Response to Reviewer 2

Nils Weitzel, Heather Andres, Jean-Philippe Baudouin, Marie Kapsch,
Uwe Mikolajewicz, Lukas Jonkers, Oliver Bothe, Elisa Ziegler,
Thomas Kleinen, André Paul, and Kira Rehfeld

We thank the Reviewer for assessing our manuscript and providing constructive feedback that will strengthen our manuscript. The following is a point-by-point response to the Reviewer's comments. Here, the comments by the Reviewer are shown italicized in teal and our responses are provided in black.

In this manuscript the authors present a new methodology to compare simulated and reconstructed sea surface temperatures and apply it to the case of the last deglaciation. The method nicely separates different aspects of temperature variability on different time-scales and will become a valuable tool to quantify model-data agreement as new transient model simulations and more proxy records of past climate intervals become available. The paper is well written and the results are clearly presented and I therefore recommend publication of the paper in Climate of the Past after some, mostly minor, issues have been addressed.

We thank the Reviewer for this positive assessment of our work.

I'm not convinced that the title properly reflects the content of the paper, namely a comparison of simulated and reconstructed sea surface temperatures. Possibly reformulate to something like:

Towards spatio-temporal comparison of simulated and reconstructed (sea surface) temperatures for the last deglaciation

We thank the Reviewer for this suggestion to improve the title of the manuscript. In the revised manuscript, we will change the title to *"Towards spatio-temporal comparison of simulated and reconstructed sea surface temperatures for the last deglaciation"*.

I realize that this is a predominantly methodological paper, but the authors could possibly consider to shorten the technical part a bit by moving some details to the supplementary, and focus a bit more on the results in terms of how well different models reproduce different aspects of the temperature evolution over the last deglaciation. For example Fig. 6 seems to add very little information and the corresponding section 4.1.3 seems very extended considering that the important messages are simply that i) the reliability and robustness of the algorithm seem to be very little affected by misspecified SNRs and ii) the effect of under- or overestimating the temporal persistence of non-climatic noise is negligible in our PPEs.

We understand the wish of the Reviewer to reduce the technical part of the paper a bit. However, we also have to give the interested reader a chance to understand how we obtain the respective results without extensively consulting the supplemental information. Therefore, we would prefer not to shorten the methods section and instead giving the reader the option to skip technical subsections, e.g. Sect. 3.1.1 to 3.1.4., if they are more interested in the results of the application of the methodology. Following the advice of the Reviewer, we will merge Sect. 4.1.2 and 4.1.3 and move the previous Fig. 6 to the supplemental information. We still think that the two relatively simple messages mentioned by the Reviewer are of relevance because they can provide some guidance for future proxy system model developments. In addition, they were not a priori (i.e., before running the pseudo-proxy experiments) obvious to us. Therefore, we will still include them in the results section and not move this part of the analysis fully into the supplemental information.

A nice addition to Figures 8 and 9 would be a figure showing also a direct comparison of simulated and reconstructed temperature (anomalies) time series for the different regions. I believe that a simple visual inspection of model and observation time series is still useful to get a first idea about model-data agreement.

We thank the Reviewer for this suggestion. We agree that a visualization of the difference is still a useful first step in model-data comparison as long as it does not lead to a rushed judgement on

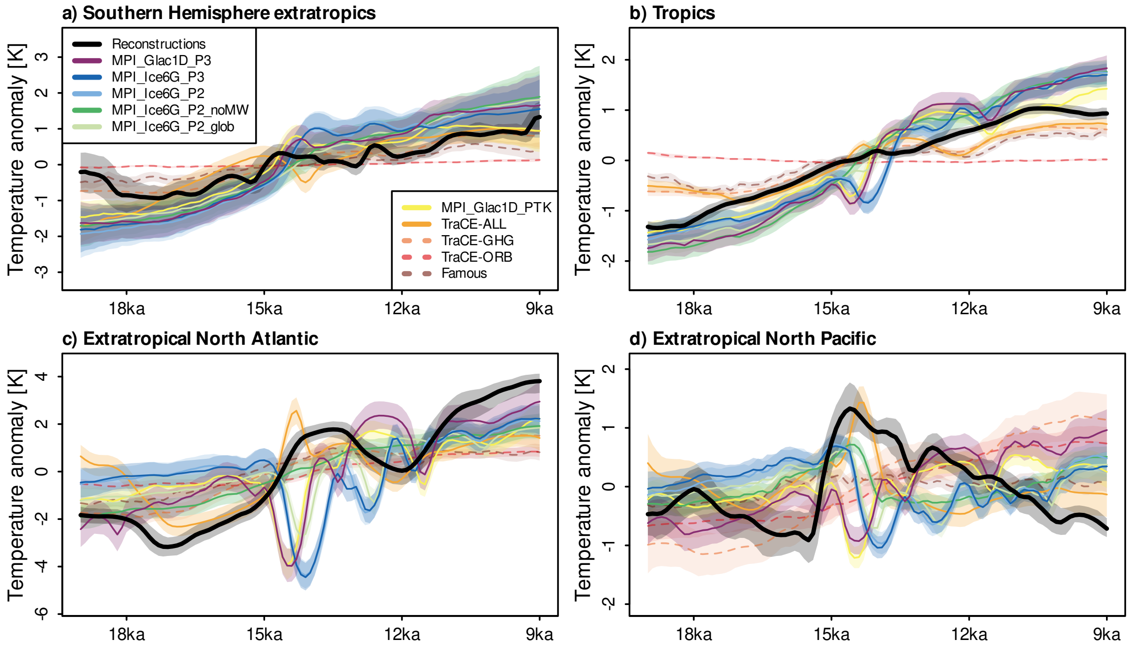

Figure 1: Regionally stacked SST variations for records in (a) the Southern Hemisphere extratropics (n=10 proxy records), (b) the Tropics (n=44), (c) the extratropical North Atlantic (n=13), and (d) the extratropical North Pacific (n=7). Black lines denote the stacked reconstructions, whereas colored lines depict the stacked forward-modeled proxy time series derived from the ten transient simulations. Shaded areas show uncertainties from chronologies and the PSM. Note that the stacks are not used in the model-data comparison algorithm, but just provide a visual impression of the reconstructed and simulated regional temporal evolution. The methodology to construct the stacks will be described in the supplemental information of the revised manuscript.

the model-proxy agreement. We did not include it in the original manuscript as the focus of the manuscript is the method development and not an all-encompassing assessment of the model-proxy agreement of the ten simulations. Nevertheless, we understand that a figure as described by the reviewer can help the reader get a better impression of the data that we use. Therefore, we created regional stacks for the four disjunct regions (Southern Hemisphere extratropics, Tropics, extratropical North Atlantic, extratropical North Pacific) of reconstructed and simulated temperatures. The construction of the stacks follows the same methodology as for Fig. 9, but without employing the timescale decomposition and the decomposition into magnitudes and patterns of variations (Fig. 1). We will include this figure as the new Fig. 6 in the revised manuscript and move the previous Fig. 6 to the supplemental information (see response to comment above).

I suggest removing the simulations which do not include all forcings, i.e. TraCE-ORB and TraCE-GHG, from the main analysis. For example, in Fig. 1a it is confusing to show simulations which do not include the full forcing as it gives the impression that the spread among models is even larger than it actually already is. I think it is perfectly fine to include those simulations to test the methodology, but not when it comes to the actual comparison of how well models simulate different aspects of the last deglaciation (e.g. lines 540-545).

We thank the reviewer for pointing out that the presentation of the simulation ensemble as currently done can lead to misunderstandings. It is not our intention to give the impression that all ten simulations are equally likely representations of the temperature changes during the last deglaciation. Nevertheless, we believe that there is value to also compare sensitivity experiments with proxy-based reconstructions, e.g., to understand for which regions and components the inclusion of all three major forcings (orbital, GHG, ice sheets) improves the model-proxy agreement the most and if the accelerated application of transient boundary conditions reduces the model-proxy agreement significantly. Therefore, we will revise Fig. 2, 7, 8, and 9 such that they contain clear visual separations between the seven simulations, which apply orbital, GHG, and ice sheet forcing without acceleration, and the three sensitivity experiments TraCE-ORB (only orbital), TraCE-GHG (only GHG), and FAMOUS (accelerated application of forcings, and neither change of land-sea mask nor Antarctic ice sheet). Examples of the adapted figures are provided in Fig. 2 and 3). Furthermore, we will adapt Sect. 2.1, 4.2, and 5.2 to better separate the sensitivity ex-

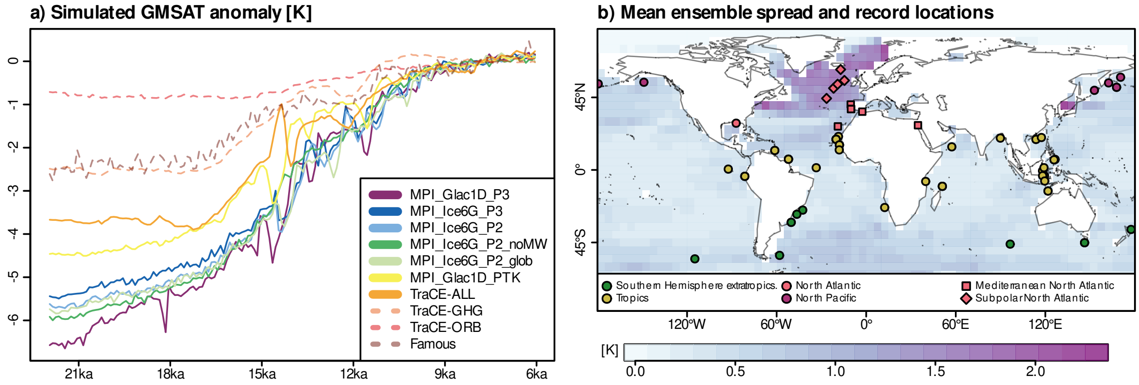

Figure 2: Revised version of Fig. 1 in the manuscript. In the revised version, the sensitivity experiments TraCE-ORB, TraCE-GHG, and FAMOUS are visually separated from the other seven simulations and the ensemble spread in (b) is only computed from the six MPI-ESM simulations and TraCE-ALL.

periments from the other seven simulations. In particular, Sect. 4.2.1 and 4.2.2 will only contain descriptions of the model-proxy agreement for the six MPI-ESM simulations and TraCE-ALL. We will add a Sect. 4.3 that assesses for which regions and components the model-proxy agreement of the three sensitivity experiments deviates the most/least from the other seven simulations.

Can the presented new methodology, which is here applied to SSTs, in principle also be extended and applied to other variables? It is mentioned that it could be extended to land temperature reconstructions, but what about very different variables like carbon or oxygen isotopes?

Yes, it is possible to extend our approach to other continuous variables (e.g. carbon or oxygen isotopes), which are simulated by Earth system models or can be linked to simulated variables through PSMs. For these variables, one can either integrate existing proxy systems models into the algorithm (e.g., BAYFOX, Malevich et al., 2019, for the oxygen isotopic composition of planktic foraminifera) or adapt the methodology from our manuscript to estimate the structure of the noise process. Employing the algorithm to proxies with binary or categorical values (e.g. occurrence of sea ice, sediment types) would require more substantial changes to the employed statistical methods. In the revised manuscript, we will add a sentence on the applicability to other continuous variables at the end of Sect. 5.2.

Some parts of the text are filled with acronyms, which makes it sometimes a bit hard to read. Please consider if the use of acronyms could be reduced, particularly also in figure captions. One example: is LD really needed?

We are sorry for reducing the readibility of the text by using many acronyms. We intended to use well-established acronyms (e.g. LGM, GHG, PSM) plus introducing some additional ones to avoid having to repeat lengthy phrases while maintaining sufficient precision of the wording. We understand that this led to some hard-to-read paragraphs. We carefully went through the list of used acronyms and will remove the acronyms LD (last deglaciation) and MC (Monte Carlo) in the revised manuscript. We will not remove the other used acronyms completely but try to replace them by alternative phrases at appropriate places.

L. 30: The references cited do not support the 3-8°C range given in the paper. From the abstracts: Tierney: -5.7 to -6.5°C; Annan: -4.5+-0.9°C.

We thank the Reviewer for correctly pointing out the inconsistency of the given range with the references. In the revised manuscript, we will correct the value to 3.6-6.5 K as the combined range of the two cited studies.

Section 2: consistently use either past or present tense

We thank the Reviewer for pointing out the inconsistent use of past and present tense in Sect. 2. The intention of mixing past and present tense was to clearly separate our choices (in present tense) and choices made during the construction of the simulations / proxy database. However, (a) this strategy was employed inconsistently by us and (b) we realize that it reduces the readability of the section and can lead to confusion. Therefore, we will use present tense consistently in Sect. 2

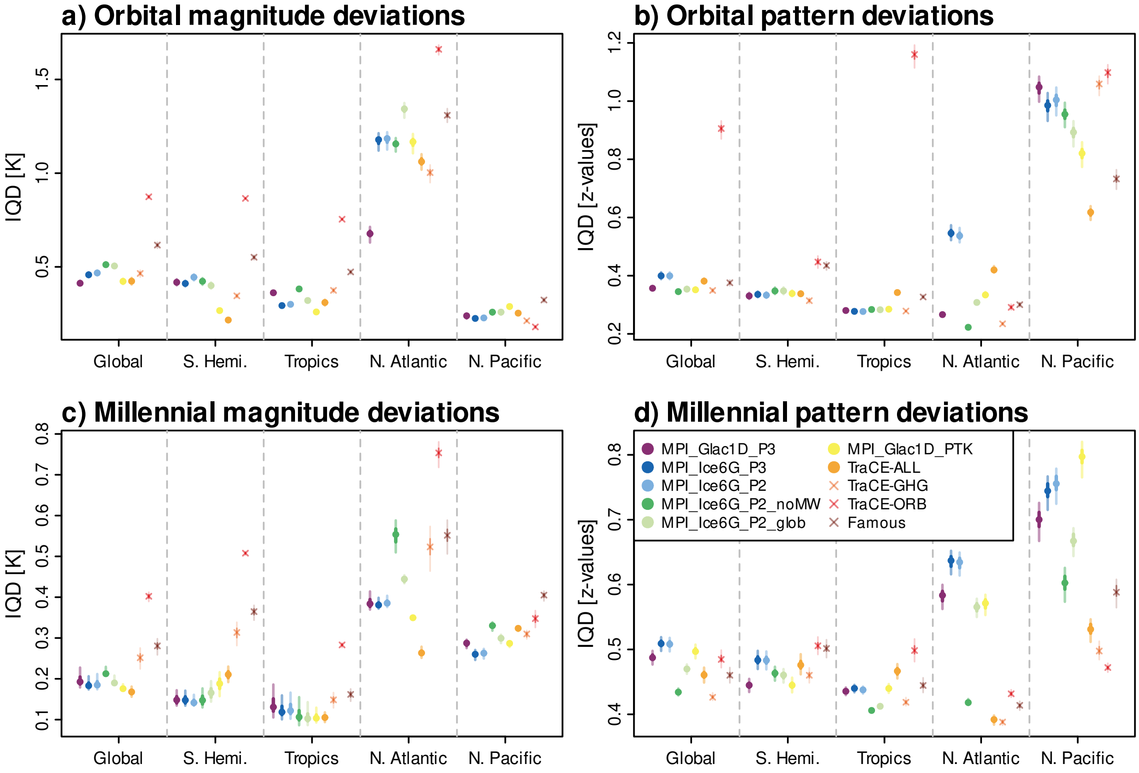

Figure 3: Revised version of Fig. 7 in the manuscript. In the revised version, the sensitivity experiments TraCE-ORB, TraCE-GHG, and FAMOUS are visually separated from the other seven simulations.

in the revised manuscript. To emphasize that all of the simulations and proxy records have been published previously, we will adapt the wording in the first sentences of Sect. 2.1 and 2.2.

L. 106: are -> is

Thanks for finding the typo. We will correct it in the revised manuscript.

L. 130-131: How is that justified? What does sub-surface mean? Is it still in the mixed layer?

It is well-established that most of the employed proxy records reflect surface or mixed layer conditions, in particular all $\mathrm{U}^{\mathrm{k}}_{37}$ and assemblage records, and the majority of Mg/Ca records (e.g., Kucera et al., 2005; Rebotim et al., 2017; Tierney and Tingley, 2018). This justifies our approach. All used sensors occupy a range of depths, which has likely changed in the past, but the environmental and biological controls of the vertical habitat variability are poorly constrained (e.g., Greco et al., 2019; Kretschmer et al., 2018). Therefore, we make the simplifying choice of comparing all records with simulated SSTs. We do not expect that this assumption has a significant effect on our results. In the revised manuscript, we will clarify that most of the proxy records reflect surface or mixed layer temperatures.

L. 396: FRPRR -> FPRR ?

Thanks for finding the typo. We will correct it in the revised manuscript.

Fig. 2 top-right, Proxy System Model: the single line representing the simulation is possibly a bit misleading, as there are 4 time series from different model grid cells that enter the PSM, if I understood correctly.

We thank the reviewer for pointing out this imprecision. The black line in the top right panel is the result from the interpolation of the simulation data to the proxy location. Thus, it is correct that time series from multiple grid boxes are combined in a weighted average in the black line. However, we believe that plotting the time series of all grid boxes used in the interpolation would not improve the reader's understanding of the algorithm as it would reduce the visual clarity of the figure. Instead, we will adapt the legend of this panel to explicitly note that the black line is the simulation interpolated to the proxy site. We believe that this is the best compromise between a detailed explanation of the steps of the algorithm and visual clarity. The legend label in the revised panel will read *"Simulation at proxy site"* and we will adapt the description of the top right panel to explain the meaning of the black line.

Fig. 2 bottom: how are the latitudinal belts defined? By the dashed vertical lines? It is not clear from the figure caption. Please repeat the text in 3.1.4.

The definition of the latitudinal bands follows the definition in Sect. 3.1.4, i.e., overlapping moving windows of 20° width which move in 5° steps. The dashed gray bars are just plotted in 30° steps to improve the visual orientation of the reader. To not lengthen the caption of the plot further, we will not repeat the text from Sect. 3.1.4 in the caption but instead refer to Sect. 3.1.4 for details on the zonal averaging procedure.

Fig. 2 Timescale decomposition panels: this is just a detail, but it would be nice and more intuitive if the lines would be plotted and shown in the legends in order of increasing 'smoothing', i.e. 1) Reconstruction, 2) Orbital+millennial, 3) Orbital.

We thank the reviewer for this suggestion which we will incorporate in the revised manuscript.

References

Greco, M., Jonkers, L., Kretschmer, K., Bijma, J., and Kucera, M.: Depth habitat of the planktonic foraminifera Neogloboquadrina pachyderma in the northern high latitudes explained by sea-ice and chlorophyll concentrations, Biogeosciences, 16, 3425–3437, https://doi.org/10.5194/bg-16-3425-2019, 2019.

Kretschmer, K., Jonkers, L., Kucera, M., and Schulz, M.: Modeling seasonal and vertical habitats of planktonic foraminifera on a global scale, Biogeosciences, 15, 4405–4429, https://doi.org/10.5194/bg-15-4405-2018, 2018.

Kucera, M., Weinelt, M., Kiefer, T., Pflaumann, U., Hayes, A., Weinelt, M., Chen, M.-T., Mix, A. C., Barrows, T. T., Cortijo, E., Duprat, J., Juggins, S., and Waelbroeck, C.: Reconstruction of sea-surface temperatures from assemblages of planktonic foraminifera: multi-technique approach based on geographically constrained calibration data sets and its application to glacial Atlantic and Pacific Oceans, Quaternary Science Reviews, 24, 951–998, https://doi.org/10.1016/j.quascirev.2004.07.014, 2005.

Malevich, S. B., Vetter, L., and Tierney, J. E.: Global Core Top Calibration of $\delta^{18}O$ in Planktic Foraminifera to Sea Surface Temperature, Paleoceanography and Paleoclimatology, 34, 1292–1315, https://doi.org/10.1029/2019PA003576, 2019.

Rebotim, A., Voelker, A. H. L., Jonkers, L., Waniek, J. J., Meggers, H., Schiebel, R., Fraile, I., Schulz, M., and Kucera, M.: Factors controlling the depth habitat of planktonic foraminifera in the subtropical eastern North Atlantic, Biogeosciences, 14, 827–859, https://doi.org/10.5194/bg-14-827-2017, 2017.

Tierney, J. E. and Tingley, M. P.: BAYSPLINE: A New Calibration for the Alkenone Paleothermometer, Paleoceanography and Paleoclimatology, 33, 281–301, https://doi.org/10.1002/2017PA003201, 2018.