[Note that the reviewer's comments are shown below in regular text and the authors' response to the reviewer's comments are shown in **bold**]

Comments from Reviewer #1

General comments

This paper provides a description of the MsTMIP effort and general design of the experimental protocol. As part of a multipart publication, it is a useful document to allow readers a more efficient way to focus on aspects of a large intercomparison such as this instead of wading through a single large paper. It is well-written (increasingly rare!) and well-organized

From intercomparison always comes the inevitable issue of assessing if some models are "better" or perhaps contain algorithmic components that can be deemed "better" or "more complete", etc. Comparison to observed quantities is often what many think of as the metric by which this is determined, but this isn't always clear because sometimes observations and model simulated quantities are not directly comparable. There may also be underlying diagnostic tests that, though not directly tied to an observable, are considered metrics to test model "realness". In atmos models, things like interhemispheric gradients of a conserved quantity might qualify. There may be similar quantities in TBMs that are emergent yet identify lack of "realness" or suggest internal inconsistencies. Worth thinking about and if the authors have such metrics, worth writing about in this paper (maybe just a para)

> **As the Reviewer mentioned, model evaluation is challenging because comparisons because of a lack of direct observations. Our analysis approach does consider some evaluation of model seasonal and diurnal cycles for different outputs (e.g., carbon and energy fluxes), which are somewhat simple analogs to the interhemispheric gradients in atmospheric models mentioned by the reviewer. However, these types of "emergent" property comparisons are limited for land-atmosphere carbon models. As such, our evaluation approach is focused more on comparing model output to a range of observational products, while also recognizing the information content limitations of each. In order to address this more clearly in the manuscript we have modified the wording in Section 4 to read:**
>
> **"There is often an inconsistency between observations and models in terms of the variables being measured/simulated, as well as the temporal and/or spatial resolution of the models compared to those of observations. Thus, the alternative is to use some combination of model-model comparisons, comparisons of model output to related observations at relevant scales, and comparisons of model output to equivalent variables/observations at somewhat mismatched scales. In recognition of these challenges, the MsTMIP benchmarking activities will emphasize a combination of benchmarking approaches, including site eddy-covariance data (e.g., NEE, latent heat, sensible heat), regional products (e.g., aboveground biomass; *Saatchi et al*., 2011) and gridded model-data products (e.g., upscaled GPP from *Jung et al.,* 2011)."**

The authors chose to share "preliminary results" in this manuscript. Though interesting, it might be unclear what the rationale for this is, especially given that the paper has the stated aim of providing an overview and description of experimental design. I do appreciate prelim results, but the question is where to draw the line? Why this set of preliminary results? Why not more or different? Are they important to the narrative about the design? It might help to motivate why you are showing these here and why this particular suite of results.

> **The authors (along with the second reviewer) feel that the manuscript adequately provides the reasoning and justification for including preliminary results. In short, the preliminary results show that although the experimental design of the MsTMIP does control for some of the variability of model results (when compared to the NACP RCIS), there is still significant inter-model spread or variability, underscoring the importance of model structure (model choice) on model estimates. Given that Reviewer #2 felt the reasoning behind showing preliminary results was made clear in the manuscript, we have chosen not to edit this section of the manuscript.**

Specific comments

I don't see how the quantification of structural differences as contributors to across-model variability will contribute to evaluation and feedback to improve the state of the art (page 3982, lines 20-25, paraphrased). It seems to me that you should be a bit more upfront about the differences between understanding the variability that emerges (tracing it to structural differences or parameter choices) and determining what choices, approaches, parameterizations are "better". The latter is challenging in more ways than one (note under general comments), but I think intercomparisons have to be somewhat "tough" in this regard, particularly since you will avail of observations that will evaluate the output of the models (and that is key). In other words, don't be too shy or overly diplomatic about that fact that some models will not perform as well as others against observations. This doesn't always mean some are better and some are worse, but it is this upfront comparison alongside the structural understanding that will lead to model progress, I believe.

> **The variability that emerges due to structural differences among the models can indeed be used to inform understanding of what choices, approaches, etc. are 'better." At a minimum, understanding how structural differences drive inter-model spread can be used to illustrate levels of uncertainty when a discriminating choice among candidate results cannot be made due to lack of available evaluation/validation data.**
>
> **In order to address the reviewers comment, we have augmented the wording at the end of Section 4 to read:**
>
> **"In addition, methods to evaluate model structural differences, similar to the dendrograms presented in this manuscript, will be used to attribute differences in estimates between subsets of models to differences in model structure. For example, do model estimates of global long-term mean GPP cluster similarly to model structural attributes? Such side-by-side comparisons will inform understanding of the drivers to inter-model**

differences in estimates of carbon fluxes and carbon pools. By better understanding the variability that emerges due to structural differences among the models, the MsTMIP activity can help inform understanding of what modeling structural choices or assumptions lead to improved model estimates. At a minimum, understanding how structural differences drive inter-model spread can help inform our understanding of model uncertainty, particularly when a discriminating choice among candidate results (e.g., which model is "best") cannot be made due to lack of available evaluation/validation data."

Page 3985, lines 5-14: It can also lead to observing systems better-tuned or optimized to test models – not to be underestimated as an added benefit of model-obs comparisons. Vice-versa is true: ensuring models generate variables that can be directly compared to observable quantities is an important goal in this subsection text.

We agree with the reviewer's comments. On lines 11 through 14 of the Discussion Manuscript, we state: "Comparing model predictions with available observations and data-driven products can help identify knowledge/information gaps in both the models and the observations, and advance process-level understanding of land-atmosphere carbon exchange." In order to clarify the points made by the reviewer, we have added the following sentence to the end of this paragraph:

"This may ultimately lead to observing systems that are better optimized for evaluating model performance, as well as encouraging models to generate output that can be more directly compared with available observations."

Parameter variation is mentioned quickly on page 3984 (near top) and one wonders if there will be some form of systematic exploration of the parameter space in addition to structural space? In this way you can generate an RMS of the internal uncertainty and the external uncertainty (model spread).

Parameter variation was not controlled for in the MsTMIP experimental design, thus a systematic exploration of parameter space is not feasible within the MsTMIP activity. As stated in the manuscript, parameter uncertainty is better assessed within a given model. Thus the MsTMIP activity complements parametric uncertainty analysis, as stated at the end of the first paragraph on page 3984 in the discussion manuscript.

Technical corrections

Page 3984, line 18: I think the spelling should be "rationale"?

Fixed according to suggestion

Page 3984, line 23: "the" at the end of the line should be removed.

Fixed according to suggestion

Page 3990, line 3: the sentence starting on this line is run-on. Break into two?

Revised the wording to at the end of paragraph 1 of Section 2.5 to read:

"In order to compare model estimates with atmospheric CO_2 concentration and FLUXNET data, carbon and energy fluxes are also collected at 3-hourly intervals for the time period of 1980 to 2010 for the reference (RG1, RR1) and baseline (BG1, BR1) simulations. For some models, however, generating 3-hourly output was not feasible. Thus, for these models, carbon and energy fluxes were collected at the finest temporal resolution possible for that model over the final 30 years of MsTMIP simulations."

Page 3990, line 23: grammar, perhaps "the" should be removed?

Fixed according to suggestion

Page 3992, line 1: perhaps "but inter-model differences remain, particularly…"?

Fixed according to suggestion