Reply to reviewer #2 by Thum et al. (The original comments by the reviewer are in violet and the replies by the authors are in black.)

Thum et al. present an interesting approach for the evaluation of soil carbon models in a land surface model by using atmospheric CO2 observations. I like the basic idea of the study and the work is methodological sound. However, I actually got bored and disappointed when I was reading the paper. This is too a large degree caused by the presentation (text and figures) of the material (I agree with all the points by reviewer 1):

We thank the reviewer for these views on the paper.

1) It is not clear what the purpose of the study is. Only in the discussion it's written that the "aim was to use atmospheric observations to benchmark soil carbon models". If this was the aim, the evaluation of just two modules of JSBACH is insufficient (and causes my disappointment). It would be better to clearly state already in the introduction that the aim was to evaluate two soil modules of JSBACH.

We have removed the word "benchmark" from the manuscript, as the aim of this paper is not to do a purely numerical ranking of two different models. We change the title to include 'JSBACH', therefore the scope of this paper should be then visible already in the title.

However, if this is the case,the manuscript would do better as a model evaluation paper in Geoscientific Model Development. Generally, the text is written as a model evaluation study and I don't find important results for the general Biogeosciences. My feeling is that the paper should go beyond a model evaluation and include some more substantial scientific questions, hypotheses and findings in order to fit into Biogeosciences.

The aim of this study was to assess, whether we can use the atmospheric CO_2 concentration to evaluate two different soil carbon models and assess, what can we by these means say about the functionality of the soil carbon models. We do address scientific questions in this study. Such as, whether the soil carbon content or environmental responses of the decomposition are more responsible for the differences seen in the results. In writing the revised version of the manuscript we will include scientific questions this study addresses to the introduction.

2) If the "aim was to use atmospheric observations to benchmark soil carbon models" (as stated), I would expect a more detailed description of the assumptions and a detailed analysis on how to use atmospheric CO2 observations for the benchmarking, including how to disentangle the contributions of GPP and Reco on CO2, the role of uncertainties in observations and atmospheric transport, and how different regions contribute to the CO2 seasonality. Especially the later points would help to potentially benchmark soil carbon simulations if different parts of the world.

We will pay more attention to the points raised by the reviewer in the revised version of the manuscript. It is not possible in our current set-up to disentangle the contributions of GPP and Reco, but the GPP is same for both of the model simulations and we re-write the text to remove using the word benchmarking, so that our main aim to see the contribution of two different soil carbon modules in the atmospheric CO_2 molar fraction becomes more clear.

We will add text on the observational and atmospheric transport uncertainties. The atmospheric transport model was the same in the both model simulations, so we don't expect this to be causing differences we see in the two different model runs. The role of different regions contributing to the seasonality of the CO_2 molar fraction can be studied by setting the land fluxes to zero at certain regions, as done earlier by Dalmonech and Zaehle (2013). We did not for now consider this to be necessary for this analysis.

3) As already stated by reviewer 1, the text needs substantial rewrite. The text has no clear structure, topic sentences are missing, some chapters are too long (especially 1and 3.1). For example, the first section of the results (3.1) report mainly minor results (including references to the supplement) but does not report the most important results. In addition, I recommend to split this section in further sub-chapters to improve the structure of the text.

Thank you for the advice in improving the re-writing of the manuscript. Concerning the order of the results section: This work was done in a step-wise manner, when first the JSBACH simulations were performed and they were then fed to TM5. Starting with the atmospheric CO_2 results (which we would consider being main results of the study) potentially makes it challenging for the reader to follow the story line. Having said that, we will bear this view in mind during re-writing and try to highlight important results before minor results.

4) Figures: I'm sorry, but reading the figures in the main text and in the supplement wasa nightmare! The figures are too small and the grey colours make it almost impossible to distinguish the different model runs and observations. Please improve all figures.

We thank the reviewer for this insight and improve the figures in the revised version of the manuscript.

5) The discussion of GPP is over-simplistic. JSBACH overestimates GPP and has in some regions shifts in the seasonality. Hence it remains unclear which soil carbon model is the better one because the comparison of CO2 seasonality is also affected by wrong simulations of GPP. Could it be an option to force more realistic GPP estimates into JSBACH or mix GPP from data-driven estimates with Rh from JSBACH in TM5?

As this study was also a methodological test for us to see, if using different soil models would indeed result in differences in the atmospheric CO_2 cycle, we wanted to only use the 'off-shelf' version of JSBACH. It is obvious that using a full land surface model brings up constraints. What was not successful in writing of the earlier version, was to bear this in mind when interpreting the results. Overall, in this model evaluation exercise, a performance competition of the two different models was not the main goal, but to find out what characteristics of the soil carbon models need improvement. For the YAS model the finding, that the precipitation should not be used as a proxy for soil moisture at this temporal and spatial resolution, is relevant and independent of JSBACHs GPP overestimation.

The reviewer has a suggestion of forcing more realistic GPP into JSBACH. This aim can be achieved by data assimilation, as done by Castro-Morales et al. (2019) where they used another, earlier version of JSBACH. However, this approach would be a study by itself to be done first.

As for using more data-driven GPP estimate, this unfortunately doesn't work in our simulation set-up, as this would break mass-balance of the terrestrial carbon cycle and the litterfall and GPP are strongly coupled in such a model. There would be several biases resulting from this. There certainly is a lot of value in more data-based approaches and that kind of work would be best done by a test-bed set-up similar to works by Wieder et al. (2018). We will now mention this point in the discussion of the revised version.

6) Please describe if permafrost was simulated in the JSBACH runs and how the simulation or non-simulation of permafrost contributed to soil carbon simulations.

Permafrost was not included in the simulations and we will mention this in the new version of the manuscript clearly. Recent model development adding permafrost to JSBACH includes only the YAS model (Castro-Morales et al., 2018). Including permafrost would increase the carbon stocks in high latitude regions. The exact influence on the atmospheric signal is speculation without the actual model runs, but it would be expected that the seasonal cycle of heterotrophic respiration in high latitude regions would be dampened, since the active layer producing heterotrophic respiration would be thinner.

Figure 1: Explain the numbers in the caption.

Thank you for noticing this was missing. We have now changed the caption to be:

Locations of GAW stations, denoted as black dots, and different TransCom regions (different numbers denote the different TransCom regions in this study) as different colors.

Figure 2: Even if the soil carbon stocks have been already evaluated, it would be still helpful to add 1 or 2 maps from an observation-based product for comparison.

Sure, we will do this.

Table 2: There seems to be a mistake in the results for TER, as those can't be the same numbers.

Yes, thank you, the YAS number should be 155 PgC.

References

Castro-Morales, K., Kleinen, T., Kaiser, S., Zaehle, S., Kittler, F., Kwon, M. J., Beer, C., and Göckede, M.: Year-round simulated methane emissions from a permafrost ecosystem in Northeast Siberia, Biogeosciences, 15, 2691–2722, https://doi.org/10.5194/bg-15-2691-2018, 2018.

Castro-Morales, K., Schürmann, G., Köstler, C., Rödenbeck, C., Heimann, M., and Zaehle, S.: Three decades of simulated global terrestrial carbon fluxes from a data assimilation system confronted with different periods of observations, Biogeosciences, 16, 3009–3032, https://doi.org/10.5194/bg-16-3009-2019, 2019.

Dalmonech, D. and Zaehle, S.: Towards a more objective evaluation of modelled land-carbon trends using atmospheric CO_2 and satellite-based vegetation activity observations, Biogeosciences, 10, 4189–4210, https://doi.org/10.5194/bg-10-4189-2013, 2013.