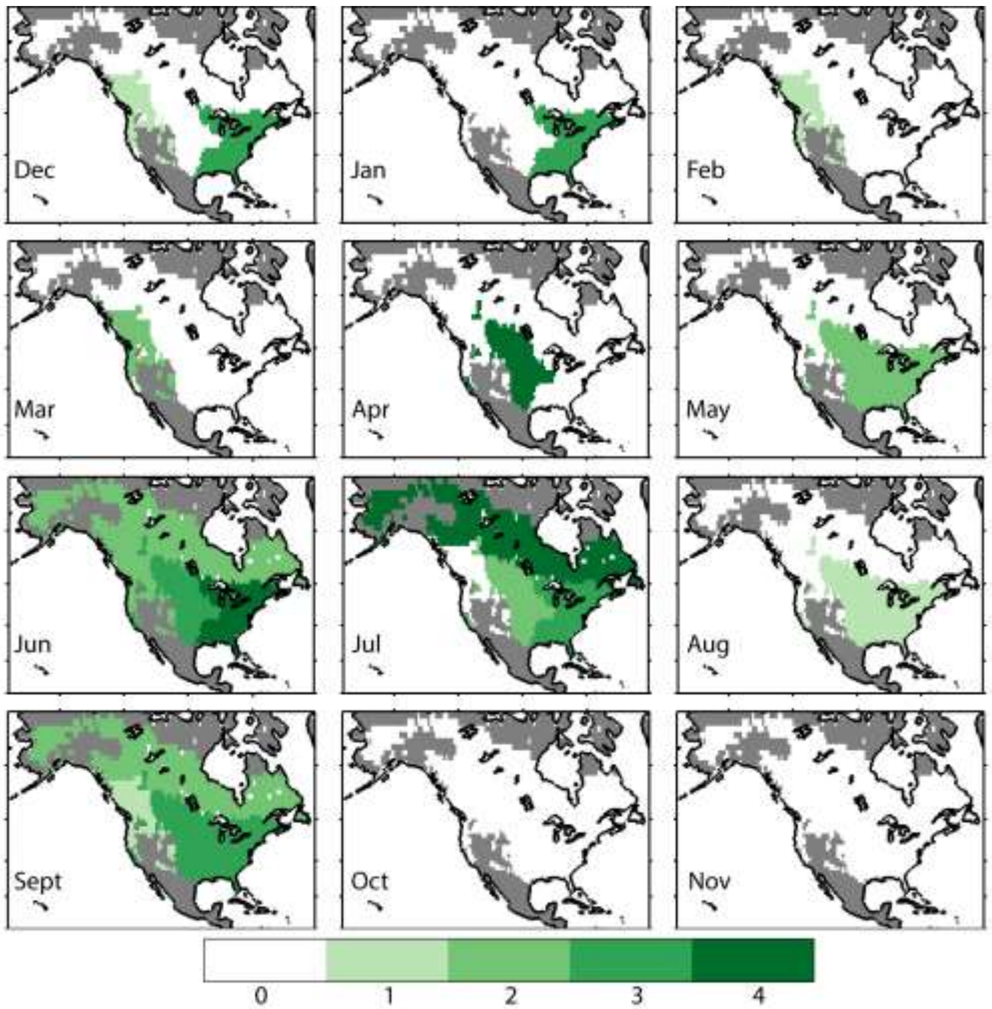
Formatted: Font: 12 pt

Figure 6



Formatted: Font: 12 pt

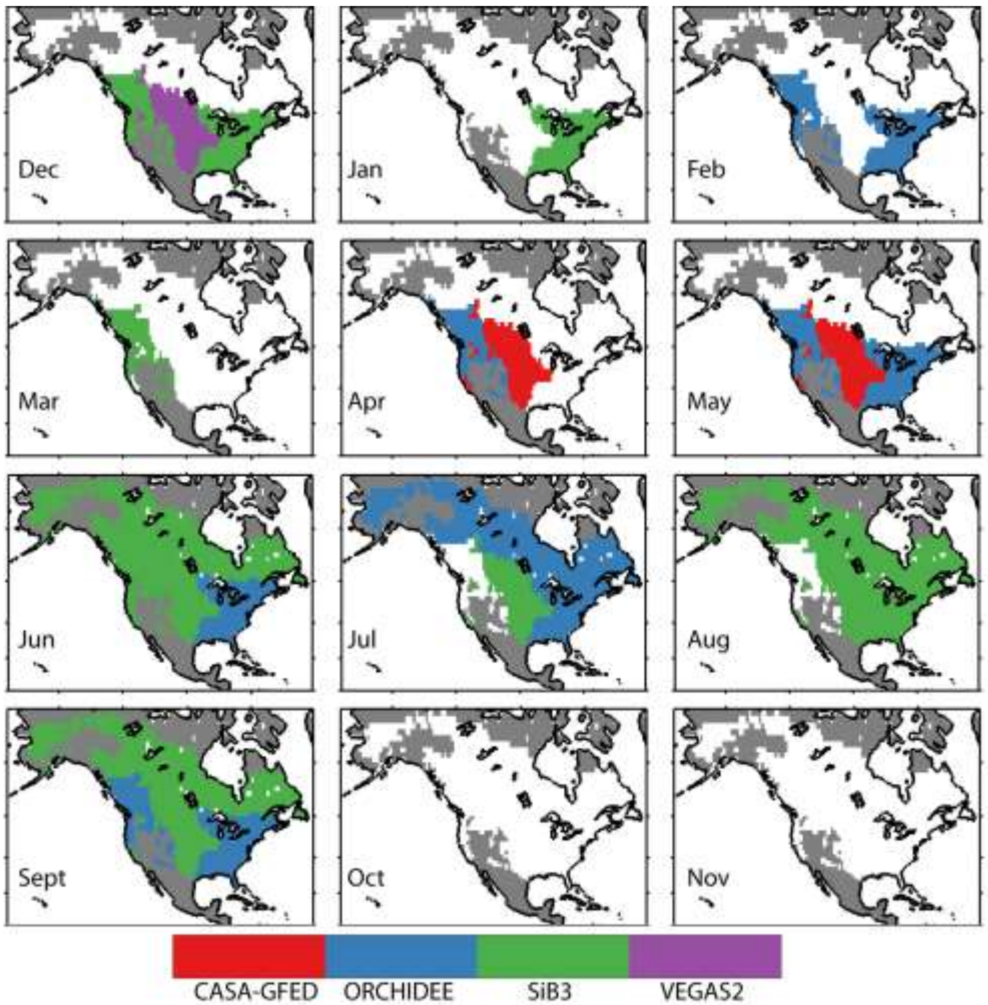
866 Figure 7



867

Formatted: Font: 12 pt

868 Figure 8



869
870

Formatted: Font: 12 pt