Response to Reviewer 1

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.

While I am not an expert in the statistical models used within the methodology, the body of work appears sound without any immediate issues.

We thank the Reviewer for this assessment of our work.

I feel the wording justifying the need for the extra complexity of using a PSM could be improved. One of the core justifications for using the PSM is it avoids the need to rely on "sparse and uncertain proxy data", yet the PSM largely degrades or reforms the model output to be more compatible with the proxy data and then uses same sparse and uncertain proxy data to benchmark the PSM created forward-modeled proxy time series. The authors also do not address why more traditional signal processing methods, such as Principal Component Analysis, could not instead be used to extract signal from the messy proxy data without the added complexity (and caveats) of using the PSM. I don't think these criticisms undercut the work in any way, only that the justification for the need of the algorithm and PSM could be framed better.

We thank the Reviewer for the suggestions. We do already discuss advantages of using PSMs in a forward-modeling approach in different parts of the manuscript. We will try to summarize all of the advantages in the Introduction of the revised manuscript. With respect to the quote included in the comment of the Reviewer, we note that the full quote about *"sparse and uncertain proxy data"* is that *"The use of proxy forward modeling avoids the need to reconstruct gridded or regional mean temperatures from sparse and uncertain proxy data"*. This statement was supposed to say that proxy forward modeling avoids the reconstruction of gridded fields or regional mean time series, not that it avoids using sparse and uncertain proxy data. Of course, any model-data comparison method using paleoclimate proxies has to handle the sparse and uncertain nature of this type of data. In the revised manuscript, we will rephrase the statement at hand to avoid misunderstandings.

The main advantages of using PSMs are that, by accounting for non-climatic processes, (1) they allow to correct for potential biases in (timescale-dependent) temperature variability estimates from proxy data (e.g., Laepple and Huybers, 2014) and (2) they can help to distinguish patterns of temperature variations from non-climatic variations in the proxy time series. Using PSMs in a forward-modeling approach, as the algorithm described in Sect. 3.1 does, avoids the need to interpolate reconstructed temperature time series to a common time axis or spatial grid from the sparse and non-uniformly distributed records. While PSMs can be inverted using Bayesian statistics to reconstruct temperature time series with a regular time axis, this inversion is computationally expensive and requires the definition of prior distributions for the temperature time series which are hard to estimate from proxy data. Therefore, fewer assumptions and computational resources are needed in the forward approach. In the revised manuscript, we will add the incorporation of the proxy understanding as an advantage of using PSMs in the abstract and conclusions.

We have not addressed problems with more established signal processing methods because, to our knowledge, none of them has been used previously for quantitative spatio-temporal comparisons of proxies and simulations for the last deglaciation. This would have led to an arbitrary choice of methods to compare our methodology to. Most traditional signal extraction methods, in particular principal component analysis, are not designed to extract signals from irregularly spaced time series with varying time axes and coverage periods. To apply a PCA, all the data needs to be interpolated to the same time axis, which requires similar temporal resolutions of the proxy records and a common coverage interval. Additionally, (1) applying PCAs to reconstructions and simulations still requires the development of a comparison metric between the resulting spatial pattern / time series, which needs to account for uncertainties in the PCAs, (2) PCAs have stricter assumptions on the distributions of the non-climatic noise than PSMs which can lead to biases if these are violated, and (3) results of a PCA depend on an a priori defined region of interest while it is possible in our method to a posteriori assess arbitrary regions. Due to the absence of previous examples that employ standard signal processing methods for spatio-temporal model-data comparison for the last deglaciation, we prefer to focus on better explaining the advantages of using proxy-forward modeling in the revised manuscript instead of discussing why more standard signal processing methods are less suited for this task.

While less text is dedicated to evaluating the model simulations against data, I disagree with one of the key findings: "Comparing the MPI-ESM and CCSM3 simulations that employ orbital, GHG, and ice sheet forcing, we find no systematic differences between the two climate models. In particular, TraCE-ALL is mostly within the IQD spread of the six MPI-ESM simulations." This is an important statement, so the language should be more precises. What specifically are the authors referring to? There is only one MPI-ESM simulation that employs orbital, GHG, and ice sheet forcing (MPI_Ice6G_P2_glob), and there is no comparable TraCE simulation since TraCE-GHG and TraCE-ORB fix the ice sheet forcing to LGM (see Table 1). When I look at Figure 7, I don't see TraCE-ALL effectively being the same as MPI with freshwater flux. Especially in the North Atlantic and North Pacific. [...] The TraCE-ALL and hosed MPI IQDs in the Figure 9 legend are largely not similar, no less TraCE-ALL bracketed by MPI. It is often difficult and nuanced to say when one model is performing better than another, but I don't think the analysis and figures here support the claim TraCE-ALL and hosed MPI are effectively the same when compared to data.

We apologize for the imprecise language in the paragraph mentioned by the Reviewer. The statement aimed at a comparison of the IQDs of TraCE-ALL with the IQDs of the six MPI-ESM simulations (which all employ orbital, GHG, and ice sheet forcing, but inject freshwater into the ocean in different ways) across regions and components of the deglacial temperature evolution. We did not want to make the statement that TraCE-ALL and the MPI-ESM simulations are indistinguishable. Instead, we wanted to say that there is currently no clear evidence to identify differences in the ability of the employed models, MPI-ESM and CCSM3, to simulate the last deglaciation, because the experiment protocols (in particular regarding the location, timing, and magnitude of freshwater injections) are too different to attribute results clearly to either model differences or differences in the experiment protocol. We will reformulate Sect. 5.2 to improve the messaging.

To compare the IQDs of TraCE-ALL and the six MPI-ESM simulations across regions and components of the deglacial temperature evolution in a more quantitative way, we now compute rank statistics that we will use in the revised manuscript (see Fig. 1). These are computed for a reduced ensemble, which excludes TraCE-ORB, TraCE-GHG and FAMOUS. Rankings are computed for each proxy record and each of the four components. The average (over all records and components) of the simulations is between 3.8 (for MPI_Glac1D_PTK and MPI_Ice6G_P2_glob) and 4.2 (for MPI_Glac1D_P3) with TraCE-ALL having an average rank of 4.0. This shows that we cannot find a systematically superior performance for any of the seven simulations across regions and components, despite the different experimental protocols and employed models. Meanwhile we find a difference when comparing the rank statistics in more detail. The ranks of TraCE-ALL are concentrated at 1 (highest agreement) and 7 (lowest agreement) whereas the MPI-ESM simulations have much flatter rank histograms. This suggests that TraCE-ALL is more often outside than inside the range of the MPI-ESM simulations even though it does not feature a systematically higher or lower rank. This is in contrast to our previous statement of *"TraCE-ALL is mostly within the IQD spread of the six MPI-ESM simulations"* which we will remove from the revised manuscript. This concentration of TraCE-ALL at extreme ranks tends to hold for all four components (see Fig. 2 - Fig. 5). At the moment, we cannot determine whether these differences originate from the varying experiment protocols or model formulations. We will include the rank statistics in the supplemental information of the revised manuscript and refer to them in the revised Sect. 5.2.

What does it mean that MPI_Ice6G_P2_noMW outperforms TraCE-ALL in the North Atlantic for orbital pattern, when TraCE-ALL is specifically designed to reproduce the reconstructed AMOC variability? AMOC variability is sub-orbital scale of course, but what does it mean that a model without hosing is capturing that scale of variability better than TraCE-ALL (or conversely, the addition of hosing degrades the orbital-scale performance)? Likewise, what are the implications of TraCE-ALL and TraCE-GHG having nearly identical millennial pattern deviations in the North

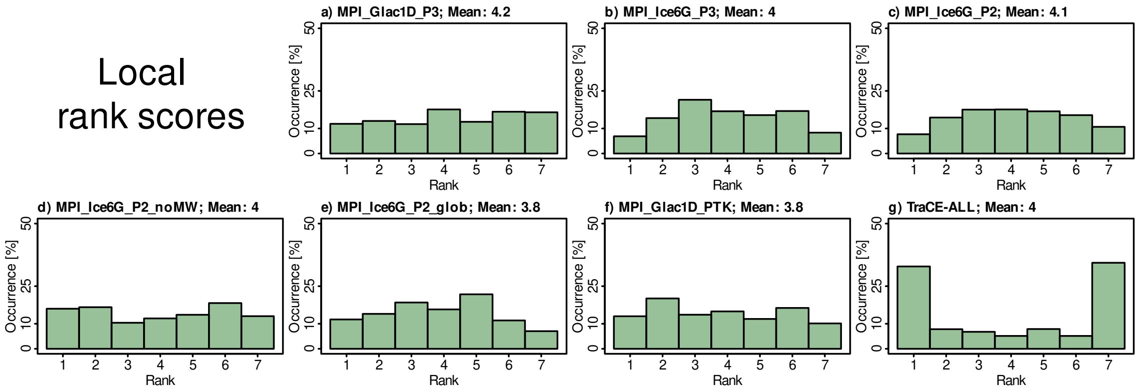

Figure 1: Rank scores for the ensemble of the six MPI-ESM simulations and TraCE-ALL. Rankings are computed for each proxy record and each of the four components. The bars depict the occurrence percentages of the ranks.

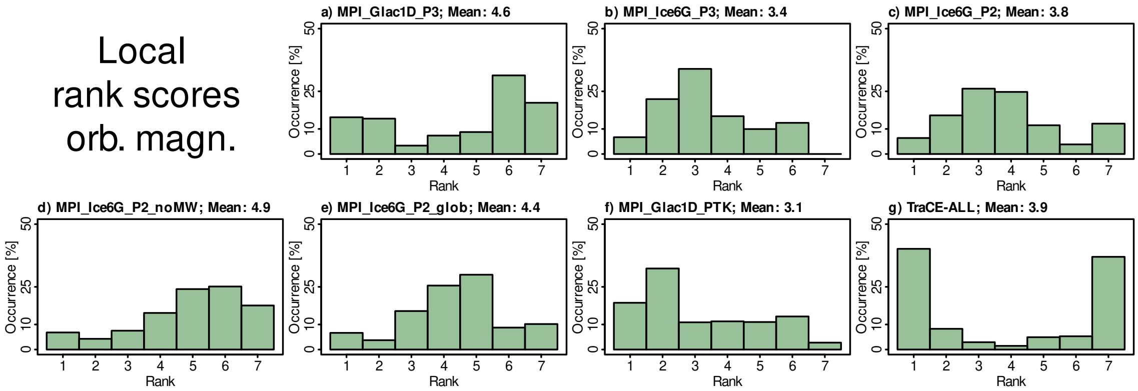

Figure 2: As Fig. 1, but restricted to orbital magnitude IQDs.

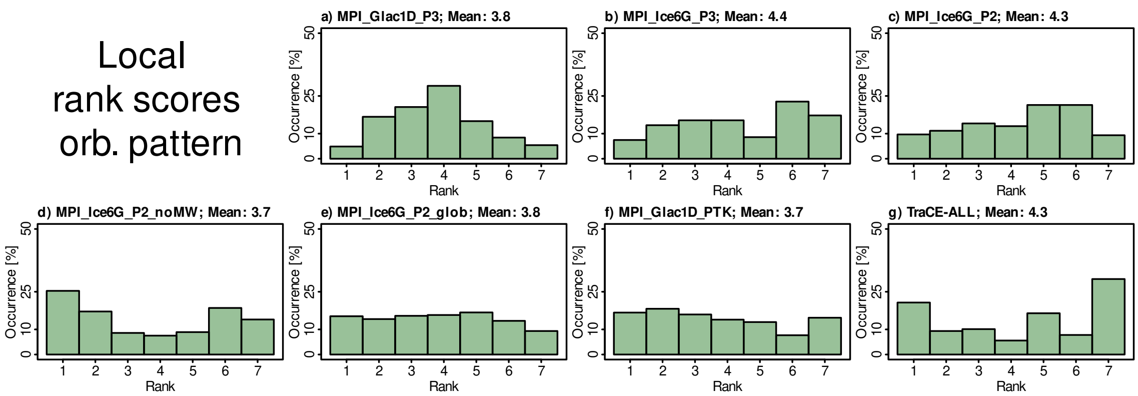

Figure 3: As Fig. 1, but restricted to orbital pattern IQDs.

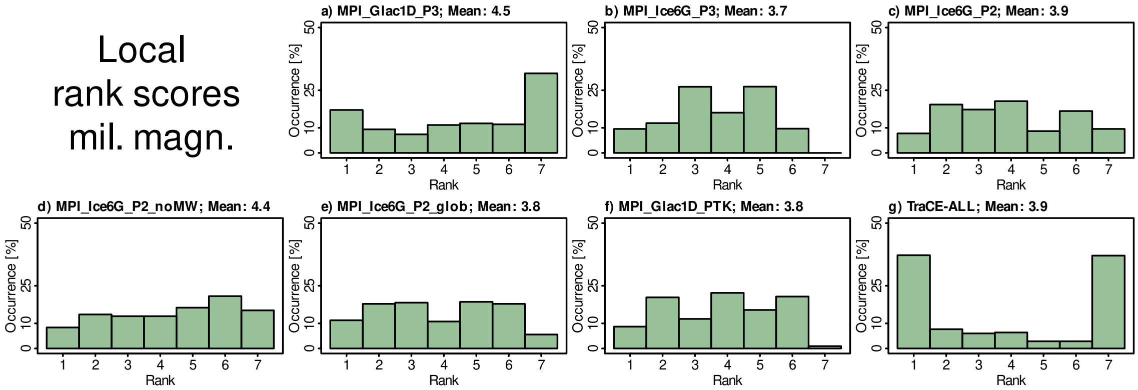

Figure 4: As Fig. 1, but restricted to millennial magnitude IQDs.

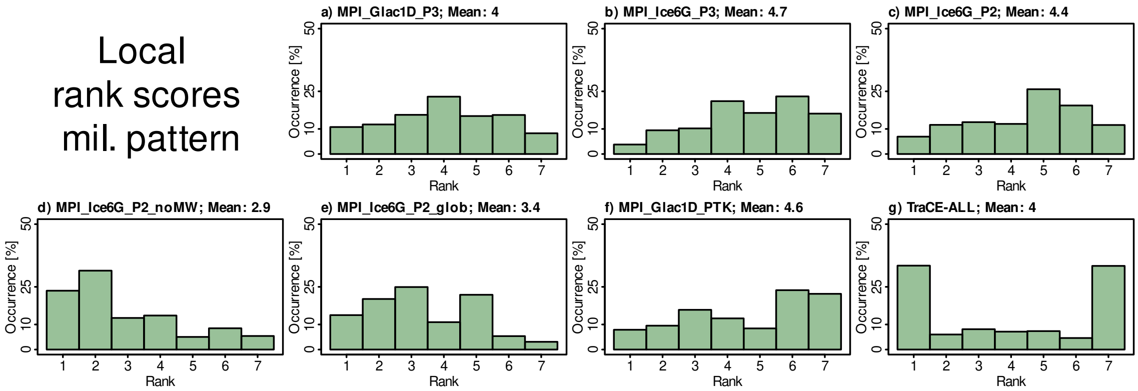

Figure 5: As Fig. 1, but restricted to millennial pattern IQDs.

Atlantic, when the TraCE-GHG doesn't include freshwater hosing?

As we explain in Sect. 4.2 and 5.2, these results originate from regional differences between the Mediterranean North Atlantic and the Subpolar North Atlantic in the reconstructions, which are not present to the same degree in the simulations. TraCE-ALL agrees the most with the reconstructions in the Mediterranean North Atlantic but its performance in the whole North Atlantic is degraded by poor agreement in the Subpolar North Atlantic. When combining magnitude and pattern metrics in biplots (see Fig. 7 and response to comment below), simulations with local freshwater injection perform the best in the North Atlantic for either timescale, MPI_Glac1D_P3 for orbital timescales and TraCE-ALL for millennial timescales.

When focusing purely on the patterns, the results mentioned by the Reviewer have several potential implications. First, a good agreement for orbital timescales does not imply a good agreement for millennial timescales and vice versa. This could be due to varying importance of forcings and internal feedback processes for the patterns of temperature changes on different temporal and spatial scales. Second, reproducing the patterns of a small set of proxies might be an insufficient strategy to capture the spatial structure of millennial-scale temperature patterns. For example, our results suggest that reproducing the patterns of a specific AMOC proxy (e.g., Pa/Th ratios at Bermuda rise), as TraCE-ALL does, will not necessarily lead to a good model-proxy agreement for millennial-scale temperature patterns across different regions. Instead, other factors such as the magnitude of the AMOC response, which cannot be directly quantified from comparing Pa/Th ratios with simulated AMOC strengths, or the background climatic state could potentially have a large influence on the regional manifestations of the temperature variability patterns. Third, the uncertainty about the factors controlling millennial-scale variability could lead to an adequate reproduction of the pattern of AMOC variability with an incorrect mechanism. In this case, the spatially varying degree of model-proxy agreement could be the result of an incorrect driver of millenial-scale variability. The uncertainties of ice sheet and freshwater flux reconstructions, and the heterogeneity of the experiment protocols make it very challenging to determine the reasons for the spatially varying performance. Addressing these challenges with designated protocols in the context of inter-model comparison projects could be a promising way forward. In the revised manuscript, we will enhance the discussion of potential implications of the varying model-proxy agreements on different temporal and spatial scales, and give more suggestions for future research to address the identified challenges.

The text notes "More generally, all simulations with meltwater input show a better agreement with reconstructions for millennial magnitudes than those without meltwater input." I don't think this is strictly true. There are cases where MPI_Ice6G_P2_noMW performs similar to, if not better than, the routed MPI-ESM simulations. In either case, this only means the millennial-scale variability is more like the data when hosing is added, not that the pattern is realistic (as noted around line 550). This is more apparent with the TraCE simulations where in some locations the addition of hosing degrades model performance. Getting the magnitude of variability correct, but the patterns (ie trends) of the deglaciation wrong isn't particularly satisfying, which could be emphasized here.

We apologize for the imprecise statement in l. 546 of the original manuscript. We aimed at referring strictly to the global averages in this statement. Indeed, MPI_Ice6G_P2_noMW performs similar or better than the simulations with local freshwater input in some regions. We will replace the sentence starting with *"More generally..."* by *"In particular, in the global average, all simulations with meltwater input show a better agreement with reconstructions for millennial magnitudes than those without meltwater input."*

We agree with the Reviewer's statement that this only implies that *"millennial-scale variability is more like the data when hosing is added, not that the pattern is realistic"*. As we state in l. 552ff, we can currently not determine the reason for the absence of pattern improvement when including local freshwater injections. Following the advice of the Reviewer, we will add a sentence in the same paragraph to emphasize this contrast between magnitudes and patterns more strongly.

Line 125: define or give examples of "sensor" for the novice.

Following the Reviewer's suggestion, we will define sensors as the organisms recording the temperature signal (e.g., planktonic foraminifera) and proxies as the measured temperature-sensitive quantities (e.g., Mg/Ca ratios, species compositions) in the revised manuscript. This definition will be given in the Introduction, when sensors and proxies are first mentioned.

Line 137: Osman, et al., 2021 uses four proxy types, so I am not sure why it cited here.

We apologize for the imprecise language in the sentence. Both cited studies use multiple, but distinct sensors (Paul et al. (2021) uses only faunal and floral assemblages whereas Osman et al. (2021) uses only geochemical proxies, i.e., Mg/Ca, U^{k}_{37}, TEX_{86} and $\delta^{18}O_c$). We included the half sentence to emphasize that we use both, assemblage-based and geochemical proxies. In the revised manuscript, we will correct the sentence by stating that Paul et al. (2021) focuses on assemblage-based temperature reconstructions while Osman et al. (2021) focuses on geochemical proxies.

Lines 177: "Computing averages in this last step instead of averaging temperature time series in the beginning avoids interpolating proxy records with irregular time axes to a common resolution." Explain this. It seems like in some portions of the analysis the data are binned to a fixed 100-yr timestep (ie Section 3.2). The time series displayed in Figure 9 are regional averages, are these first binned to 100-yr interval or are they somehow calculated on irregular time spacing for multiple records and ensemble members?

The Reviewer is correct that the used data is interpolated to a regular time axis with 100 yr time steps in Sect. 3.2 (Estimation of proxy system model parameters) and to construct the regional stacks in Fig. 9. We want to note that we consider neither the estimation of the proxy system parameters nor the construction of the regional mean stacks as part of the model-data comparison algorithm. Only the four steps described in Sect. 3.1 are part of the algorithm. Therefore the statement in line 177 refers just to the procedure described in those four steps and motivates the averaging of IQDs obtained for individual records instead of averaging temperature records before computing the deviations between reconstructions and simulations.

The interpolation of the reconstructions to a common time axis in the estimation of the PSM is suboptimal. As the core topic of the paper is the development and testing of the new algorithm, additionally developing a more sophisticated method, which avoids interpolation, to estimate the PSM parameters is beyond the scope of our manuscript. We will mention this limitation explicitly in the discussion (Sect. 5.2) of the revised manuscript and recommend the development of more advanced methods to estimate the temporal structure of the non-climatic noise process in addition to the integration of process-based PSMs.

Fig. 9 only serves illustrative purposes to aid the interpretation of the quantitative model-data comparison results. The IQD values in the legends of the respective panels are the averaged IQDs of the records included in the construction of the stacks, but they are not computed by comparing the reconstructed and simulated stacks quantitatively. We are sorry for not explicitly noting this in the previous manuscript version and will adjust the caption of Fig. 9 to emphasize the role of the regional stacks.

Lines 220: magnitude is defined as the standard deviation of each ensemble member (for the decomposed time series). Since standard deviation is an absolute value, doesn't this fail to discriminate between trends in opposite directions (ie data is cooling when model is warming)?

Indeed, the magnitude metric does not discriminate between trends in opposite directions but only measures the strength of variations. However, opposing trends lead to very high pattern IQDs (e.g., orbital pattern IQDs for TraCE-ORB in the Southern Hemisphere) as opposing trends create large deviations of the 'pattern' time series. To evaluate strength and direction of trends together, one needs to jointly consider magnitude and pattern IQDs. As the Reviewer suggests below this can be visualized by using biplots of the two metrics in which simulations that match the reconstructed strength and direction of trends appear in the lower left (see Fig. 7). We note that trends during the deglaciation are mostly non-linear. Therefore, reducing the orbital signal purely to the strength and direction of trends does not seem sensible. This motivates our separation into magnitudes and patterns. In the revised manuscript, we will try to better explain the separation into magnitudes and patterns in Sect. 3.1. Additionally, we will emphasize that a single metric is likely insufficient for fully capturing the deviations between simulations and reconstructions in an interpretable way. The use of biplots as suggested by the Reviewer below will be mentioned explicitly as a way to jointly assess magnitudes and directions (or, more generally, patterns) of trends.

Throughout the text "magnitude" is used to denote the degree of variability in the decomposed time series, which is just saying the strength of variability. It may not be obvious to the reader what the utility of this metric is. We tend to think in terms of time series, so "pattern" (as defined here) is far more intuitive.

Indeed, we use the term 'magnitude' to denote the strength of variability. We do somewhat disagree with the notion that patterns are *"far more intuitive"* than magnitudes. Patterns are mostly meaningful if variations are externally forced and if there are sufficiently tight constraints on the boundary condition reconstructions such that models can be expected to reproduce the observed patterns of variation. As we discuss in l. 583-589 of the original manuscript, whether these two conditions are fulfilled, is debated. Models forced with local freshwater injections computed from established ice sheet reconstructions do not reproduce the observed millennial-scale patterns in the North Atlantic (Kapsch et al., 2022) while another model (MIROC) produced a spontaneous abrupt warming similar to the transition into the Bolling-Allerod (Obase and Abe-Ouchi, 2019). Therefore, we think that comparing patterns of time series alone is an insufficient metric to comprehensively evaluate simulations of the last deglaciation. Instead, we attempt to combine this pattern approach with methods developed for analysing unforced variability in simulations and proxies for e.g. the LGM, the Holocene, and the last millennium (e.g., Laepple and Huybers, 2014; PAGES 2k Consortium, 2019; Rehfeld et al., 2018). In addition, comparing magnitudes of variations is important for assessing the strength of feedback processes, which in turn can be relevant to constrain the projected response to future emission scenarios or the impact of potential tipping points in the Earth system. We will try to better motivate the separation into magnitudes and patterns in Sect. 3.1 of the revised manuscript.

Lines 235: Since N is either 100 or 1000, I assume an empirical probability distribution is used rather than a fitted distribution.

Yes, we are using empirical distributions to approximate the analytically intractable distributions P and Q. We note this in l. 244 of the original manuscript, when stating *"We compute the IQD using a MC approximation of Equ. (2) with the MC samples from step 2."*. We apologize that this sentence alone has not clarified that we use empirical distributions. In the revised manuscript, we will rearrange the paragraphs such that the Monte Carlo approximation is mentioned immediately after explaining the different terms in Equ. (2). In addition, we will add a sentence to state that we approximate P and Q by empirical distributions.

Line 244: How does IQD integrate differences in time series? Each forward-modeled proxy time series is on the same irregular age model spacing as the proxy data, but how is the time series translated to distributions used in the IQD equation?

In the Monte Carlo approach, we obtain N pairs of realizations of the reconstructed and forward-modeled proxy time series. Each realization has M samples and each of the M samples corresponds to one depth in the proxy record. Thus, the realizations can be interpreted as an empirical, M-dimensional probability distribution with correlations between the samples due to auto-correlation in the time series. The IQD computes the difference between these M-dimensional distributions by using a Monte Carlo approximation of Equ. (2), which we describe in more detail in the supplemental information. In the revised manuscript, we will try to better describe this translation between time series and multivariate probability distributions in Sect. 3.1.2 and 3.1.3.

Lines 254: Zonal IQD seems to only be used in part of Figure 2, which is a flowchart of the analysis. If it is not used in the results section, could it be removed?

As the Reviewer correctly points out, we do not use zonal IQDs in Sect. 4.2. This is due to differences between the North Atlantic and North Pacific detected in an initial visual inspection of the reconstructions and simulations. However, we do use zonal mean IQDs in the pseudo-proxy experiments of Sect. 4.1. The evaluation of the pseudo-proxy experiments is meant to assess the general properties of the algorithm, whereas the regions chosen in Sect. 4.2 are more targeted towards the specific data at hand. As we show in Sect. 4.1, the spatial aggregation is an important step to improve the reliability and robustness of the results, independent of the specific partitioning into regions. Therefore, we think that the visualization of this step should not be removed in Fig. 2. As zonal averages are an easy to understand and often-used separation into regions, we decided to use these in Fig. 2 and Sect. 4.1. and do not think that a change to different regions in Fig. 2 is needed to improve the understanding of the steps of the algorithm. In the revised manuscript, we will mention that this is just an exemplary partition into regions in the cpation of Fig. 2.

Section 4.2 Comparison of simulations against SST reconstructions: I think it would be really useful to the reader to plot orbital + millennial time series for the regions summarized in Figure 7 (ie this new plot should come before Figure 7). I envision something like Figure 9, which would give the reader a feel for what the models are simulating (relative to the data) prior to the decomposition. Those raw trends are somewhat abstracted away by plotting orbital and millennial time series separately. For example, in Figure 9 MPI_Ice6G_P3 has a cooling trend around 14 – 13 ka in both the orbital and millennial scales. If it is caused by the injection of freshwater forcing, I would expect it to only be in the millennial-scale (also perhaps implying freshwater forcing is showing up in the orbital-scale decomposition).

The Reviewer suggests an additional visualization of reconstructions and simulations prior to the quantitative model-data comparison. We did not include a visualization as suggested by the reviewer 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. Nevertheless, we understand that a figure as described by the Reviewer can help the readers to 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 forward-modeled temperature time series. The procedure is similar to Fig. 9, but for the four disjunct regions from Fig. 7 and before employing the timescale decomposition and the decomposition into magnitudes and patterns of variations (Fig. 6). 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 comments of Reviewer 2). As noted by the Reviewer, strong freshwater-water induced perturbations can have an imprint on the orbital-scale signal, albeit weaker than in the millennial-scale signal. This is the case when the perturbations are large enough to substantially influence time-averages on timescales longer than the smoothing frequency used to separate orbital-scale and millennial-scale variations. The cooling trend around 14 ka in MPI_Ice6G_P3 is one example of this phenomenon. We will add a sentence in Sect. 5.2 of the revised manuscript to discuss this interaction between orbital- and millennial timescales.

Figure 7: It would be very verbose, but would it be worth plotting a magnitude versus pattern IQD scatter plot? There are too many combinations for the main text, so perhaps an example (perhaps millennial magnitude versus pattern for the North Atlantic)? The best models should converge in the lower left of the plot near the plot origin (0,0).

We thank the Reviewer for this very insightful visualization idea. In particular since magnitude and pattern scores provide different information on the model-proxy agreement, combining these two aspects in one plot can be a helpful visualization. As the Reviewer correctly notes, the best performing models will be located in the lower left corner. We have decided to not include this figure in the main text as it does not add a new information compared to Fig. 7. However, we

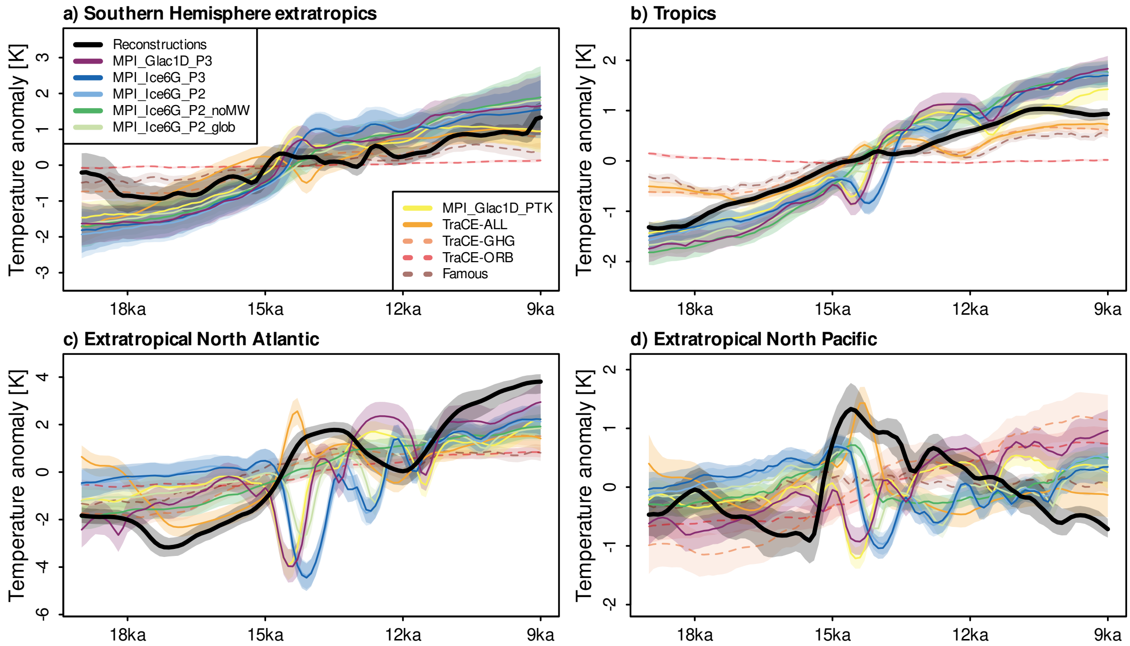

Figure 6: Regionally stacked SST variations for records in (a) the Southern Hemisphere extratropics (n=10 proxy records), (b) the Tropics (n=44), (c) the extratropical North Atlantic (n=13), and (d) the extratropical North Pacific (n=7). Black lines denote the stacked reconstructions, whereas colored 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.

will include it as supplemental figure for globally and regionally averaged IQDs and reference it in Sect. 4.2. We include one version with all ten simulations and one version without the three 'sensitivity experiments', which either change just one boundary condition transiently (TraCE-ORB, TraCE-GHG) or apply boundary condition changes with a 10x acceleration (FAMOUS) (see response to Reviewer 2 for more information on this separation in the revised manuscript) (Fig. 7 and Fig. 8). The latter plot is meant to improve the visualization of differences between the remaining seven simulations in cases where the performance of at least one of the sensitivity experiments differs strongly from the six MPI-ESM simulations and TraCE-ALL (e.g.. TraCE-ORB for orbital variations in the Tropics).

Line 597: "To avoid the need to reconstruct gridded or regional mean temperatures from sparse and uncertain proxy data, the algorithm applies proxy system models to simulation output and quantifies the deviation between the resulting forward-modeled proxy time series and temperature reconstructions". Doesn't Figure 9 create regional stacks? I understand mean IQD is used to summarize regions (as explained in section 3.1.4), but how are time series of regional averages constructed?

As the Reviewer points out correctly, we construct regional stacks in Fig. 9. As mentioned above, these are not part of the quantitative model-data comparison but are only aiding the interpretation of the regional mean IQDs. The regional mean IQDs in Fig. 7 and in the labels of Fig. 9 are averaged IQDs from each proxy record located in the respective regions. We mention this as *"The numbers in the legends next to each simulation are the averaged IQDs over all records in the respective stacks"*. As using the word *'stacks'* in this statement might have led to confusion, we will change it to *'regions'* in the revised manuscript. As we point out in Sect. 3.1.4, averaging as the last step of the model-data comparison avoids the creation of regional mean time series / gridded reconstructions, since we only have to average numbers instead of time series. To formulate this point more clearly, we will replace *"reconstruct gridded or regional mean temperatures"* by *"reconstruct gridded fields or regional mean temperature time series"* in the revised manuscript. We will describe the construction of the regional stack in the supplemental information of the revised manuscript.

The regional stacks are not used to compute the IQDs in the labels of Fig. 9.

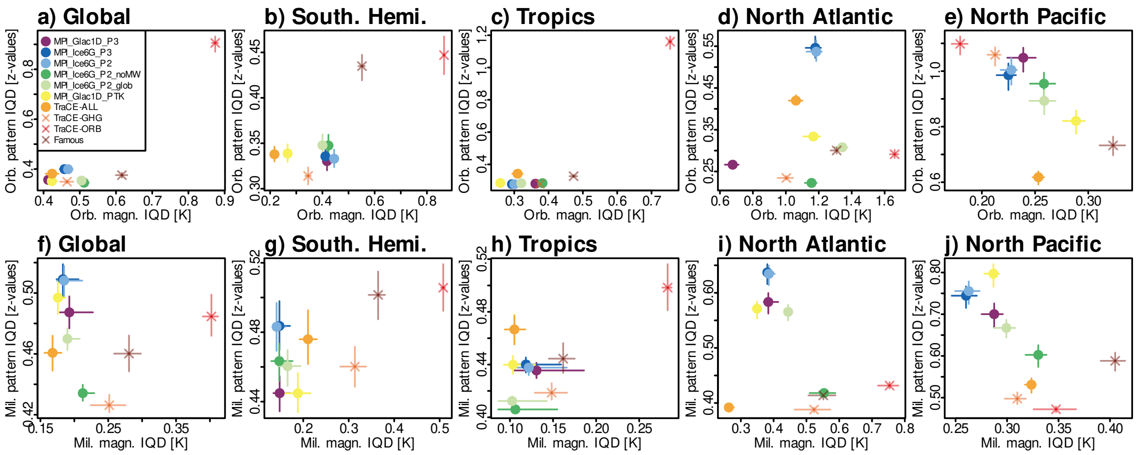

Figure 7: Biplots of IQDs for orbital-scale and millennial-scale variations. The magnitude IQDs of variations are plotted on the x-axes and the pattern IQDs on the y-axes. Lines indicate uncertainties from varying the PSM parameters. Dots in the lower left corner indicate simulations with the highest model-proxy agreement for magnitudes and patterns.

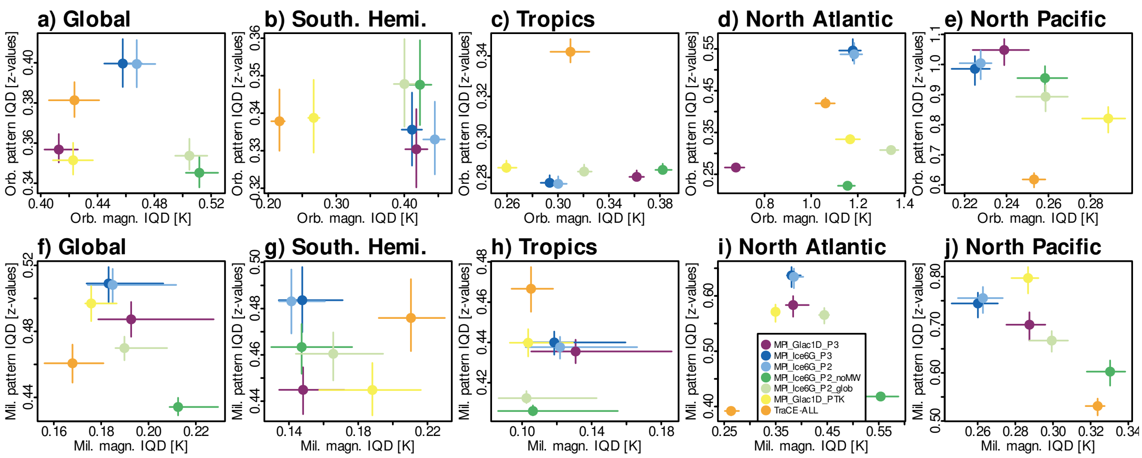

Figure 8: As Fig. 7 but without the three sensitivity experiments TraCE-ORB, TraCE-GHG, and FAMOUS.

Figure S2. Many of the plot titles in Figure S2 are identical. I assume this is depicting multiple records from the same core site. Perhaps this could be denoted better in the plot titles or figure caption. Also, how are the regional stack time series in Figure 9 made when not all records in Figure S2 span 19 – 9 ka?

We thank the Reviewer for this observation. The assumption is correct that the identical plot titles occurred in the case of records belonging to the same core. In the revised Fig. S2, we will include the IDs of each record from Table 2 in the plot titles, which identify each record uniquely from the combination of core name and sensor.

For the regional stacks, we are considering records only for the periods that they cover, which can differ depending on the selected age ensemble member. This can lead to artefacts, mostly at the edges of the period 19-9 ka. To reduce these artefacts, we restrict the coverage period of the stacks to 19-9 ka whereas the model-data comparison algorithm includes all samples in the interval 22-6 ka. This is not a major issue because the stacks are not used to compute the IQDs but just support the understanding of the model-data comparison results. We apologize for not describing the methodology of computing the stacks before, because they are not part of the model-data comparison methodology. In the revised manuscript, we note explicitly in the figure caption that the stacks are not used to compute the IQDs and describe their construction methodology in the supplemental information.

References

Kapsch, M., Mikolajewicz, U., Ziemen, F., and Schannwell, C.: Ocean Response in Transient Simulations of the Last Deglaciation Dominated by Underlying Ice-Sheet Reconstruction and Method of Meltwater Distribution, Geophys. Res. Lett., 49, https://doi.org/10.1029/2021GL096767, 2022.

Laepple, T. and Huybers, P.: Ocean surface temperature variability: Large model–data differences at decadal and longer periods, Proc. Natl. Acad. Sci. U.S.A., 111, 16 682–16 687, https://doi.org/10.1073/pnas.1412077111, 2014.

Obase, T. and Abe-Ouchi, A.: Abrupt Bølling-Allerød Warming Simulated under Gradual Forcing of the Last Deglaciation, Geophys. Res. Lett., 46, 11 397–11 405, https://doi.org/10.1029/2019GL084675, 2019.

Osman, M. B., Tierney, J. E., Zhu, J., Tardif, R., Hakim, G. J., King, J., and Poulsen, C. J.: Globally resolved surface temperatures since the Last Glacial Maximum, Nature, 599, 239–244, https://doi.org/10.1038/s41586-021-03984-4, 2021.

PAGES 2k Consortium: Consistent multidecadal variability in global temperature reconstructions and simulations over the Common Era, Nat. Geosci., 12, 643–649, https://doi.org/10.1038/s41561-019-0400-0, 2019.

Paul, A., Mulitza, S., Stein, R., and Werner, M.: A global climatology of the ocean surface during the Last Glacial Maximum mapped on a regular grid (GLOMAP), Clim. Past, 17, 805–824, https://doi.org/10.5194/cp-17-805-2021, 2021.

Rehfeld, K., Münch, T., Ho, S. L., and Laepple, T.: Global patterns of declining temperature variability from the Last Glacial Maximum to the Holocene, Nature, 554, 356–359, https://doi.org/10.1038/nature25454, 2018.