Responses to comments of Referee #3

Gao et al compared PMIP4 simulations for the Last Interglacial (LIG, 127 ka) with existing paleoclimate syntheses of sea and air temperatures, and sea ice concentration. The authors found that the warming recorded in the paleoclimate data cannot be captured by LIG model simulations. Aiming to explain the large model-data discrepancy, the authors also performed a North Atlantic freshwater hosing simulation on 128 ka and found that this simulation better agrees with the paleoclimate data syntheses.

The authors are trying to reconcile the data-model discrepancy for the LIG, which is suitable for Climate of the Past. However, both parts of the study have some major issues that need to be addressed before being considered for publication.

Response: We are thankful for your constructive comments. Please check below the responses to each comment.

Data syntheses

The authors acknowledge some limitations of using different LIG data synthesis (Sec 4.1). However, the quality of the employed data syntheses (mainly SSTs) needs to be substantially improved to make the model-data comparison meaningful. In the supplementary data table, 127-ka annual SST anomaly appears to be unrealistic (11.5 C at site ODP 1089, >5 C at E49-17 and MD02-2588, -6.8 C at MD73-025, 7.7C at MD97-2120). The SST anomalies at the same site based on different studies are drastically different (e.g., DSDP 594, MD88-770, MD97-2120, etc).

Response: We fully understand the concerns on the quality of the employed data syntheses. We noted the large SST anomalies in the Hoffmann et al. (2017) dataset, and the large differences in the reconstructions from the same cores between this dataset and other two datasets. This indeed motivated us to adopt multiple syntheses and focus on regional averages as also recommended in Capron et al. 2017. We still consider the adopted strategy here to be effective for model-data comparison. In fact the PMIP4 community thought so as well while recommending the use of these dataset for the lig127k model-data comparison exercises (Otto-Bliesner et al. 2017). We agree with Reviewer 3 that a single synthesis resolving all the issues is highly desirable, however 1) this is beyond the scope of our study and 2) we are aware of a group of experts already working on it at the moment.

The quality of the data syntheses can be significantly improved by revisiting the original data to resolve potential issues associated with inconsistent age models, different proxy calibrations, and the core-top SST values. These issues, mentioned by the authors, needed to be addressed.

Response: We fully agree that the data syntheses we looked into in our manuscript could be further improved and we are aware of on-going initiatives led by another group investigating the different proxy calibrations and the core-top SST values to build a new LIG data synthesis. Still, significant improvements in the syntheses we are using have been made (in particular related to the age models) while detailed PMIP4 model-data comparisons have not been made yet for individual models. Hence, we are convinced that our model-data comparisons with this particular selection of data syntheses is of added value. As a matter of fact, the data syntheses that we are using are the ones that have been recommended by the PMIP community itself (please see the recommendations formulated by Otto-Bliesner et al. 2017). In addition, we consider that it is already a great improvement compared to what had been done as part of the PMIP3 exercise where only peak-centred data syntheses could be used to evaluate the model simulations (Lunt et al. 2013). In the revised manuscript, we try and stress more clearly the limitations of the current data synthesis and the need for a data synthesis that will tackle some of the identified limitations in Section 4.1 Limitations of the LIG data syntheses:

While it would be helpful to provide a single unified LIG Southern Ocean and Antarctic data synthesis for benchmarking PMIP simulations, this would be a substantial additional piece of work. In the meantime, there are minor shortcomings in the four syntheses which could be addressed in the future. Firstly, the preindustrial values were derived from a gridded dataset rather than reconstructed from core top measurements, which are not available for some cores. It would be helpful to check the implications of this approach. Secondly, the spatial coverage of the proxies is still rather uneven. For example, a lack of southern Pacific Ocean sites will cause large uncertainties if the LIG climate were to be explored using data assimilation techniques, as was done for the Last Glacial Maximum (Tierney et al., 2020). Thirdly, some discrepancies exist in the reconstructions from the same cores between different datasets, particularly between the Hoffman et al. (2017) dataset and two other datasets in Table A1. Indeed this would be one of the main challenges to compile a single unified synthesis. Fourth, Capron et al. (2014) and Hoffman et al. (2017) did not use consistent calibration functions for each proxy as Chandler and Langebroek (2021a) did. Then, Chandler and Langebroek (2021a) and Chadwick et al. (2021) provided age uncertainties and reconstruction uncertainties separately, but did not estimate uncertainties accounting for both dating and reconstruction errors. It would also be useful to revisit the calibration of Antarctic ice core temperatures in light of recent work (Sime et al., 2009). Lastly, it would be beneficial to build a synthesis with a coherent chronology independent from climate assumptions. Overall, it would be most helpful if a future synthesis could address these issues.

The SST reconstruction at any point e.g., at 127 ka can be subject to uncertainties associated with measurements. Using the average over a period (e.g., 125-128 ka) can reduce the impact of such an influence.

Response: It is true that all reconstructions are associated with dating and reconstruction uncertainties. This was considered by both Capron et al. (2017) and Hoffmann et al. (2017) to provide quantified uncertainties associated to both the dating and the method used for SST reconstructions. In addition we would like to note that the reconstructions of Capron et al. (2017) at 127 ka represent an average value between

126-128 ka. We are not sure at this stage whether a longer average period would be better, since 127-ka was experiencing millennial-scale climate changes.

To derive the 127-ka SST anomaly compared to pre-industrial from paleo data, Holocene SST changes need to be considered too. This is because the core-top ages at many sites are not late Holocene.

Response: We are sorry we are not sure to understand the comment from the reviewer. However we would like to mention that we are aware that core top SST reconstruction might diverge from the PI values inferred from HadISST and this may introduce systematic offsets in temperature anomaly calculations. However, it was noted in the studies publishing the data synthesis that using core top SST as PI reference is complicated by the perturbation or the loss of the most recent sediments during the coring procedure, making it very difficult to date the core-top.

For SSTs, I suggest focusing on 1 data synthesis taking some of the above issues into account, and adding more sites following the same criteria.

These are additional work but are necessary to make the model-data comparison meaningful.

Response: As mentioned, while we agree a future synthesis addressing all the issues is desired as discussed in section 4.1, the current model-data comparison still provides valuable insights. For example, all four syntheses suggest a warmer Southern Ocean and Antarctic at 127 ka, and all model simulations without hosing cannot reproduce the magnitude of warming. This contrast indicates a robust finding that some processes are missing in the model simulations. Using a hosed simulation, we argue that meltwater input from northern ice sheets can be a candidate to explain the model-data mismatch.

In addition, since any new LIG data synthesis will likely be associated with some subjective decisions, e.g. criteria to include a given record or not, choice on the method used to derive the SST reconstruction and dating strategies, comparing model simulations with several syntheses relying on a coherent temporal framework has the added value to provide a broader overview and different possible scenarios. As far as we know it is common practice to evaluate simulations or present-day climate to multiple observation datasets considered independently because of different blending methods and assimilation systems. This is a point that we mention in the revised version of the manuscript.

“It is indeed a common practice to evaluate model simulations against multiple observation datasets that are considered independent because of different compilation methods.”

Hosing model simulation

The hosing experiment indeed reduced the RMSE between model and data. However, this improvement cannot be attributed to the 3,000-year hosing, without comparing 128 ka simulation without hosing with paleo data. And the length of the hosing period cannot be justified without comparing the 1,600-year hosing experiment (Holloway et al 2018 from the same group) to the same paleo data.

Response: We appreciate these points. We did compare a 128-ka simulation without hosing with the paleo data, which gives similar results as the 127-ka simulation. We mentioned it in the text as below.

“For comparison, we also run a standard 127-ka simulation of HadCM3 (HadCM3_127k), which gives qualitatively consistent results with a standard 128-ka simulation of HadCM3 by Holloway et al. (2018, not shown).”

We also compared the 2000-year hosing simulation with the paleo data, and it shows qualitatively consistent results with the 3000-year hosing simulation. This is because the climate tends to reach an equilibrium state after around 2000 years, as shown in Fig A1. However, we still consider the length of 3000 years to be valuable, as it provides a guidance for future studies on what are the optimal modelling length considering both computational costs and expected signals.

Additionally, I doubt if RMSE is suitable for evaluating model-data agreements. It appears that the larger RMSE is driven by the systematic offset between data and model. Investigating RMSE ignores spatial patterns of warming in different regions.

Response: We consider RMSE to be a suitable index to measure exactly “the systematic offset between data and model”. It is true that RMSE cannot reveal spatial patterns, as it is mainly regional “average”. We refer to the maps regarding spatial patterns of climate anomalies.

Some minor points

Line 76: more details are needed. How do these parameters differ from pi control?

Response: We added the following sentence:

The greenhouse gas concentrations were lower at 127 ka than preindustrial: 275 vs. 284.3 parts per million (ppm) for carbon dioxide, 685 vs. 808.2 parts per billion (ppb) for methane, and 255 vs. 273.0 ppb for nitrous oxide. At 127 ka, the Earth’s orbit was characterised by a perihelion close to the boreal summer solstice, larger eccentricity, and higher obliquity than preindustrial (Berger and Loutre, 1991). Such configuration affected the seasonal and latitudinal distribution of solar insolation at the top of the atmosphere, resulting in a small positive annual insolation anomaly at 127 ka than preindustrial at high latitudes (Otto-Bliesner et al., 2017).

Line 79: why mention the CNRM model specifically here?

Response: Because this is the only model that does not use greenhouse gas concentrations at 127 ka for *lig127k*.

Line 89: Annual SST and summer SST in the paleo data syntheses are often derived from the same dataset but with different calibrations (e.g., alkenone, see Chandler and Langebroek 2021). Therefore, these two SSTs in paleo data sets are not independent. This point should be made in the methods.

Response: We appreciate this point and added the following sentence in Line 100:

"Note that annual and summer SST could be reconstructed from the same proxy using different calibration functions (Chandler and Langebroek, 2021b)."

Line 149: Forcing parameters for the 128 ka simulation need to be described in detail for comparison with the 127 ka simulation

Response: We added the following details in Line 167:

"The greenhouse gas concentrations in this simulation are close to those set by the PMIP4 lig127k guideline: carbon dioxide at 275 ppm, methane at 706.8 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)."

Table 3: Good to add mean deference between pi control and HadSST1. From Fig. 2, the mean difference can be large for some models. How does this contribute to offset between the model (comparing 127 ka with pi) and data (comparing 127 ka with HadlSST1)?

Response: Thank you. We calculated mean differences between piControl and HadISST1, but we consider RMSE to be a better measure of model bias here, as mean differences are affected by the compensation between positive and negative bias. We added the following sentence in Line 191:

"We used RMSE to measure model-data discrepancies rather than mean differences to avoid compensation between positive and negative bias."

It is a very good question about how model bias in the preindustrial condition affects the simulated anomalies at 127 ka. While it is obvious that large warm bias in MIROC-ES2L undermines its applicability for a warmer climate (e.g. no sea ice to reduce), it is more complicated to draw any conclusions for other models. We also do not find a systematic relationship between the model bias and simulated anomalies. It is because of the model bias that we decided to focus on climate anomalies, rather than climate states at 127 ka.

Line 201: if you mean statistically significant, show the statistics.

Response: We did the student's t-test on the differences between annual SST in lig127k and piControl simulations in Fig. 3. Any differences we show in colour in Fig. 3 are statistically significant at 5% level.

Original text: “While the magnitude of simulated climate anomalies at 127 ka is small, the differences between lig127k and piControl are generally **significant** (Fig. 3).”

Line 215: How many is “a few”

Response: We modified this sentence:

“ACCESS-ESM1-5 and FGOALS-g3 show reduced September SIC over the southern Indian Ocean (Fig. 6a and 6f).”