Dear Anne and coauthors,

first of all, I acknowledge the delay for this manuscript dues to the finalization of the PhD thesis of the lead author and proceeds with this manuscript in February 2020. I hope that your defense of the thesis was successful and that you find more time for this manuscript now.

I am summarizing below the agreements between you (the authors) and the TE (Kirsten Elger) for this manuscript (based on several email exchanges and a skype call in late November 2019):

Agreed Workflow suggested by the authors:

- We write a public response to reviewer #3 (So that the interactive discussion would show a RC3 and AC3);
- More reviewers can then review the current version of the manuscript (suggestions below*);
- Then we would, if necessary, make a new version of the manuscript/dataset after the reviews (=reviewer #3 + some more).
- Let us know whether this would work for you. We would then need the Copernicus system to show the RC3 and get a possibility to upload a AC3.

Detailed answers to specific requests below. Please proceed with the revision of this manuscript at your earliest convenience.

TE:

All three reviewers request a significant change of your data and paper. There is a general concern about the very little number of input variable (due to the use of the surface variables and not the 3D fields of CMIP5/PMIP3 data for which there would be many more data available - this is the main question of Reviewer #2).

I have also contacted Reviewer #1 and asked whether a new setup of your model would be possibly improving your data quality. The answer here was "It's possible the group would be willing to include more models and variables, though it could be a significant time investment, hard for me to say. My guess would be that their results would change significantly, which is one of the reasons I'm skeptical about this endeavor. The CORE2 forcings have been a powerful tool because they combine different kinds of information and converge on a reasonable set of atmospheric forcings; if more observations were added, the results probably wouldn't change too much. At the LGM, there is no guarantee that averaging models will converge on anything at all, let alone the "right" annual cycle. So while there's value in what they're trying to do, estimating the LGM annual cycle and its uncertainty is a deep question and I worry that characterizing the present treatment as an LGM CORE2, even with more models added, risks oversimplifying the problem."

Authors comments:

(1) Use of multi-model mean fields

Regarding Reviewer #1, and the comment on the robustness of the dataset: We show that after adding the GISS model (difference between version 1 and version 2 of the dataset), the

mean anomaly fields look almost identical (i.e., the figures in v2 of the manuscript are nearly the same as in v1), showing that adding more data will likely not significantly change our results. Regarding Reviewer #3, providing individual anomalies for each of the models is something we could certainly provide.

Regarding the comments of reviewer #1, we stress (again) that we do not try to provide a true estimate of the LGM annual cycle, nor are we able to provide a LGM "normal year" forcing – this is because available model output or other data is too limited, not because we are reluctant to take into account suggestions to improve our data set. As described in our author response, we aim to provide a large scale monthly mean anomaly field that can be used with any type of pre-industrial ocean forcing (but formatted for easy use with CORE forcing, see also the first paragraph of Sect. 2), to obtain an estimate of the LGM atmospheric state. We believe that this is currently (given the available data) the best estimate we can produce. Regarding the comments on averaging over multiple models (Reviewer #1 and #3), this is an accepted practice in model intercomparison studies (e.g. HappiMIP uses multi model average SSTs, Mitchell et al. (2017)) and to force stand-alone models (Muglia et al., 2015/2018), as well as outside of the paleoclimate and MIP community (f.e., Chowdhury and Behera, 2019). There are hundreds of papers analyzing the multi model mean and spread of CMIP5 models. We have discussed and defended the inconsistencies that will occur through such averaging in our earlier author response and incorporated that in the main text (Sect. 2). Nevertheless, as mentioned earlier, we could include individual model anomalies in an updated version of our data set.

(2) Use of 3D model output and calculation of 10m temperature and humidity

Reviewer #2 shared some of his/her remaining concerns with you, as far as we understand after reading our author response. We understand this is mainly about the use of the 2D surface fields. First, we would like to note that the use of 3D output would actually not lead to many more available data, as it would be just seven models (adding CCSM and MPI-ESM) as compared to 5 models in version 2 of our dataset.

Second, unfortunately, we will not be able to use 3D model output, because the consistent calculation of 10m temperature and humidity would require the implementation of parts of the boundary layer scheme for each model, which is beyond what we could do. But, more importantly, we cannot agree with the reviewer (#3) that such a procedure would result in more accurate 10 metre temperature and humidity after taking into account limitations due to data availability. After receiving the reviews we re-discussed this issue with an expert in boundary layer meteorology. The bottom line is that an inconsistency with the original model is unavoidable as the calculation is done off-line (and due to data availability: only climatological monthly means are available in PMIP), but the error made will be generally much smaller by using our method (shifting from 2 to 10 meters with an inconsistent method) than doing what the reviewers ask for (shifting from the lowest 3D model level, i.e. over a factor of 10 larger vertical distance, with a still inconsistent method as it is done off-line and on climatological monthly mean data). We would be able to address the point of reviewer #3 about estimating the impact of using monthly averages instead of higher resolution data in our calculations.

(3) Inclusion of river input

Regarding river fluxes (variable ‘friver’), an anomaly field would not be meaningful due to differences in the land-sea mask. Nevertheless, we see an opportunity to provide the user with (basin-scale) LGM-PI changes in river runoff based on the different models, which we expect will satisfy the reviewer.

TE: Additional comment:

Seeing the third review (who reviewed already the revised version) also mentioning major concerns with the rescaling method for temperature and humidity and the gerenal interest of the dataset as "the use of an average of CMIP5 ocean models to force ocean model might not be a relevant experiment protocol".

Authors: this will be addressed in a direct answer to Reviewer #3 in the public discussion

Many thanks and best regards,

Kirsten Elger