Responses to Report #2

I would like to thank the authors for their thoughtful replies and corresponding revisions following the first review. The points I raised have mostly been addressed. Just one point remains, which in my opinion was not adequately addressed and I really would like to see clarified. It concerns the decision to not use the Turney et al. 2020 SST synthesis, on the basis that it lacks a chronological framework and that it is a peak-warmth synthesis.

Response: We appreciate the efforts and time of Referee #2. We fully understand the concerns on the usage of the Turney et al. (2020) SST synthesis. Detailed arguments are presented in the following responses.

I agree that a peak-warmth synthesis is in general not appropriate for a model-data comparison, especially at global scale. However, was it not the case that Capron et al. 2014/2017 dated their Southern Ocean LIG SST records by aligning SST with the EDC temperature record? For a time slice close to the EDC temperature peak, the Capron method implicitly creates a peak-warmth synthesis because the SST temperature peaks are all aligned to the EDC temperature peak at ~128 ka. Capron et al. justified this methodology by assuming regionally synchronous SST changes through the LIG. Quoting from Capron et al. (2014):

"We follow the strategy of Govin et al. (2012) to align marine records onto the AICC2012 ice core chronology. It is based on the assumption 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 over inland Antarctica (respectively Greenland)."

Turney et al. 2020 use the original chronologies, which again is an approach I don't think is appropriate if using a global dataset, and I agree with the authors this is in some ways a "step back" in terms of progress addressing chronological frameworks etc (see the authors' reply to reviewer #1). But on the other hand, you would not be using all the Turney et al records, as you would only be using a Southern Ocean subset. When selecting only Southern Ocean records, the Turney et al method of finding peak warmth in the period 124-129 ka is conceptually very similar to Capron et al. 2017 in which Southern Ocean SST records are aligned by peak warmth, as a means of obtaining the 127ka SST. If Capron et al can argue that Southern Ocean SST peaks were synchronous, why reject that same argument from Turney et al?

In my view, for the 127 and 128 ka time slices that are so close to the Antarctic Ice Sheet LIG temperature peak, then either (i) both the Capron methodology and that of the Turney et al "peak warmth" are acceptable for comparison with model data in the Southern Ocean, or (ii)

neither are. Hence, the decision to use Capron et al. 2017 but not Turney et al. 2020, needs to be much stronger and not dependent on the argument against the peak warmth approach.
Response: Thank you for this thoughtful comment. Firstly, we would like to clarify that the dating strategy of the Capron et al. (2014, 2017) dataset is different from the peak warmth approach used in the Turney et al. (2020) dataset. It is true that the Capron et al. (2014, 2017) dataset align the Southern Ocean marine SST record onto the EDC water isotope profile to get the marine records on the AICC2012 timescales using the assumption the referee copied above. However, in practice they did not use the peak warmth to define a tie point for the climatic alignments. Instead, they used the mid-point during the warming phase over Termination 2 and the mid-point during the cooling over the glacial inception before the first millennial-scale climate variability (see Fig. 2 in Capron et al. 2014). This means that the Capron et al. (2014, 2017) dataset did not fix by construction the timing of the peak warmth in the Southern Ocean to the EDC peak warmth.

We do not think the Turney et al. (2020) dataset can be treated equally as the current four syntheses, because it completely ignores age scale issues (they take records on their published original chronologies) and does not provide age uncertainty estimates. It is true that the other datasets are associated with some limitations (e.g. based on climatic assumptions that are not fully satisfactory), but they made efforts in homogenising the chronologies onto a reference age scale and in quantifying the associated uncertainties. It is the way forward as clearly stated in many papers from the large paleoclimate communities over the past few years including the Otto-Bliesner et al. (2017) paper presenting the PMIP4 lig127k guidelines. Moreover, unlike the Capron et al. (2017) and Hoffman et al. (2017) datasets that provide 127 ka time slice reconstructions representing 126-128 ka with age and reconstruction uncertainties, the Turney et al. (2020) dataset provides peak reconstruction values during 129-124 ka. If we use the Turney et al. (2020) dataset to evaluate the model simulation of 127 ka climate, the potential age differences (among the records in the Turney et al. (2020) dataset and between the average age of the records in the Turney et al. (2020) dataset and the 127 ka) would prevent us from drawing any conclusions about model-data differences.

We add the following paragraph in Section 4.1 to reasoning our choice of not including this synthesis:
"We note that a recent synthesis by Turney et al. (2020) compiles maximum annual SST estimates during the early LIG (129-124 ka), of which 28 records are located south of 40° S. The results from the model-data comparison using this recent dataset are similar to those obtained using the Capron et al. (2017) and Hoffman et al. (2017) syntheses. However, we do not include them in our study considering the strong limitations associated to the Turney et al. (2020) compilations related to the fact that the compiled records are kept on their original age scales and peak values are provided without quantitative age uncertainty estimates,. The

potential age differences among the records in the Turney et al. (2020) dataset, and between the average age of the records in the Turney et al. (2020) dataset and 127 ka would prevent us from drawing robust conclusions about model-data discrepancies."
Minor points... line numbers from the tracked changes version.

New Fig 2 Taylor diagram: suggest to avoid green and red, for colourblind readers.
Response: Thank you. As mentioned, we only show the Taylor diagram as responses to comments, not in the manuscript.

L58 "while applying" change to "after applying"
Response: changed.

L88 grammar, "… resulting in a small positive annual insolation anomaly at 127 ka than preindustrial at high latitudes".
Response: changed to "…, which results in a small positive anomaly of annual insolation at 127 ka compared to preindustrial at high latitudes."

L94 grammar "…we use 100-year simulation from the end of each model integration period" change to "…we use 100-year simulations from the ends of each model integration period" or simply "we use the last 100 years of each simulation".
Response: changed to "we use 100-year simulations from the ends of each model integration period".

L383 Antarctic ice sheet should be in capitals (Antarctic Ice Sheet).
Response: changed.

L384 "may contribute to " change to "may have contributed to".
Response: changed.

Responses to Report #3

Gao et al improved the manuscript a lot in this revision. I am convinced of the main conclusion that freshwater input may be a key factor to simulate the warm conditions during the Last Interglacial. However, I think a few issues still need to be addressed.

Response: We are grateful for the comments of Referee #3, which help us improve the manuscript.

Data synthesis

I still believe and the authors probably agree that a better data synthesis would help with the model data comparison a lot. I also agree that publishing of the manuscript should not be prevented from not making a new data synthesis, if it is the convention in the field, as the authors explained. I do think, however, that the manuscript may be improved without too much effort in data compilation. Maybe it is good to at least get rid of some inconsistency between different compilations by getting averages for SSTs at the same sites.

Response: We fully agree that a better synthesis is extremely beneficial for model-data comparison, as discussed in Section 4.1. We understand it is not optimal to keep the inconsistency between different syntheses in the manuscript, but unfortunately, we are not in a position to favour one published SST reconstruction over another. The various methods used in each synthesis could explain some of the differences, and we believe more thorough work should be conducted to homogenise the SST reconstruction from proxies. We think simply averaging the reconstructions would mask the underlying issues potentially arising from different age scales, calibration functions, and reconstruction methods.

Null hypotheses

Following a point raised by Reviewer 2, the null hypothesis is not clearly described in the revision, and further clarification is needed.

I thought the null hypothesis presented by the author is comparing reconstructions at 127 ka with HadlSST (bottom line in Table 4), with the assumption that simulations by models are the same at 127 ka and pi condition. This point should be clearly mentioned somewhere in section 3.3. In this section, the authors compared the means between reconstructions with HadlSST, without setting the scene for comparing RMSEs against null hypotheses..

However, given the very large difference between your simulations and HadlSST, maybe an alternative set of null hypotheses is comparing reconstructions with simulations in both 127 ka and pi conditions? In this case, there are 12X4=48 null hypotheses associated with a different RMSE for SST. I am not saying this alternative is better, but the authors should better clarify the null hypothesis and consider this alternative way to set up the comparison.

Moreover, it seems to me that when comparing RMSEs from model-data comparison with the null hypothesis, some kind of statistical test is needed. For example, looking at the

comparison between models and SST from EC2017 against the null hypothesis, I am not sure how confident the authors are in the statement that MIROC-ES2L (RMSE=4.1) outperforms the null hypotheses while ACCESS-ESM1-5 (RMSE=4.3) does not. If the test is similar to the t-test, then I think you can also take into consideration the number of observations in the reconstruction in assessing the robustness of the conclusion.

Response: Thank you. Firstly, we would like to clarify that we introduce a Null Scenario, not a hypothesis. In the Null Scenario, the 127 ka climate is assumed to be the same as the preindustrial climate, so the SST anomalies at each core site are zero in the Null Scenario. Then, when we compare the Null Scenario with a synthesis, e.g. annual SST from the Capron et al. (2017) dataset, we obtain a RMSE of 4.2°C. Intuitively, for a model simulation to be useful, when compared with a synthesis, it must demonstrate a lower RMSE than the Null Scenario.

The student t-test may not be applicable here. Firstly, it tests the significance of differences in mean values, not RMSE. Secondly, it normally requires row data with more than 30 entries, but we have only limited records for each dataset. We are not aware of any other suitable statistical tests, but we consider the comparison provides valuable qualitative insights.

We added the following sentence in line 242 to clarify on the Null Scenario: "Then when we compare the Null Scenario with a synthesis, e.g. annual SST reconstructions from the Capron et al. (2017) dataset, we obtain a RMSE of 4.2°C, which serves as a baseline to evaluate model performance"

Detailed comments

Line 144: mentioning HadlSST before introducing to readers what this is.

Response: We introduce HadISST1 now at line 114 when we first mention it: "the HadISST1 dataset (1870-1899), which contains global monthly SST and SIC on 1°×1° grids from 1870 to present and is constructed by the UK Met Office using multiple observational data sources (Rayner et al., 2003)."

Section 2.4 Not sure why you compare model results with SST from the European Space Agency Climate Change Initiative rather than HadlSST during the same period?

Response: Thank you for this comment. The SST dataset from the ESA CCI project has much finer temporal (daily) and spatial (0.05 degree) resolution. Although the spatiotemporal resolution of HadISST1 is already enough for model evaluation, we consider checking different datasets to be a good way to indicate the robustness of our results.

We modified the following sentence in line 183: *"Indeed, we also compared annual SST between 1982-2014 in CMIP6 historical simulations of the 12 models with a SST dataset*

from the European Space Agency Climate Change Initiative (Merchant et al., 2014), ***which has much finer spatial and temporal resolution than HadISST1****."*