Responses to Reviewer #1 (Reviewer's comments are shown in *Italic*)

We thank the reviewer for their valuable comments. From these comments, we realized that some aspects of the original manuscript were unclear, especially with regard to the goal of our study. We address the reviewer's concerns in this letter, and corresponding changes will be made to improve the manuscript.

We begin by summarizing the main high-level issues brought up by the reviewer, based on our understanding of the review:

1. The reviewer found that the actual evaluation of model output using real data is brief and that the explanation of the performance of the different models is not sufficient, especially with regard to difference between EK and LUE models.

Response (1):
We agree with the reviewer that the evaluation of model output using real atmospheric observations is brief.

The main purpose of this paper is to present, validate and test the application of a new method for evaluating the spatiotemporal variability (at a resolution of 1°✕1° and 3-hourly) of NEE simulated by Terrestrial Biospheric Models (TBMs). Interpreting model performance is not the main goal of this paper and is thus not fully discussed.

The ultimate goal of the real data experiments in Section 6 is to further test of the actual application of our method. As a consequence, we only evaluate the performance of a small subset of the NACP models that have been studied extensively in literature. Our discussion in Section 6 is used to show that (1) our results are consistent with earlier literature and (2) that our results complement the existing literature because of the unique model features that we assess. We will revise Section 6 to reflect these goals more clearly.

As we only evaluate a small number of models that use different protocols, our explanations about model performance are hypotheses rather than definite conclusions (see Line 13-16 on Page 9234 in Section 7). We will clarify this in Section 1, 4, 6 as appropriate during revision.

We also state in the manuscript that a future application of our method to a large ensemble of models using a uniform protocol would allow us to identify a cleaner connection between model performance and model structure (see Line 16-18 on Page 9234).

2. The reviewer was concerned about whether our method is a test of the terrestrial biosphere model or a test of the spatial coverage of the observation network.

Response (2):
Our method is designed to evaluate the spatiotemporal variability of NEE simulated by TBMs.

Like any other model evaluation method, our method only works when and where there are sufficient data constraints.

In order to ensure that our results reflect model performance, we conducted thorough synthetic data experiments to test the sufficiency of current data constraints to assess model performance (see Section 5). Through these synthetic data experiments (in particular, the first four paragraphs of Section 5), we find that atmospheric data provide sufficient constraints in four major biomes of North America, so that our results there reflect model performance. Consequently, we discuss results from the real data experiment only in those biomes where the selection of a TBM reflects its performance in simulating the spatiotemporal variability of NEE.

In the revised manuscript, we will clarify that results for all biomes are affected by both the coverage of the network and the capability of the TBM to simulate NEE variability. However, our synthetic data cases suggested that current data constraints are sufficient to assess model performance only in four biomes.

3. The reviewer was also concerned about the complexity of atmospheric transport and mixing processes and the uncertainties associated with boundary conditions, the transport model, and fossil fuel emissions.

Response (3):
We agree with the reviewer that atmospheric transport and mixing processes are complicated. It is also true that this study, just as other regional inverse modeling work, is subject to uncertainties related to fossil fuel emissions, the transport model and boundary condition.

These concerns motivated us to design synthetic data experiments with varying levels of complexity, in which we know the true underlying flux patterns, to test whether our method works in evaluating the simulated NEE variability. These experiments are also conducted to test the sensitivity of atmospheric data to underlying flux patterns. We will clarify in the revised manuscript that all SD experiments are designed both to validate our methodology and to test the sensitivity of atmospheric data. We will revise Section 5, including its title, to reflect this.

Even in the most sophisticated experiment (SD-one-**ξε**) which considers realistic levels of model-data mismatch errors (including errors in transport, boundary conditions, fossil fuel emissions, and measurement and aggregation errors) and spatially-correlated flux residuals, our method can still identify TBMs that simulate a substantial portion of the NEE spatiotemporal variability using atmospheric CO_2 measurements in many biome-months. See Line 21-24 on Page 9221 and Section S3 in supplementary material for the explanation and definition of model-data mismatch term; see Line 18-21 on Page 9225 and Line 1-15 on Page 9228 for the description and results of SD-one-**ξε**, Line 1-14 on Page 9226, 18-25 on Page 9229 and 1-19 on Page 9230 for the description and results of RD one-**ξε**.

Comparing results from SD-one-**ξε** and RD one-**ξε**, a TBM being selected in SD-one-**ξε**, for a given biome-month while not being selected in RD-one-**ξε** suggests that this TBM does not represent the true flux patterns. We acknowledge that even the most sophisticate SD case (SD-one-**ξε)** does not fully address the complexity of the real world, because we assume that errors are independent and follow a Gaussian distribution. This is the reason for which, when conducting the RD experiments, we intercompare results extensively with the existing literature. By comparing our evaluation results with the literature, we find consistent results (i.e., better performance during the growing season, poorer performance during transition seasons etc.), further supporting the idea that our method is able to assess model performance despite the extenuating factors.

Results previously reported in the literature also support the idea that the factors described by the reviewer, although important, are unlikely to alter the main results presented in this work. Specifically:

Fossil fuel emissions:
It is a common practice to assume that uncertainties in fossil fuel emissions are “small” when using atmospheric inverse modeling to estimate net ocean and land fluxes; the uncertainty associated with fossil fuel emissions is not a unique challenge to our method, but is a common problem for atmospheric inversions of CO_2 [Peylin et al., 2011]. In addition, fossil fuel inventories for North America are more sophisticated and quite likely more accurate than those for other regions, further limiting the possible impact of errors in fossil fuel emissions. Third, our method aims to identify TBMs that can represent a portion of the variability in atmospheric observations, not the entire variability. Therefore, even if there are some residual FF signals, one should still be able to see the biospheric patterns. As a result, we expect that directly subtracting the fossil fuel signal from atmospheric measurements is not a major impediment to the application presented here.

Boundary condition:
As Gourdji et al. [2012] showed, while different boundary conditions affect the absolute magnitudes of regional NEE inferred by regional inversion, they yield similar spatial patterns of NEE. We thus expect that different boundary conditions would have a smaller effect here than in inverse modeling studies that focus on flux magnitudes.

Transport model uncertainties:
Pillai et al. [2012] pointed out that the flux differences between two inversions, using footprints from two different atmospheric transport models (Eurlerian v.s. Lagrangian), are a factor of two smaller than the inversion-observation mismatch and a factor of three smaller than the mismatch between TBM and measurements. Hence, we expect that transport model uncertainties, although important, do not make it impossible to discern flux patterns from atmospheric observations, at least for the four biomes that are identified through the SD test cases as being better constrained by the available atmospheric observations.

We will revise the manuscript to reflect the discussion above.

Having addressed the high-level issues, as we understood them, we now address the reviewer's individual comments in detail.

General Comments: (Overall Quality of Paper)
The purpose of this manuscript is twofold: first it tests a methodological approach that compares continuous CO2 observations against simulated NEE fluxes, and second applies this methodological approach to evaluate 4 models from the Regional Continental Interim Synthesis (RCIS) across 35 towers within North American during 2008.

Response (4):
As mentioned in Response (1), the purpose of this paper is to present, validate and test the application of a new method for evaluating the spatiotemporal variability (at a resolution of 1°×1° and 3 hourly) of NEE simulated by TBMs. The goal of both the synthetic and real data test cases is to validate the methodology and test its applicability under realistic conditions.

The strength of this paper is a test of the methodology, in part established by Gourdji and others and combined with the Bayesian Information Criterion, for selection of TBMs that best match variation in atmospheric CO2. The authors rigorously test the design of experiment with pseudo-data to evaluate the effectiveness of the method for idealized situations. The authors find that only a subset of biomes (e.g. grassland, conifer and deciduous) provide a large enough biogenic signal to make atmospheric CO2 a useful diagnostic for surface fluxes. This is not surprising considering the method is incapable of detecting skillful model performance in poorly sampled regions or biomes with minimal variation in seasonal NEE (e.g. tundra).

Response (5):
We are glad that the reviewer finds the test of the methodology to be a strong aspect of the paper. Testing the methodology is indeed the main goal of this paper.

It is true that, due to the insufficient data constraints, we are not able to draw conclusions about model performance in some biomes, including Tundra.

When applied to real model output, the authors found that two EK models (SIB3, ORCHIDEE) perform better than two LUE models (CASA-GFED, VEGAS2), which the manuscript states is most likely the result of the time step and number of PFTs included in these models. In general, the models perform better during the growing season and poorer during the transition seasons.

Response (6):
We want to point out that TBM output applied in both the synthetic and real experiments are the same "real" model output. The only difference between the synthetic and real data experiments is in the atmospheric measurements: in the SD cases, synthetic measurements are generated using simulated NEE, while in the RD cases actual

atmospheric measurements are used.

We also want to clarify that we do not suggest that EK models perform better than LUE models in the paper. In fact, we show that an EK vs. LUE formulation is unlikely the reason for differences in model performance (please see Line 16-17 on Page 9234); we hypothesize that better model performance may be attributable to the number of PFTs and the model time step, both of which are not distinguishing features of EK vs. LUE models. We will clarify our discussion regarding EK and LUE models during revision.

Drawing a definite conclusion on the causes of model performance is not the main goal of this methodological paper. We will make this clear throughout the paper during revision. Please also refer to Response (1).

Overall, the evaluation of real model data is brief, with limited interpretation, and largely reiterates findings of previous manuscripts within the NACP synthesis. This is the main weakness of the paper. This reviewer recommends including more years and models to improve the scope of the paper, with more focus upon the attribution of model performance.

Response (7):
We agree that the evaluation of models using real data is brief. We intend to keep it simple, as the purpose of this paper is to present, validate and test the application of our methodology to evaluating the spatiotemporal variability of simulated NEE.

The RD experiments are conducted as a test of the application of our methodology in the real world, where we do not know the true fluxes. To do so, we only apply our method to evaluate a small subset of NACP models that have been extensively examined in the literature.

In particular, for the RD experiments,
(1) We want to show that our methodology finds reasonable results, in agreement with earlier findings where/when results are indeed comparable. We therefore discuss our findings in the context of the existing literature, i.e., reiterate findings from previous work and compare it with our findings, see Line 3-11 on Page 9230 and Line 14-17 on Page 9231.
(2) We want to show that our methodology provides further insights relative to earlier literature, as it focuses on a new model feature that has not been evaluated before. We therefore discuss the spatiotemporal similarity between various model pairs and its connection with the performance of these models, and are able to attribute observed model performance to three aspects of model structure, see Line 1-15 on Page 9232).
(3) We also want to highlight the potential of our methodology to inform process-level model evaluation, because the feature of modeled NEE that we target (i.e., spatiotemporal variability) is sensitive to spatiotemporal variability in environmental drivers (see Line 24-28 on Page 9231).

Please also refer to Response (1) and the fourth paragraph of Response (3) for additional details.

This reviewer also recommends comparing the TBM model performance against those findings in Schwalm et al. 2010 (site runs) and Raczka et al. 2013 (regional runs) as part of the NACP Synthesis. This should better identify consistencies/ inconsistencies across the different model validation approaches.

Response (8):
We agree with the reviewer that a comparison between different methods is valuable.

However, as model features evaluated in those studies are different from our study, a direct comparison is not possible. For example, Schwalm et al. [2010] and Raczka et al. [2013] evaluate monthly and/or annual mean carbon fluxes at a site level, while this study evaluates the spatiotemporal variability (at a resolution of 1°✕1° and 3 hourly) of NEE.

Nevertheless, the relative performance of TBMs across biomes and seasons are comparable. For example, we show in the paper (see Line 2-8 on Page 9230) that TBMs have problems during transition seasons, consistent with what was found in Schwalm et al. [2010]; however, Raczka et al. [2013] suggested that models have good performance in simulating seasonality when evaluating their performance in simulating interannual variability. In addition, our hypothesis that the number of PFT and the model time step may affect model performance (in Line 18-25 on Page 9232) is also consistent with what was suggested by Schwalm et al. [2010]. Furthermore, our finding of a better performance over Broadleaf and Mixed Forests is also shown in Raczka et al. [2013] and Schwalm et al. [2010]. Finally, our finding that the SiB model is selected more often is consistent with results in in Raczka et al. [2013] and Schwalm et al. [2010], despite different model features being assessed in those studies. We will discuss these comparisons in Section 6.

As the feature of modeled NEE that we target (i.e., spatiotemporal variability) differs from other studies, our results provide additional insights relative to the existing literature. For example, we find SiB and ORCHIDEE are more similar and perform consistently better, which enables us to tentatively attribute model performance to model structures shared only by these two models. Another example is related to poor model performance during transition seasons. We find models have problems in simulating the fine-scale (at a resolution of 1°✕1° and 3 hourly) NEE spatiotemporal variability during transition seasons, while earlier studies show problems in capturing the timing of phenology using monthly mean carbon fluxes. Both results then suggest that the transition seasons pose particular challenges for TBMs, although for different reasons. Therefore, it is likely that certain common model features may contribute to both problems.

Specific Comments: (Individual Scientific Questions/Issues)
It was unclear to what extent this method is a test of the terrestrial biosphere model vs. the spatial coverage of the observation network. A part of the methods identifies biome-months that were not well constrained by atmospheric concentrations and eliminates

them from further consideration, nevertheless, the approach is influenced by the observation network.

Response (9)
Our method is developed to examine the spatiotemporal variability of NEE simulated by TBMs. However, the reviewer is right that this method is affected by the observational network, just as any other model evaluation method would be. For example, some biome-months in which a TBM is not selected could be due to insufficient data constraints or poor TBM performance (see Line 15-19 on Page 9224).

In order to ensure that our results reflect model performance, we conduct thorough synthetic data experiments to examine the sensitivity of atmospheric data to NEE variability and to validate our methodology. From those experiments, we find that atmospheric data provide sufficient constraints in four major biomes. Over those biomes, synthetic data cases indicate that the selection of a TBM with our method directly reflects model performance. As a result, results from the real data experiments are only discussed over those four biomes, where our method has proven capability to test model performance. We will clarify those points during revision. Please also see Response (2).

Furthermore, the method attempts to tease out the behavior of biogenic fluxes, when the variation of real data is due to boundary conditions, fossil fuel emissions and function of model transport. The method 'subtracts out' the influence of boundary conditions and fossil fuel emissions, but still the strength of the biogenic signal is weakened because of these confounding influences. The method requires fitting of the TBM NEE to a statistical model, and it is this statistical model, which is tested for how well it can simulate the observed atm. CO2. This seems like a challenging undertaking to connect the model surface fluxes to network of atm. CO2 observations.

Response (10)
We agree with the reviewer that this study, just as other regional inverse modeling work, is subject to uncertainties related to fossil fuel emissions, the transport model and boundary conditions.

However, even with our most realistic SD case (SD-one-**ξε** which considers realistic model-mismatch errors (which include errors in fossil fuel emissions, transport and boundary conditions), our method is still able to identify TBMs that represent a substantial portion of the NEE variability. We therefore expect that our major results are not likely affected by those uncertainty sources. Please also refer to Response (3) for details.

This reviewer remained skeptical how a method that emphasizes the ability to assess fine-scale processes, can use atmospheric CO2 observations, a state variable that is inherently dependent upon an integral of influence across a large spatial area, as an effective way to accomplish this. Even though the TBMs provide fluxes at relatively fine spatial/temporal resolution, the measured flux tower CO2 can be the result of many confounding, fine scale processes within the footprint.

Response (11)
We agree with the reviewer that atmospheric transport and mixing processes are complicated and solving an inverse problem related to those processes at fine resolution is challenging. However, the existing literature shows that atmospheric CO_2 observations are sensitive to fine-scale NEE spatial and temporal variability [Huntzinger et al., 2011] and that atmospheric inversions can take advantage of those measurements to infer regional NEE at fine scales [Göckede et al., 2010a; Göckede et al., 2010b; Gourdji et al., 2010; Gourdji et al., 2012]. We will add those references in the manuscript to show the potential of using atmospheric measurements to infer NEE at fine scales.

In order to test whether atmospheric data can be applied to infer the fine-scale spatiotemporal variability, we designed synthetic data experiments with varying levels of complexity, in which we know the true underlying flux patterns, to test whether our method works in evaluating the simulated NEE spatiotemporal variability. These synthetic data experiments suggest that our method can identify TBMs that simulate a substantial portion of the NEE spatiotemporal variability over four major biomes. We will clarify the capabilities of our method during revision. Please also see Response (3) for details.

It is unclear to this reviewer whether this approach offers any additional insights into model processes other than what has already been demonstrated from previous RCIS publications. The most important findings that the models perform better during the growing season, rather during the transition seasons is already established by works such as Richardson et al. 2012. Although the author attempts to make a distinction between 2 types of models (enzyme kinetic, light-use efficiency), there is limited attribution (e.g. model time step, # pfts) to why the performance is different.

Response (12)
The main purpose of this paper is to present, validate and test the application of our methodology to evaluate the simulated spatiotemporal variability of NEE. Understanding and interpreting model performance is not the main goal of this paper.

Our approach aims to evaluate the simulated spatiotemporal patterns of NEE, a model feature not addressed by the existing literature. Therefore, our results complement earlier literature that focuses mainly on aggregated magnitudes of NEE fluxes, or on site-specific NEE fluxes. For example, regarding model performance during growing season, our results, when examined together with those of Schwalm et al. [2010] and Richardson et al. [2012], suggest that models perform poorly in simulating the timing of the seasonal cycle and the fine-scale spatiotemporal variability of biospheric carbon fluxes.

In Section 6, we present a few hypotheses based on our approach about the attribution of model performance to model internal structure (EK vs. LUE; number of PFTs; the model time steps) based on evaluating the a small number of NACP models. Future application of our method to a large ensemble of models using a uniform protocol will allow us to maker more direct connections between model performance and model structure. Please also refer to Response (3).

In general, TBMs perform the best at simulating diurnal and seasonal variation, which it seems this methodology is sensitive to. Models do not simulate magnitude of integrated carbon fluxes well, nor do they simulate interannual variation in flux well. This is something this method/paper does not address. It is unclear to this reviewer whether this approach distinguishes between models that have large biases in the carbon flux, or only evaluates model skill through variation in model flux.

Response (13)
The reviewer is right that our paper does not evaluate the magnitude of integrated carbon fluxes or the interannual variation of carbon fluxes.

Simulated carbon fluxes need to be evaluated in terms of their magnitude, as well as their spatial and/or temporal variability at a full spectrum of scales. We evaluate the spatiotemporal patterns of NEE at a resolution of 1°×1° and 3 hourly within each month and biome. That is to say, we examine whether TBMs can simulate the spatiotemporal patterns of NEE within each month-biome combination. We do not evaluate the seasonality of carbon fluxes; however, our results reflect seasonal differences of model performance in simulating such fine-scale NEE variability.

We will clarify our focus in the revised manuscript.

The author provides cursory explanation of the difference in model performance between growing season and transition season. The author does not make a clear distinction between phenology (e.g. leaf timing, leaf activity) and model representation of photosynthetic/respiration processes influenced by environmental variation. They can be viewed as distinct entities. In other words, do the models perform more poorly during the transition seasons because simply the leaf timing (phenology submodels) are wrong, or because other physiological processes are not well represented during this time?

Response (14)
To explain what causes different model performance between growing season and transition seasons is not the main purpose of this methodological paper. We will further clarify our goal in the revised manuscript. Please also see Response (1).

It was difficult for this reviewer to reconcile why SIB3 and ORCHIDEE (EK models) performed better than CASA-GFED2 VEGAS (LUE models). Given that the models tended to perform worse in the transition seasons likely from the poor phenology. Presumably LUE models (remote sensing of phenology) would do better than EK models (internal prediction of phenology) during transition seasons.

Response (15)
We clarify that we do not suggest that EK models perform better than LUE models. In fact, we show that EK vs. LUE formulations are not likely the reason for differences in model performance (please see Line 16-17 on Page 9232). We will clarify our point

during revision. Please also refer to Response (1) and (6).

In addition, we are not sure that we agree with the reviewer that LUE models use remote sensing phenology while EK models do not. According to our knowledge, SiB3 (as an EK model) and CASA-GFED (as a LUE model) use remote sensing phenology, while ORCHIDEE (as an EK model) and VEGAS (as a LUE model) do not [Huntzinger et al., 2012; Richardson et al., 2012; Schaefer et al., 2012].

This reviewer was left with the question of how do we improve the models? How does this method, in particular, provide an upgrade in the diagnosis of model behavior from the classical approaches of tower-site evaluation vs. aggregated region analysis between bottom up, inversion and inventory approaches? I think the method is interesting and has potential, but the payoff was not clear at least for this application.

Response (16)
The reviewer asks a good question. The ultimate goal of the proposed approach is to inform the improvement of TBMs. This paper is the very first step towards that goal. It describes a new method for evaluating model performance. To inform model improvement is not within the scope of this paper. Once the method is established, it can then be used to evaluate a large ensemble of models to understand their performance and inform the improvement of these models.

In the introduction, we state that information from classical tower-site assessments is only representative of small spatial scale processes while evaluations based on aggregated regional analysis do not provide directly useable information regarding the environmental processes driving carbon exchange (Line 8-18 on Page 9217).

Our method aims to evaluate the spatiotemporal variability of carbon fluxes on biome level (i.e., whether TBM can simulate the fine-scale spatiotemporal variability of NEE over each biome). Such variability has been shown to be sensitive to spatiotemporal variability in environmental drivers (see Line 24-28 on Page 9231); therefore, our method can be potentially applied to inform process-level model evaluation. We will make this potential clearer during revision.

Technical Corrections:
Page 9216, Line 21: "...differences among the examined models are at least partiallyattributable to their internal structures". The RCIS from which the TBM output was taken from, was not based upon a controlled protocol. It is conceivable models of identical internal structure could provide different results just from differences in driver data.

Response (17)
We agree with the reviewer. We show in the paper that the differences are only "partly" attributed to the internal model structure. The other "part" could be attributed to differences in driver data, as suggested by the reviewer. We will add this point in Section 6.2.

Page 9217, Line 28: "…novel path…" Assessing the spatial/temporal variability of carbon fluxes to evaluate TBMs is not a 'novel' approach. It has been done quite often. This particular inversion methodology seems to be the 'novel' part.

Response (18)
We agree with the reviewer and will make corresponding changes to our manuscript.

Page 9227, Line1: Seems contradictory to say that the tundra biome has a weak biospheric signal, but it plays an important role in the carbon cycle in climate. I think you need to be clear that it is expected to play a more important role in future.

Response (19)
We agree with the reviewer and will make corresponding changes to our manuscript.

Page 9231, Lines 17-24: Not sure what this means. What specifically are the models not simulating correctly?

Response (20)
Here model limitations include incorrect timing of phenology based on evaluating monthly carbon flux data [Richardson et al., 2012; Schaefer et al., 2012; Schwalm et al., 2010] and the inconsistent fine-scale carbon flux variability in the transitional months with that inferred by atmospheric data (our study). We will clarify this during revision.

Page 9234, Lines 7-9: This is vague. What are the seasonal differences in performance attributable to?

Response (21)
We attribute better/poor performance during growing/transition seasons to different model capabilities in representing dominant environmental controls on fluxes across seasons. For example, Mueller et al. [2010] showed that radiation is the dominant driver during the growing season for a temperate forest, while temperature is the dominant driver during non-growing and leaf-out seasons. Our hypothesis is that models may better be able to simulate the effect of variation in radiation than that of variation in temperature. We will clarify this point during revision.

Page 9234, Lines 16-19: It would be more worthwhile to perform this analysis as part of the MsTMIP (Huntzinger) experiment, which includes a controlled protocol.

Response (22)
We agree with the reviewer. Applying our method to evaluate MsTMIP models is in our research plan.

Figure 1: The stars denoting the location of the CO2 observations are hard to read, especially in dark-colored biome regions.

Response (23)
We will make corresponding changes in the figure.

Figure 2: Shouldn't 3rd row be labeled as 'correlated flux residuals' not just 'flux residuals'?

Response (24)
We will make corresponding changes to the figure.

Figure 3 caption: Unclear what the 'grey' threshold (average number of months) is for map figures, which were determined to not be well constrained by atmospheric data.

Response (25)
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.

References:

Göckede, M., A. M. Michalak, D. Vickers, D. P. Turner, and B. E. Law (2010a), Atmospheric inverse modeling to constrain regional-scale CO2 budgets at high spatial and temporal resolution, Journal of Geophysical Research: Atmospheres, 115(D15), D15113, doi:10.1029/2009JD012257.

Göckede, M., D. P. Turner, A. M. Michalak, D. Vickers, and B. E. Law (2010b), Sensitivity of a subregional scale atmospheric inverse CO2 modeling framework to boundary conditions, Journal of Geophysical Research: Atmospheres, 115(D24), D24112, doi:10.1029/2010JD014443.

Gourdji, S. M., A. I. Hirsch, K. L. Mueller, V. Yadav, A. E. Andrews, and A. M. Michalak (2010), Regional-scale geostatistical inverse modeling of North American CO2 fluxes: a synthetic data study, Atmos. Chem. Phys., 10(13), 6151-6167, doi:10.5194/acp-10-6151-2010.

Gourdji, S. M., et al. (2012), North American CO2 exchange: inter-comparison of modeled estimates with results from a fine-scale atmospheric inversion, Biogeosciences, 9(1), 457-475, doi:10.5194/bg-9-457-2012.

Huntzinger, D. N., S. M. Gourdji, K. L. Mueller, and A. M. Michalak (2011), The utility of continuous atmospheric measurements for identifying biospheric CO2 flux variability, Journal of Geophysical Research: Atmospheres, 116(D6), D06110, doi:10.1029/2010JD015048.

Huntzinger, D. N., et al. (2012), North American Carbon Program (NACP) regional interim synthesis: Terrestrial biospheric model intercomparison, Ecological Modelling, 232(0), 144-157, doi:http://dx.doi.org/10.1016/j.ecolmodel.2012.02.004.

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.

Peylin, P., et al. (2011), Importance of fossil fuel emission uncertainties over Europe for CO2 modeling: model intercomparison, Atmos. Chem. Phys., 11(13), 6607-6622, doi:10.5194/acp-11-6607-2011.

Raczka, B. M., et al. (2013), Evaluation of continental carbon cycle simulations with North American flux tower observations, Ecological Monographs, 83(4), 531-556, doi:10.1890/12-0893.1.

Richardson, A. D., et al. (2012), Terrestrial biosphere models need better representation of vegetation phenology: results from the North American Carbon Program Site Synthesis, Global Change Biology, 18(2), 566-584, doi:10.1111/j.1365-2486.2011.02562.x.

Schaefer, K., et al. (2012), A model-data comparison of gross primary productivity: Results from the North American Carbon Program site synthesis, Journal of Geophysical Research: Biogeosciences, 117(G3), G03010, doi:10.1029/2012JG001960.

Schwalm, C. R., et al. (2010), A model-data intercomparison of CO2 exchange across North America: Results from the North American Carbon Program site synthesis, Journal of Geophysical Research: Biogeosciences, 115(G3), 566-584, doi:10.1029/2009JG001229.