Responses to Reviewer #2 (Reviewer's comments are shown in *Italic*)

We thank the reviewer for their valuable comments. These comments are very constructive, and will help us to improve the manuscript, specifically in terms of clarifying our methodology and the goal of this paper. We address the reviewer's concerns in this letter, and corresponding changes will be made to improve the manuscript.

General Comments: (Overall Quality of Paper)

This is mostly well-written paper describing a novel approach to the recently often-pondered problem of reconciling multiple constraints to understand the uncertainty in the magnitude and patterns of ecosystem CO2 fluxes. Because of the need to evaluate land surface models in order to make more reliable estimates of future change, this study is valuable for its contribution of an alternative, useful and complementary method of comparison. The approach is notable for its emphasis on the seasonal and spatial variability in fluxes, which may at the current state-of-the-science be a good balance for model evaluation between more commonly used comparisons of flux magnitudes that are either regionally aggregated or compared at specific sites.

Response (1)

We thank the reviewer for this positive feedback. The reviewer understood the main goal of our manuscript and the implications of our method.

Though the progress made in the application here is still modest, it can be considered an important step toward bridging the gap between a relatively sparse network of atmospheric concentration measurements and highly uncertain gridded flux simulations.

The lack of insight provided here into the reasons behind differences in the performance of particular models is mostly for reasons beyond the scope of this study. This point is touched on somewhat in the discussion, but it could be better emphasized that the objective of the paper is (or should be, in this reviewer's opinion) to demonstrate a method that can make use of atmospheric observations to evaluate model performance from a new perspective – rather than a detailed study of particular modeling structures themselves. In other words, this reviewer's eyes typically glaze over at the use of "made up" data, but in this case the synthetic experiments were much more informative than the "real" ones.

Response (2)

The reviewer has a good understanding of the goal of our manuscript. We agree that the synthetic data experiments are a key aspect of this analysis. The real data experiment is intended, for the purposes of this study, primarily as a further evaluation of the approach.

We appreciate the reviewer's constructive suggestion which will help to clarify the goal of the manuscript and avoid confusion from the readers. As suggested, we will clarify that the objective of the paper is to demonstrate a method that can make use of atmospheric observations to evaluate model performance from a new perspective – rather than a detailed study of particular models themselves.

The background and impetus for the study are well described and justified in the introduction. Other than the aforementioned unnecessary strain in an attempt to explain differences in individual model performance, the discussion is solid and the conclusions sound. The section that needs some attention is the methods; it could use more clarity in explanation for not only the individual steps but also how they all fit together. Otherwise, this is a nice paper overall and an important contribution. It could be improved relatively easily by the authors giving it another run through and taking opportunities to focus on the value of the approach and what it can (and can't) tell us about reconciling models with observations.

Response (3)

We agree with the reviewer that additional details regarding the methodology would be useful. In the interest of being succinct, we referred to other papers for aspects of the model setup that were common with previous work, but we will re-examine the balance between completeness and brevity in the revision.

At the same time, interpretation of the "real data" experiments is still useful and necessary, but it should be couched more as a demonstration or proof-of-concept with respect to the types of analysis that could be done with the methods, rather than a futile attempt to say something about particular model structures of specific models performing better or worse on certain indicators. Again, the reasons why are not necessarily the responsibility of the authors here, but instead they can take this opportunity to call out these issues and challenge the modeling community to address them going forward.

Response (4)

The review understood our intention well. We intended to use real data experiments as a further test case of our method. We will rebalance the discussion of the real data experiment in Section 6 to make it clear that it was intended as a proof-of-concept of the approach.

Specific Comments: (Individual Scientific Questions/Issues)

Section 1

Overall a good set up of the issues and good justification for the work presented here. The fifth paragraph (9218: 16 – 23) could use a little more explanation to help convey the importance (along with the caveats and uncertainties) of the observations and inverse modeling to this study. How is this done, and why? Some additional detail here would balance better with the more lengthy discussion of TBM issues above. One suggestion would be to have an explicit statement of objective(s) in there somewhere; perhaps at the top of the 6th paragraph (9218: 24), e.g.

"what is needed is a method that: : :" This where you can make clear that the emphasis of this paper is on the value of this method.

Response (5)

We thank the reviewer for their careful thoughts and constructive suggestion to improve our paper. We will add a further discussion of caveats and uncertainties in this Section during revision. We will also include an explicit statement of objectives at the beginning of paragraph 7 as suggested by the reviewer.

The 7th paragraph seems a bit unconventional; it is not common to have this kind of "table of contents" paragraph. Not that there is anything wrong with non-conformity when needed, it's just that this isn't.

Response (6)

We understand the reviewer's suggestion, and will consider removing the road map if we decide that it is not necessary to ensure that the readers understand the structure of the paper.

Section 2

Figure 1 should be referenced in the first sentence (9220: 18 – 20) to illustrate the locations of the towers in relation to the biomes. This could also use a quick note in the text about how well distributed (or not) these towers are across biomes. On the figure itself, consider different symbols for the towers and/or color schemes for the biomes; the tower locations are difficult to see.

Response (7)

We thank the reviewer for this suggestion. We will revise the text and the figure as suggested.

There is a general problem in this section in terms of passing off potentially important details to previous publications. For example, 9220: 22 -23 talks about processing and sub-selection. Can this be expanded to two sentences with a little more explanation of each, at least within the context of this study? How many observations are in the sub-sample? Is that important or an issue to deal with in the model selection? There could also be a little more explanation on the inverse method, for non-experts. What is this doing and what is the product?

Response (8)

We agree with the reviewer, and we will revisit this Section to include additional details while at the same time striving to keep it succinct.

Fossil fuel emissions impacts are pre-subtracted (9220: 2 – 3), but not fire emissions (like CarbonTracker)? Is this an issue that might impact your analyses? Similarly, there is no *discussion in section 2.3 about how the models define NEE and which components are included or not. For example, CASA-GFED includes the impacts and emissions from fire while SiB does not. Is this important?*

Response (9)

The reviewer brings up an important point. Relative to Fossil Fuel emissions, fire emissions estimates are less accurate, and are therefore not presubstracted. There were indeed a few significant fires occurring in North America in 2008: including summer California fires (May-Sept), Evans Road Wildfire in Eastern NC (July), and Targo Fire in New Mexico (Apr-May).

As we focus on the variability of carbon fluxes at the resolution of 3 hourly and 1×1 degree, we expect that the impact of fires on the spatiotemporal flux patterns would be small relative to that of other biospheric fluxes especially these fires occurred during the growing season. In addition, our method intends to identify whether a TBM explains a portion of the variability, rather than the total variability. For TBMs whose NEE includes no fire emissions (such as SiB3, ORCHIDEE and VEGAS), we expect it to be selected if it represents a substantial portion of biospheric signals; for TBMs whose NEE include fire emissions (such as CASA-GFED), we expect it to be selected if it represents a substantial portion of biospheric signals or a mixture of biospheric and fire signals (while the latter component is expected to be smaller). The results presented here are therefore not expected to be substantially impacted by fires that occurred in 2008.

We will add a brief discussion about fire emissions as an additional source of uncertainty during revision.

Section 3

This section overall was very difficult for a non-expert (admittedly not as fluent in matrix algebra as one would like) to grasp what was happening even after several re-reads. Could this explanation improve with more clarity in text and/or a visual rather than a reliance on equations? Using words like "true fluxes" (9221: 17) can become confusing without more explicit definition.

Response (10)

We agree with the reviewer. We will try different ways, including using a diagram, to make the text clearer and easier to understand during revision. We clarify that "true flux" means the underlying flux that causes the variability in actual atmospheric CO_2 measurements.

The most troublesome was the seeming mismatch in resolution of a model based on 'biome-month' to predict "true" fluxes at 1x1 and 3-hrly resolution. This is so key to the whole thing that it probably should be spelled out and/or illustrated more clearly and in more detail. On the other hand, it was curious that the BIC was defined / described in such detail (rather than just saying what it means for model selection here) compared to the relative lack of explanation of *inverse methods, transport models, the derivation and character of the predictor and dependent variables, etc.*

Response (11)

We agree with the reviewer that "the seeming mismatch in resolution of a model based on 'biome-month' to predict "true" fluxes at resolution" is a difficult concept to convey. Meanwhile, we realize that this concept is also one of the most important details of the whole approach. We will work out a way to make it crystal clear. As a start, we can describe the method as evaluating the spatiotemporal variability of NEE "within" each biome-month, rather than "for" each biome-month, because then it makes it clearer that we are not aggregating / averaging fluxes to biome- or monthly-scales, instead, the method evaluates the 1x1 and 3-hrly NEE patterns within each biome and within each month. Some figures could potentially help to illustrate what we mean. Below is a possible example to show the spatiotemporal variability of NEE within the Boreal Forests and Taiga biome in July.

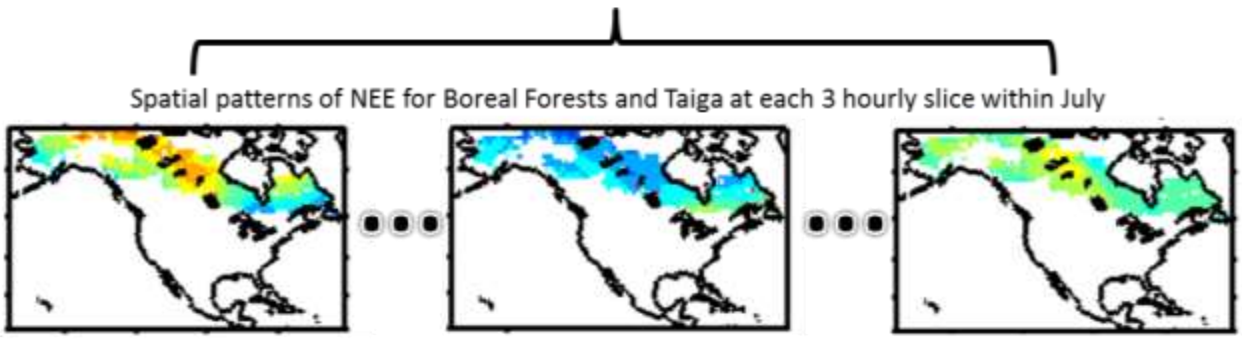

Figure 1. Illustration of the potential schematic to show the spatiotemporal variability of NEE simulated by one TBM within Boreal Forests and Taiga in July (i.e., one ancillary variable as in a column of **X**); each panel shows one example of simulated NEE spatial pattern at the 1×1 degree resolution within this biome at one 3-hourly time slice within July.

Section 4

The synthetic and real data case studies are well defined, explained and justified, and Figure 2 is helpful to see the linkages. A sentence or two that gives the basic jist on how the mismatch errors and residuals were synthetically generated would be helpful here (9225: 2 – 25), rather than leaving it all to the Supplement.

Response (12)

We agree with the reviewer, and will consider moving some necessary details about errors and residuals into the main text.

Section 5

This section is well-written, clear and does well highlighting the key aspects of this study. The synthetic experiments show that this method works for what it is designed for, and it can provide useful and complementary information in combination with other established approaches. That they highlight the biomes where (and why) CO2 fluxes are not well constrained lends further credence to the need for additional observations in sparsely sampled biomes.

Response (13)

We agree with the review, and thank the reviewer for their nice summary.

Section 6

The more interesting results of the "real data" experiments are when general insight can be drawn from the performance of the models as an ensemble. While the conclusions drawn here seem well-founded (e.g. better in the growing season and in particular biomes that are probably better-studied), the potential reasons for the differences in performance of the individual models (e.g. "capturing seasonal variation", "internal structures") are vague and purely speculative. This paper would be just fine without this inter-comparison piece, in fact it's removal or strong, caveated de-emphasis would improve the paper with respect to clarity / proper focus on the method and less on anything specific about these four (of many) particular biospheric models.

Response (14)

We agree with the reviewer about the well-founded and yet vague aspects of the discussion presented in Section 6. We find the suggestions here to be very constructive and we will follow these comments to revise Section 6, focusing explicitly on a further clarification and a strong, caveated de-emphasis of this Section. This comment from the reviewer is also in line with concerns voiced by Reviewer #1, so please see our responses there for further details.

Technical Corrections:

9217: 1, remove "the" in "the carbon cycle science"

9218: 29, remove the "n" in "North American" 9219: 26, change to ": : :gives a more realistic: : :"

9220: 2, should CarbonTracker be cited here?

Response (15)

We thank the reviewer for pointing out these technical issues. We will make corresponding changes to our manuscript during revision.