Responses to comments of Referee #2

Last interglacial climate simulations provide an interesting insight into a warmer climate state, albeit under different orbital and greenhouse gas forcings than those of present-day. Here the authors address a notable issue in the PMIP4 last interglacial simulations, which show very little Southern Ocean SST warming in comparison to proxy records. The limited improvement between PMIP3 and PMIP4 in that respect highlights the work still to be done by both the proxy and modelling communities to resolve these differences. However, by adding additional freshwater into the North Atlantic in a classic hosing experiment, the authors find considerable improvement in the simulation of last interglacial Southern Ocean SST, at least in their model (HadCM3).

Response: Thank you for your valuable time in reviewing this manuscript. We have considered each comment carefully and responded to each one as below.

This is an interesting experiment fitting the scope of CP, and is of particular relevance given increasing interest in hosing experiments for both future and past climates. Nevertheless there are some important weaknesses in the design of the experiment that I believe should be addressed, as follows.

(1) Antarctica will also have contributed anomalous freshwater fluxes during the LIG. These may actually act against your northern influence, for example by cooling SST (Mackie et al. 2020 and several similar studies). Recent work has suggested Antarctica's sea-level contribution was early in the LIG, e.g. before 126 ka (Barnett et al, 2023). In that case the Antarctic freshwater fluxes may have been just as important as North Atlantic freshwater fluxes in their influence of Southern Ocean SST. Even though I understand the logistical advantage of simply continuing an existing model run, and the computational constraints, this point needs some careful discussion in your manuscript as it is (in my view) a major weakness of your experiment. How do the magnitudes of Arctic freshwater-induced cooling compare with Antarctic freshwater-induced warming in the Southern Ocean? Previous modelling studies can help answer this. Bearing in mind that sea-level rise could also help explain the apparent cold bias in PMIP4 lig127k Southern Ocean SST (Zhang et al. 2023).

Response: We really appreciate these insights and we fully agree that our idealised scenario does not account for the full range of processes. We added the following paragraph in the Conclusion at Line 356 to discuss potential future work on this.

"It is important to note that the hosed 128-ka HadCM3 simulation only represents an idealised scenario. The impacts of ice sheet meltwater would depend on its location, magnitude, and timing (He and Clark, 2022), as well as on climate background states (Pöppelmeier et al., 2023; Lynch-Stieglitz et al., 2014). Given that the Antarctic ice sheet may contribute to the peak early LIG global sea level (Barnett et al., 2023), the corresponding freshwater input could also influence the early LIG climate (Mackie et al., 2020). These processes should be investigated with coupled ice-sheet-climate models."

We added a sentence in Introduction at Line 51 to discuss the study of Zhang et al. (2023).

"While Zhang et al. (2023) found that increased global mean sea level warms southern mid-to-high latitudes at 126 ka using a climate model NorESM1-F, the root mean squared errors (RMSE) between temperature anomalies at 126 ka relative to preindustrial from the simulations and the Chandler and Langebroek (2021a) dataset were only reduced by ~10% while applying a 5-m or 10-m sea level rise."

(2) Comparison of simulated and reconstructed SST. See Section 2.5. Here the implication is that the simulated SST is being tested against the "truth", which here comprises SST reconstructions. However, reconstructions themselves have considerable uncertainty as discussed in Chandler & Langebroek (2021a,b). Their recommendation was to focus comparison on regional averages, rather than site-by-site. I would suggest to follow that approach in this paper and (for example) use the proxies to get three regional SST anomalies (Atlantic, Indian, Pacific sectors) then evaluate your results on a regional basis rather than site-by-site basis.

Response: We fully agree that the reconstructions may deviate from the "ground truth". Indeed we also compared regional averages between Table 2 and 3 (LIG-PI columns). And by calculating RMSE across sites, it takes into account regional variations, though it is not regional averages.

For preliminary analysis we also calculated regional averages for southern Atlantic, Indian, and Pacific oceans for model output. However, we chose not to show them for two reasons: 1) The differences between different ocean sectors in model-data comparisons are small. 2) Each ocean sector contains only a few records. Splitting available records into three subsets may introduce statistical bias in subsequent evaluation.

(3) Use of HadISST1 1870-1899 as a PI reference dataset. Probably this has some precedent in other studies, but as acknowledged in Section 2.4 there are very sparse observations for the Southern Ocean during 1870-1899. Consequently this comparison is not very convincing or useful without a lot more information about the errors in this region, in HadISST1. I suspect they will be fairly substantial! Why not use a more recent period? For example CMIP historical simulations. I'm fairly sure the models contributing to lig127k all have a historical simulation using the same configuration as PI.

Response: We understand the concerns on the quality of HadISST1 over the Southern Ocean during the preindustrial period. As suggested, we compared CMIP6 historical simulations of the 12 models with the SST product from the European Space Agency Climate Change Initiative (ESACCI) version 2.1 (1982-2014). As shown in the following figure 3, the model bias is quite similar to that while comparing piControl simulations with HadISST1 (Fig. 2). Considering that the data syntheses estimated LIG climate anomalies relative to preindustrial, we show the comparison between piControl simulations and HadISST1 in the manuscript.

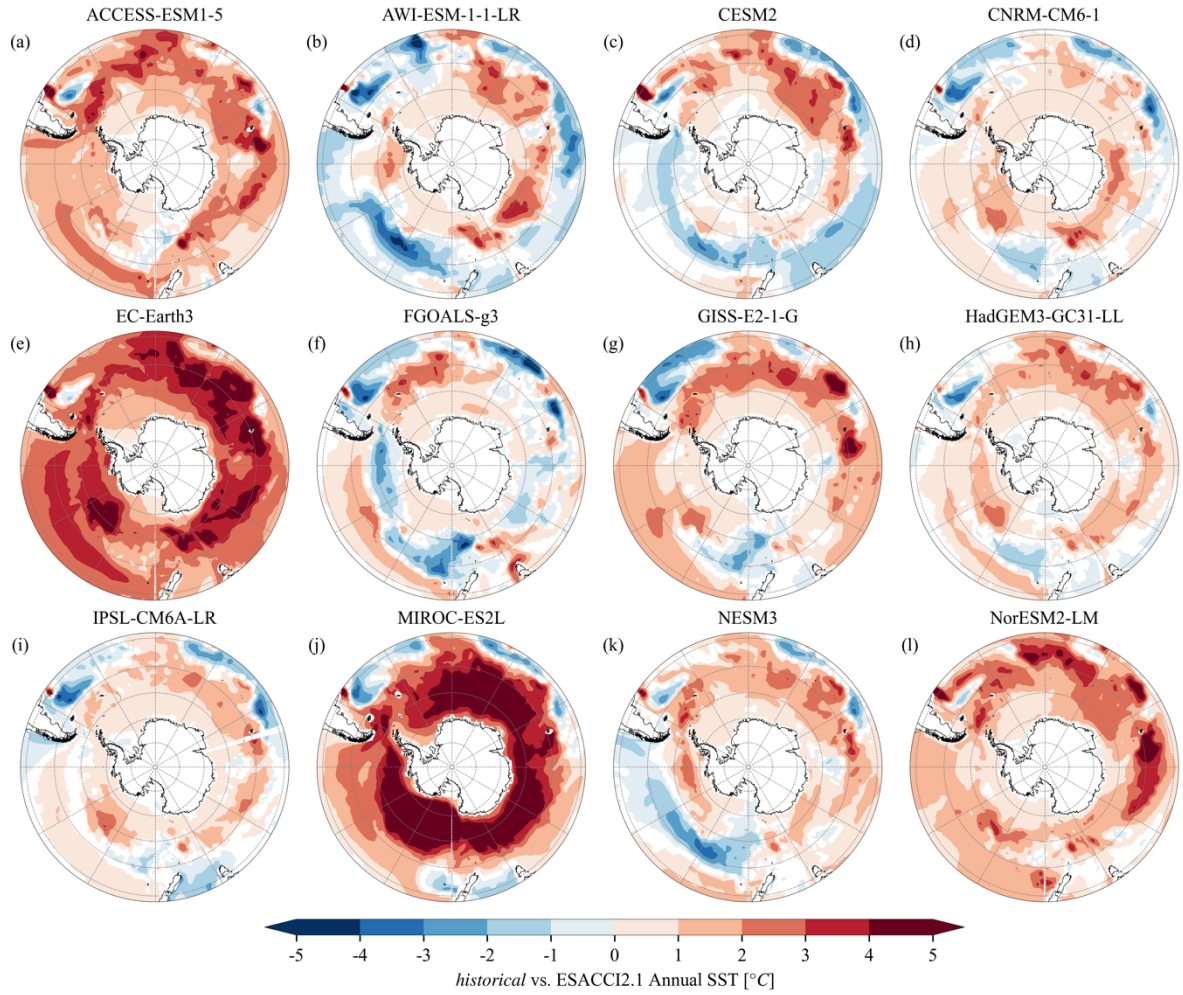

Figure 3: Differences in annual SST between historical simulations by the 12 models listed in Table 1 and the ESACCI SST dataset version 2.1 (1982-2014).

(4) Again in the comparison with reconstructions, the Turney et al 2020 ESSD paper is a surprising omission here. At first it might seem the Turney et al. methodology makes direct comparison with lig127k a bit trickier (Turney et al. report a "LIG average", not a time slice). However, they also report a 129-124 ka peak warmth. I would suggest using this, since it's actually very similar to the Capron et al. methodology in practice. This is because Capron et al. aligned their SST records to EDC air temperature, under the assumption of synchronous changes across the entire region. Consequently, their SST peaks are synchronous across all records and so the peak warmth in their study is always at 128 ka. This lies within the 129-124 ka window used by Turney et al. If you prefer to avoid too many comparisons, maybe swap the Capron et al dataset for the newer Turney et al. dataset.

Response: We thank Reviewer 3 for their suggestion and we now refer to the Turney et al. (2020) dataset in the revised version of the paper. We also explain below why we think that the Turney et al. (2020) dataset is not appropriate for our model-data comparison exercise at 127 ka (much of this argumentation is also provided to answer comments from Reviewer 1 and Reviewer 3 on this topic).

The major issue of the Turney et al. (2020) is related to the fact that the authors present a data compilation based on paleoclimatic records kept on the original chronologies. They do not attempt to build a coherent temporal framework between the different paleoclimatic records nor to provide an assessment of the large chronology errors that are associated to marine sediment records across the LIG. This is hugely problematic considering that using different dating strategies for paleorecords across the LIG could lead to age discrepancies of up to 6 ka as detailed in Govin et al. (2015). Hence, their strategy goes against the large efforts put together over the past years to guide the community towards being careful with age models for paleoclimatic archives during the LIG. Indeed, it is widely recognized now how key it is to harmonize paleorecord chronologies when building data compilations in order to provide (1) a realistic representation of the LIG climate and (2) appropriate benchmarks to evaluate the LIG model simulations (e.g. Capron et al. 2014, Govin et al. 2015, Stone et al. 2016, Hoffman et al. 2017, Capron et al. 2017, Otto-Bliesner et al. 2017, Menviel et al. 2019).

As a result showing a model-data comparison for this dataset would be a step-back for the data communities working on improving the spatio-temporal representation of the LIG climate since as previously mentioned, it ignores the current understanding of chronological uncertainties for the LIG and it would lead to misinterpretation and misused of this synthesis by other scientists part of the model community who might not necessary familiar with chronology-related subtleties, when performing model-data comparisons for the LIG.

Having said that we are aware that the syntheses that we use in our manuscript for the model-data comparisons still are attached to some limitations and improved syntheses should be developed in the future. We stress this point more clearly in the revised manuscript.

Other points follow below.

L5: are all 99 reconstructions independent, i.e., are they 99 different proxy records?

Response: No, some of the reconstructions use the same marine sediment cores as listed in Table A1. This is indicated at Line 135.

“Among these 99 records, there are 20 pairs of reconstructions based on the same marine sediment cores in at least two different datasets (Table A1).”

L10: Again from earlier comment, note that anomalous FW into SO might contribute some cooling (Mackie et al. 2020 & many others)

Response: We fully agree and it is stressed in the Conclusion.

“It is important to note that the hosed 128-ka HadCM3 simulation only represents an idealised scenario. The impacts of ice sheet meltwater would depend on its location, magnitude, and timing (He and Clark, 2022; Roche et al., 2010), as well as on climate background states (Pöppelmeier et al., 2023; Lynch-Stieglitz et al., 2014). Given that the

Antarctic ice sheet may contribute to the peak early LIG global sea level (Barnett et al., 2023), the corresponding freshwater input could also influence the early LIG climate (Mackie et al., 2020). These processes should be investigated with coupled ice-sheet-climate models."

L30, L88: I'd agree that estimating LIG warmth by compiling peak temperatures is not an appropriate approach. However, the SST records used by Capron et al. 2014/2017 were aligned to the EDC ice core temperature record, such that their synthesis also implicitly represents a synthesis of 'peak warmth' as noted in my earlier comment. Hoffman et al. followed a mixed approach (three key records representing three main ocean basins were aligned to EDC, but other records in each basin aligned to key record by d18O. Specifically from Capron et al: "*Marine records are transferred onto AICC2012 by assuming that surface-water temperature changes in the sub-Antarctic zone of the Southern Ocean (respectively in the North Atlantic) occurred simultaneously with air temperature variations above Antarctica (respectively Greenland*)".

Response: We acknowledge that making climate assumptions to infer age models is not ideal. However, in the case of the assumptions that Capron et al. (2014, 2017) used for the hemispheric alignments, it appears to be reasonable since the timing of climatic changes and the synchronicity between those observed in the ocean and in the ice core records could be checked while the different records were dated independently ((see Capron et al. 2014). Still, we propose to add the following sentence at Line 324 in Section 4.1 to stress this limitation.

"Lastly, it would be beneficial to build a synthesis with a coherent chronology independent from climate assumptions."

As mentioned before, we avoided using the Turney et al. (2020) dataset as they did not use consistent chronology and the peak warmth may occur much earlier than 127 ka. For comparison, Capron et al. (2014) found that peak temperatures occur at 129.3±0.9 ka in the Southern Hemisphere.

L118: Extraction of 127 ka anomaly. Why use 128 ka for the 127 ka anomaly? Surely better to use an average of 128 and 126 ka?

Response: We considered to use 128-ka values (20 and 21 records for annual and summer temperature, respectively), 126-ka values (15 and 17 records), and the average of 128- and 126-ka values (11 and 13 records). The average annual and summer temperature anomalies are similar: 2.2°C and 2.2°C for 128-ka, 2.4°C and 2.4°C for 126-ka, 2.5°C and 2.5°C for the average of 128- and 126-ka, respectively. We decided to use 128-ka values as a conservative estimate of the temperature anomalies with more records.

L145: Different characteristics of Chadwick et al 2021 might also reflect that they only used diatoms, whereas other datasets used multiple proxies.

Response: We added this point:

"The Chadwick et al. (2021) dataset suggests the smallest regional summer SST anomaly (1.2±1.1°C), which might result from the more southerly site locations (Fig. 1b) **and the fact that they used only diatoms for reconstructions**."

L156: I'd suggest to recap what are the key parameters/modelling choices used by Holloway et al. 2016 – presumably you keep these the same?

Response: Yes, we added the following sentence:

"The greenhouse gas concentrations in this simulation are close to those set by the PMIP4 lig127k guideline: carbon dioxide at 275 parts per million (ppm), methane at 706.8 parts per billion (ppb), and nitrous oxide at 266 ppb. The vegetation, aerosol, and ice sheets were set identical to the corresponding preindustrial simulation (Tindall et al., 2009)."

Fig 1: Can this be split into four rows, for the four studies, otherwise symbols plot over each other in a jumble.

Response: We agree that different records overlap in Fig. 1. We did try to plot each synthesis separately, but it does not look much better because records in the same synthesis also overlap.

Table 2 caption: useful to specify here again what is the PI reference used for the temperature anomalies.

Response: Added: "The preindustrial reference values are derived from HadISST1."

L164 'afflicted'... 'affected'?

Response: Changed.

L181: largest error, not largest bias. RMSE and bias are not the same thing. Which is an important point: you are ranking the datasets on their RMSE, not their bias. I think the bias should also be reported along with the RMSE.

Response: We changed "largest bias" to "largest RMSE". We focused on qualitative positive/negative bias, rather than delving into details of quantitative differences.

Table 3: what are the reported error bounds in the lig127k vs piControl anomalies?

Response: These are the standard deviations of the temperature anomalies over the region south of 40 degree south. We modified the caption:

"One standard deviation of the temperature anomalies over the region is given after mean temperature differences."

Table 4 headings: suggest to include RMSE specificically in the headings, i.e, Annual SST RMSE, Summer SST RMSE, etc otherwise this looks like a table of actual temperatures rather than temperature errors.

Response: Added.

Figs 3,4,5,6: “We only show differences that are significant at 5% level based on the student’s t-test with Benjamini-Hochberg Procedure controlling false discovery rates (Benjamini and Hochberg, 1995).” What does this refer to? I can’t see where differences are illustrated.

Response: “differences” is changed to “temperature/sea ice anomalies”

L218 Null scenario: ("*To benchmark model performance, we introduce a Null Scenario, where the climate at 127 ka is assumed to be the same as the preindustrial climate. … To demonstrate a better performance than the Null Scenario, the model simulations must have a smaller RMSE when evaluated against the climate syntheses than the Null Scenario*"). Can demonstration of the ‘Null Scenario’ be defined more clearly? In particular there could be some confusion here about what is being compared with what. Is the Null Scenario the piControl simulation compared with HadISST1? This approach is problematic because of potentially large errors in HadISST1 1870-1899 PI as noted above for Sec 3.1. Overall this confusion makes subsequent discussion somewhat difficult to evaluate.

Response: Thank you and no, the Null scenario is not the piControl simulation compared with HadISST1. The text is modified to clarify it:

“To benchmark model performance, we introduce a Null Scenario, where the climate at 127 ka is assumed to be the same as the preindustrial climate. **Therefore, in the Null Scenario, temperature and sea ice anomalies at 127 ka relative to preindustrial are zero.** The concept is similar to that of a persistence forecast obtained by persisting the initial conditions in weather forecasting (Murphy, 1992). To demonstrate a better performance than the Null Scenario, **i.e. to be useful,** the model simulations must have a smaller RMSE when evaluated against the climate syntheses than the Null Scenario.”

Fig 8: vertical error bars for the PMIP3 and PMIP4 ensembles would be useful, e.g. to show the standard deviation.

Response: This is a very good idea. The following figure 4 is an example for the PMIP4 ensemble of summer SST. The vertical blue lines show one standard deviation across 12 individual models. As the inter-model spread is relatively small, one standard deviation is normally around 0.5 degree. We added this for PMIP4 model ensembles, but not for PMIP3 model ensemble as individual model simulations are not available for PMIP3.

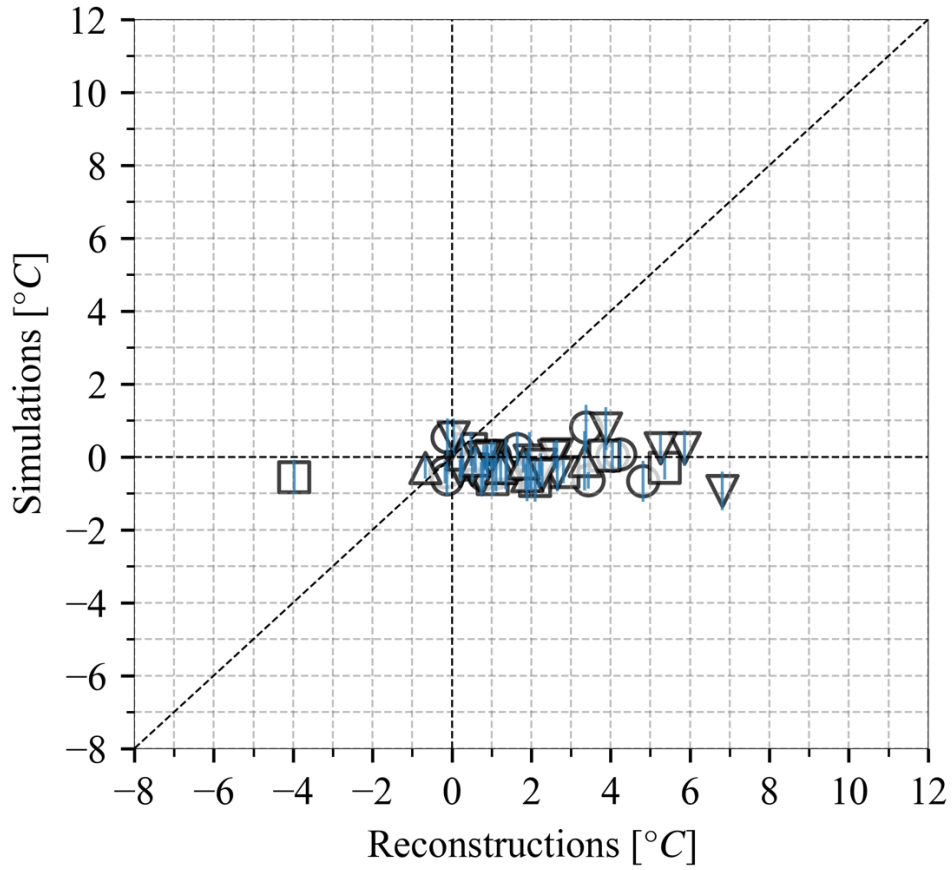

Figure 4. Reconstructed and simulated summer SST from the PMIP4 ensemble. The vertical blue lines represent one standard deviation among 12 models.

L272: Again evaluation against 1870-1899 HadISST1 is not a good benchmark given the likely errors in that dataset (particularly lack of data in S Ocean) & I would suggest to evaluate CMIP historical simulations against more recent observations in

Response: Please check the response above.

L286: "The ocean south of 40S becomes warmer…". Here and possibly several other places it is better to stick to "SST" rather than "ocean temperature" as it's only the sea-surface temperature that is being analysed. Suggest: "SST south of 40S becomes warmer…" etc.

Response: Thank you. It is changed as suggested.

L299: Pros/cons of using core-top reconstructed SST or observed SST … Of particular relevance here, not all the LIG SST reconstructions have a core-top sample.

Response: Modified: "Firstly, the preindustrial values were derived from a gridded dataset rather than reconstructed from core top measurements, **which are not available for some cores**."

L368: Acknowledgements – as well as acknowledging PMIP contributors, maybe nice to also acknowledge the many authors who have made their SST reconstructions available on public repositories.

Response: yes, that is true. We added: We acknowledge the PMIP modeling groups that contributed LIG simulations, the authors who generated the four palaeoclimate syntheses, and all scientists who produced climate reconstructions from paleoclimatic archives.

Refs

Barnett et al. 2023, Science, https://doi.org/10.1126/sciadv.adf0198.

Capron et al. 2017, QSR, https://doi.org/10.1016/j.quascirev.2017.04.019.

Chandler & Langebroek (2021a), QSR, https://doi.org/10.1016/j.quascirev.2021.107190.

Chandler & Langebroek (2021b), QSR https://doi.org/10.1016/j.quascirev.2021.107191.

Mackie et al. 2020, J. Clim, https://doi.org/10.1175/JCLI-D-19-0881.1.

Turney et al. 2020, ESSD, https://doi.org/10.5194/essd-12-3341-2020.

Zhang et al. 2023, Nat Geos, https://doi.org/10.1038/s41561-023-01153-y.