Reply to reviewer 2

We thank the referee for the detailed and constructive criticism and the appreciation shown for our work. In the following, we address the comments on a point-by-point basis. We also indicate the changes we will implement in our manuscript upon revision. In summary, in response to the reviews 1 and 2 we will

- introduce our goals as key hypotheses to frame the manuscript
- reword individual sentences and shorten text in Methods and Results.
- remove Fig. 11 from the manuscript, include melt-water forcing in an additional panel in Fig. 1, rearrange Fig. 12 and modify labels in Figs. 2 and 3.

We are confident this will improve the scientific relevance, impact and readability of the study.

Response to scientific comments

Referee: "The manuscript presents a detailed comparison of transient model simulations since the last Ice Age, using a hierarchy of physics-based 2D or 3D models, from simple energy balance models to Earth system models. The work is rigorous, well-motivated and well-written, and I am generally supportive of publication once a few key issues have been addressed."
Reply: *We thank the referee for this positive assessment of our work and address the issues raised in the following.*
Anticipated changes: None.

Referee: "1.1 Death by moments
I appreciate the authors' effort in setting up a general framework by which to analyze this ensemble of simulations. However, having to compare 4 moments across 15 simulations makes for a very large volume of information, bordering on overwhelming (witness the 41 figures in the supplement). This would be warranted if there was significant dynamical/physical insight to be gained from this assessment, but I found the paper lacking in this regard. For instance, what do changes in kurtosis (of temperature or precipitation) tell us about? Can this be tied to particular dynamics (e.g. convective clouds vs stratiform clouds produce precipitation distributions that can be distinguished by these moments). If not, is this really useful? In many places the insistence on painstakingly documenting the minute details of all four moments in these models makes for very bulky prose, from which this reader derived very few insights (e.g. L543-555). A statement like "The chosen ice sheet reconstruction has a limited impact on temperature kurtosis on all timescales analyzed here" does not help explain any dynamics. It also seems (e.g. L503) that it is broadly insensitive to many factors, so perhaps it is not a very useful indicator of anything? If so, why fill the paper with it?
I recommend using higher-order moments only when they can be connected to identifiable processes, and/or if they can be constrained by paleoclimate observations (proxies). Otherwise, this reads like a gratuitous exercise that multiplies figures without simultaneously enriching the content."

Reply: *We thank the referee for the comment and share the concern about the volume of information provided. We attempted to separate significant from non-significant changes, providing the latter only in the supplement as a point of reference. We acknowledge that this separation could be more strictly implemented.Rigorously studying how distributions of temperature and precipitation change on the long-term has never been attempted. We therefore chose to provide comprehensive results of the analysis, even those that are inconclusive or show no difference with timescale, state or forcing. This is to provide a basis for future analyses, exploration and exploitation. As these*

plots are only part of the supplement, we think keeping them is acceptable. For the main text, we will further narrow in on the relevant changes as suggested by the referee. To this end, we will move Fig. 11 (kurtosis of precipitation) into the supplement (and narrow in on the timescale-dependent changes by combining it with Fig. S24 and S25 and showing this for only one model). We will further tighten the text and reduce the level of detail provided.
Considering the heterogeneity of the results between models and simulations, a certain amount of information is necessary to highlight relevant processes and dynamics. On interannual to centennial scales, surface temperature and precipitation are the end members of a chain of dynamical processes in the climate system. It is difficult, at this point, to confidently establish links between processes and all aspects of surface climate variability. We do identify a range of processes that could link dynamics to the distributions of surface temperature and precipitation, but there is strong model diversity. We agree that pinpointing dynamics and feedbacks is of interest. Further clarification is hopefully possible when differences in model setups, boundary conditions and forcing could become a target for the next generation of model intercomparison projects.
A more in-depth comparison of simulated to reconstructed surface climate variability would be desirable. However the coverage in time and space of proxy records is limited, especially for precipitation-related variables. The LGM reanalysis is the only field reconstruction that covers the whole time period. Uncertainties in model-data comparison on variability are large and narrowing down the role of dynamical climate processes by proxy forward modeling, single-model-perturbed-parameter experiments and coordinated protocols is future work we will highlight as important in the revision.

Anticipated changes: We will move Fig. 11 to the supplement and combine it with Fig. S24 and S25. We will remove detail in Sec. 4.3 – 4.5 to narrow in on the most notable features of the analysis.

Referee: "1.2 Chapter or Paper?
The methods' description, as well as the background, read more like a thesis chapter than a paper in a specialized journal, where some common understanding exists. Recalling the mathematical definition of the various moments (section 3.1), or of spectral quantities (section 3.2), seems overkill here. Either readers know it, or they can look up a standard stats textbook. Similarly, the opening paragraph of section 3 is appropriate for a thesis chapter, but superfluous in a paleoclimate journal, where everyone either already understands this mathematical framework, or doesn't care enough about it for this exposition to matter. I recommend stripping down this exposition to the bare minimum, particularly if the code will be made available. The authors only need describe what they did in very general terms, and interested readers can go look at the code if they want to reproduce any of the results"

Reply: *We thank the reviewer for the feedback and agree that the methods section is intentionally detailed. We observed that for those who are unfamiliar with higher-order moments this section is helpful, for those who are familiar it may not be necessary for an intuitive understanding for following the manuscript. We do think having some explanation readily available instead of relying on textbooks is beneficial. Still, we will condense the section, and move parts of it to the supplement. Sections 3.1 and 3.2 will then focus on the details vital to understanding the results.*

Anticipated changes: We will shorten sections 3.1 and 3.2 and move some details to the supplementary material.

Response to editorial comments

Referee: "L25 Insert full stop before 'rather' and start a new sentence there."
Anticipated changes: We will make this change as suggested.

Referee: "L114 "has been shown to reduce seasonal to interannual standard deviation" missing "the" before "seasonal""
Anticipated changes: *We will add the "the" as suggested.*

Referee: "L149 "normality assumptions might break down" → this runs counter to my experience, whereby on long timescales, any averaging process (which is common in climate) makes things more Gaussian, by virtue of the central limit theorem. Can the authors explain why they expect normality to break down here? "
Reply: In the stationary system considered in an equilibrium climate model simulation fluctuations will indeed, at some point, average out (Central Limit Theorem). In reality, the Earth system is not stationary, as forcings and boundary conditions change over time. As we state in lines 327/328 we assume weak stationarity of our signals after detrending. Abrupt shifts, land-surface and boundary condition changes may impact shorter than millennial timescales, and thus the shapes of the distributions, which become non-Gaussian.
Anticipated changes: We will clarify the statement to read "normality assumptions might break down under a non-stationary climate evolution".

Referee: "L173-174 : "They vary regarding simulation setup, applied forcings and model complexity." This is redundant with the previous paragraph."
Anticipated changes: We will edit this to remove the redundancy.

Referee: "L189 "present-day": please define. The present is a singularly ambiguous notion in paleoclimatology [Wolff, 2007]."
Reply: *To account for this ambiguity, we have defined present-day in the beginning of the introduction, on p. 2 footnote 1.*
Anticipated changes: None.

Referee: "L219 "2 kyr AP". Presumably "AP" stands for "after present", but since the present was not defined, this does not help. Also, this is the first time I see this acronym, so the authors should add a footnote to explain what they mean."
Reply: *AP was defined in the first footnote together with BP.*
Anticipated changes: None.

Referee: "Fig 2, caption "with respect to the past 2 kyr": Is this the reference for everything? If so, please state explicitly in the text, upstream of this caption, so there is no ambiguity"
Reply: *Whenever discussing mean changes in the models, we do indeed calculate anomalies with respect to the past 2 kyr. We will make this clear in the revised version of the manuscript.*
Anticipated changes: We will insert a statement about mean anomalies being calculated with respect to the past 2 kyr in the text and adapt the text in the figure.

Referee: "Fig 3a this is a very creative and helpful way to represent model complexity and help compare the various models used here. Most of the dimensions are qualitative, but atmospheric and oceanic resolution are two dimensions where one could be quantitative. Can the authors at least give a sense of the range of resolution spanned by the model ensemble?"
Reply: *We thank the referee for the feedback. Indeed, the dimensions regarding resolutions are*

quantitative based on horizontal and vertical resolution as outlined in Sec. S2.2 of the supplement. We agree that highlighting this in the figure would be helpful. We will update the figure accordingly, while maintaining legibility.
Anticipated changes: We will add numbers to the levels for resolution in Fig. 3 and labels for the common models resolution to place them in the hierarchy.

Referee: "L296 "and EBM" : suggest "and *the* EBM", since there is only one here"
Anticipated changes: We will make this change as suggested.

Referee: "L312 "from the end of the LGM to PI". Is PI the same as "the present"? See comments above, and please be consistent in what baseline is used for the modern era."
Reply: *We agree that a consistent terminology is important and will clarify the definition of PI used in Osman et al. (2021), which is 1000–1850 CE.*
Anticipated changes: We will change the sentence to read "The resulting reanalysis estimates a global warming of $7.0 \pm 1.0°$ C from the end of the LGM to PI (with PI defined as 1000–1850 CE), as it is contains a LGM state colder than reconstructed elsewhere (c.f. Annan et al., 2022; Tierney et al., 2020; Shakun and Carlson, 2010).".

Referee: "L317-321 superflous paragraph"
Anticipated changes: We will shorten the methods as discussed above.

Referee: "L325 : "we remove the trend from the timeseries using a Gaussian filter" → presumably this is a high-pass filter? Note that for discrete data a binomial filter is more justified."
Reply: *Indeed, we apply the Gaussian filter as a high-pass filter. We appreciate the notion of the binomial filter being useful for use with discrete data. For reasons of simplicity and the benign and conservative properties of the Gaussian filter, we apply it for the removal of the trends. As a Gaussian filter is also suitable for irregularly sampled timeseries, we hope that using it further simplifies future comparison to proxy data.*

Referee: "Sections 3.1, 3.2 tighten up so they read more like a paper and less like a dissertation chapter."
Anticipated changes: We will shorten the methods as discussed above when addressing the corresponding scientific comment.

Referee: "L398 "a frequency range of [2ts,1000]" → technically, this is a range of periods"
Reply: *We thank the referee for catching this and will change this upon revision.*
Anticipated changes: We will change "frequency range" to "period range".

Referee: "L411 "Overall, MPI-ESM r1–r7 exhibit the largest temperature difference" Can this be tied to the equilibrium climate sensitivity of the various models?"
Reply: *Unfortunately, the equilibrium climate sensitivity (ECS) is not documented for all simulations. The temperature difference depends on model tuning and ECS, but also boundary conditions (as indicated by the* 1 *degree difference between ICE6G and GLAC1D runs where the same model version was used). Thus we cannot infer ECS based on the LGM-to-Holocene temperature difference, although we agree that the role of climate sensitivities in fostering climate variability would be interesting to study.*
Anticipated changes: None.

Referee: "L422: "standard deviation provides a measure that increases with the spread of the

distribution" → technically, is IS a measure of the spread of a distribution. I find this wording needlessly mathematical."
Anticipated changes: We will change this sentence to "Standard deviation represents the spread of the distribution (c.f. Sect. 3.1).".

Referee: "L437 : "On centennial scales, this lack of skewness agrees with the results for the LGMR." It should be pointed out that LGMR uses a Kalman Filter that assumes Gaussian state vectors, so no skewness could be reconstructed, even if it was there."
Reply: *Indeed, a Kalman filter's assumption of normality will affect the results of the data assimilation (depending on the timescale of assimilation versus that of the final reconstruction). This effect is complicated by the impact of model prior and resolution of the individual proxy records. As such, our statistical analysis of the LGMR are to be interpreted with caution. The impact of data assimilation reconstruction methods on higher moments requires further research. We will clarify this in the text when discussing the LGMR in Sec. 5.2.2.*
Anticipated changes: We will add to the discussion of the LGMR in Sec. 5.2.2 and its comparison to highlight potential uncertainty due to potential impacts of the reconstruction method.

Referee: "Fig 5 : would it make sense for the y axes to have units"
Reply: *Since units are not common across rows or columns, it would be complicated to indicate them in the figure at is stands. Instead, we suggest adding the units to the figure caption.*
Anticipated changes: We will change the second sentence in the figure caption to "For all simulations standard deviation (left column, in units of °C for temperature and $\frac{mm}{d}$ for precipitation), skewness (middle column, dimensionless) and kurtosis (right column, dimensionless) are shown.".

Referee: "L465 "standard deviation changes only locally" -¿ What does this mean? There are definitely some continental or basin scale patterns in some of the panels."
Reply: *We agree that "locally" may not be the best word choice as it suggests the grid box level. We meant to indicate that across models most areas see only small changes with changing ice sheets (Fig. 6, panels a-f). We will change the sentence to better reflect the results and thank the referee for pointing out this imprecise wording.*
Anticipated changes: We will change the sentence to: "Changes in the spread of the distribution, as expressed in standard deviation, are regionally limited in response to the prescribed ice sheet reconstruction (Fig. 6 a–f)."

Referee: "L485 "differences in parametrization" influence the skewness. Which parameterizations might be responsible here? It may not be easy to guess, but if the authors have candidates in mind, they would be helpful to lay out here"
Reply: *Here, specifically the difference in parametrizations between the model versions of MPI-ESM relate to the formation of clouds. The changed parameter (csecfrl) describes the threshold between cloud water and ice according to the Wegener-Bergeron-Findeisen (WBF) process. This describes the preferred growth of ice crystals at the expense of large water droplets in supercooled and supersaturated conditions. The modification leads to warmer PI and LGM climates in r5 and r6 in comparison to r1 and r2. r1 and r2 were found to be too cold in comparison to reconstructions (for further details see supporting information of Kapsch et al. 2022). We will specify that the changed parametrization relates to cloud formation.*
Anticipated changes: We will change "This indicates that the differences in parametrization influence the skewness." to "This indicates that the difference in cloud formation parametrization influence the skewness.". We will further clarify the description of the simulations in L189-191 to say "They use different sets ice sheet reconstructions – GLAC1-D or ICE-6G_C (in the following

ICE6G, Peltier et al., 2015) – and vary by meltwater scenario. Further, a parameter for cloud formation was changed in r5–r7 to remove a cold bias found in r1–r4 (as detailed in the supporting information of Kapsch et al. 2022)."

Referee: "L540; L554 : the word "significant" is used on those 2 lines. Do the authors mean "statistically significant" (if so, by which test?) or do they mean something like "substantial"? Please clarify."
Reply: *This does indeed mean statistically significant. It refers to the results of the significance tests as described in Sec. 3.1, (t-tests testing for the hypothesis of normality). The results of the significance tests are summarized as the percentages of grid boxes with positive or negative significant changes below skewness and kurtosis maps.*
Anticipated changes: None.

Referee: "Fig 12 is very dense, and too small to distinguish many of the curves. The 6 PSD panels on the left have a curve/envelope in light gray, but there is no corresponding entry in the legend."
Reply: *The curves in grey correspond to the sensitivity set as described in the caption. We will remove them from the figure in the revised version of the manuscript. To enlarge the panels, we will further move both columns with the spectral ratios below the other panels.*
Anticipated changes: We will remove the sensitivity set from the figure and move panels d,e,i and j below the rest of the figure.

Referee: "L588 "we find temperature spectra that increase towards longer timescales" → this certainly appears to be the case. Can the authors be more quantitative here? What scaling exponents are involved? How does it compare to observationally-derived exponents? Is there evidence of multifractality? If so, are the regime transitions occurring at the right timescales?"
Reply: *Indeed, an analysis of the scaling by computing scaling factors and identifying scale breaks would be of great interest and could contribute to the debate on the nature of scaling, the existence and potential timescale of scale breaks and how these relate to glacial versus interglacial conditions (Huybers & Curry, 2006, Nilsen et al., 2016, Lovejoy, 2015, Rypdal et al., 2013). An in-depth analysis of simulated scaling, its dependence on mean state, region and forcing as well as a comparison the results to reconstructions (as e.g., in Ellerhoff & Rehfeld, 2021 for the last 2000 years) is, unfortunately, out of the scope here. We add this as a potential future target of investigation to the discussion.*
Anticipated changes: We will add to the discussion of scaling with respect to possible future analysis of scaling.

Referee: "L594 "relates to the simulated El Niño-Southern Oscillation". This seems to imply that those simulations exhibit enhanced ENSO activity with LGM boundary conditions. Is there any published explanation for this?"
Reply: *Indeed, the MPI-ESM simulations exhibit an enhanced ENSO signal during the LGM. However this has been noted elsewhere (e.g. Ellerhoff & Rehfeld 2021). None of the other simulations suggest a similar feature.*
Anticipated changes: We will note that the multi-model ensemble does not suggest higher ENSO activity.

Referee: "L602-603 : again, this can be quantified with scaling exponents."
Reply: *We agree that such an analysis is a promising direction for future work (see response to change suggested for L588*

Anticipated changes: See response to comment on L594.

Referee: "L625 : "into the tropic" → into the tropics"
Anticipated changes: We will change this as suggested.

Referee: "section 4.6.4 : LGMR provides an ensemble of Kalman Filter samples. Is this taken into account here, or are the authors only considering the ensemble median?"
Reply: *Indeed, we consider the ensemble mean. We will clarify this upon revision of the manuscript.*
Anticipated changes: We will add to Sec. 2.3, the caption of Fig. 5 and Sec. 4.6.4 to clarify that we use the ensemble mean for comparison.

Referee: "L672/3 : "Simulations that differ only by ice sheet reconstruction diverge most on long timescales, although differences can be found even for annual variability". This suggests that those boundary conditions affect the entire climate continuum, which is profound and deserves some commentary."
Reply: *We agree that this is an intriguing result, deserving further discussion. It clearly highlights that complex interactions of slow-changing boundary conditions can impact shorter timescales in the system. This was implicitly mentioned in Sect. 4.3 and 5.3 and the conclusions. We will highlight this upon revision of the manuscript.*
Anticipated changes: We will highlight that slow components can impact variability on faster timescales in Sect. 5.3 explicitly.

Referee: "Fig 13 : it took me a minute to figure out that the two gray lines correspond to PMIP end-members. The authors might want to clarify this in the figure caption."
Reply: *Indeed, the figure caption is incomplete and we thank the referee for catching this. We will change it in the revision.*
Anticipated changes: We will adapt the last sentence of the figure caption to read: "In panel d, Rehfeld et al. (2018)'s estimated range of the multi-centennial to millennial LGM-to-Holocene variance ratio based on proxy reconstructions (reconstructed) and interannual variability based on the PMIP3 ensemble (PMIP3) are marked for comparison."

Referee: "L758-759 : "Further, as a reanalysis product, the LGMR uses model simulations as priors and thus might be affected by a lack of variability in models." This intuition is incorrect. In the offline data assimilation flavor of ensemble Kalman FIlters used in products like LGMR or LMR, all temporal variability comes from the proxies; the model priors are used only to link variables across space, or to link one variable to another (e.g. surface temperature to sea-level pressure). So this comment would only be true if it applies to spatial variability"
Reply: *We thank the reviewer for this comment. We will adapt the discussion of the LGMR.*
Anticipated changes: We will change the sentence to "Further, as a reanalysis product, the LGMR uses model simulations as priors, which might affect the spatial patterns found.".

Referee: "last sentence reads very generic. Please make more substantive, or dispense with it."
Anticipated changes: We will replace the sentence: "Interactions between dynamics, forcings and mean state lead to complex changes in the distributions of surface climate variables. This implies potential changes to extremes on timescales from years to centuries, requiring further investigation."

References

Ellerhoff, B., and Rehfeld, K. (2021). Probing the timescale dependency of local and global variations in surface air temperature from climate simulations and reconstructions of the last millennia.

Physical Review E, 104(6), 1–14. https://doi.org/10.1103/PhysRevE.104.064136

Huybers, P. and Curry, W.: Links between Annual, Milankovitch and Continuum Temperature Variability, Nature, 441, 329–332, https://doi.org/10.1038/nature04745, 2006.

Kapsch, M. L., 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, Geophysical Research Letters, 49, 1–11, https://doi.org/10.1029/2021GL096767, 2022.

Lovejoy, S. (2015). A voyage through scales, a missing quadrillion and why the climate is not what you expect. Climate Dynamics, 44(11–12), 3187–3210. https://doi.org/10.1007/s00382-014-2324-0

Nilsen, T., Rypdal, K., and Fredriksen, H. B.: Are There Multiple Scaling Regimes in Holocene Temperature Records?, Earth System Dynamics, 7, 419–439, https://doi.org/10.5194/esd-7-419-2016, 2016.

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.

Rypdal, K., Østvand, L., Rypdal, M. (2013). Long-range memory in Earth's surface temperature on time scales from months to centuries. Journal of Geophysical Research: Atmospheres, 118(13), 7046–7062. https://doi.org/10.1002/jgrd.50399