Short summary on anticipated changes to the manuscript:
- We will update the title to "Testing the reliability of global surface temperature reconstructions of the last glacial cycle using pseudo-proxy experiments"
- We will clarify in the data section that we only use the records metadata from the proxy database, and not the SST estimates.
- We will improve the readability of the plots.
- We will enhance the paper with the points the reviewer suggested.

Response to comments from Kaustubh Thirumalai

Major Comment 1

Bioturbation mixing depth parameterization*: The authors conclude that "Our results show also the existence of a trade-off between the inclusion of a large number of records, which overall reduces the uncertainty, and of only the highest quality records (low age uncertainty, high accumulation rate, high resolution), which improves the reconstruction of the short timescale" where the latter refers to reconstructions being relatively free from bioturbation- and age-model-related smoothening to sufficiently preserve short-timescale (or multicentennial-to-millennial) climatic signals. The manuscript (Table 1 and Figure 2 caption) suggests that the authors used a constant (presumably?) sediment mixed-layer depth (Table 1 only indicates 'bioturbation' and would benefit from more explicit details about what is being parameterized) of 10 cm. In my opinion, and according to my cursory assessment of global bioturbation rates (I realize that argue otherwise—but I'd like to see the numbers) of the data presented in Teal et al. (2010)—10 cm is far too high for average global mixing depths, particularly given J20's bias towards tropical and near-coastal proxy locations. I would like to see how a value of, for example, 5 cm would perform for the Full PP and related experiments. I understand that the authors have attempted to parse the sensitivity of 'age uncertainty' versus 'bioturbation' and other associated parameters in their analysis, but this does not address the entire hierarchy of choices with a lower bioturbation rate. If feasible, I'd recommend that the authors perform such an experiment (Full PP with reduced bioturbation rate) and check whether more high-frequency information is retained in the associated spectra.*

This is a valid point, we did not further investigate the impact of different bioturbation depth on the signal. We agree that, since Boudreau 1998 evaluated mixing depth at about 10cm, more recent studies have suggested lower mean values (Teal et al. 2008, Zhang et al. 2024). One justification to our use of a conservative estimate is to account for temporal variability, and the possible occurrence of higher values. Even if episodes of deep bioturbation last a short time, they may have a large impact on the signal if the sedimentation rate is low. A second justification is the weak impact of smaller depths on our result as shown in the Fig.1 below. In this figure we provide a comparison of pseudo-proxy experiments with 10 cm and 5 cm bioturbation layers. Looking first at the effect of bioturbation only, we find as expected that the drop in the coherence occurs at half the timescale than if we use a 50% smaller bioturbation depth. However, if we consider all factors (Full PP), the change does not sensibly affect our results. This is because bioturbation is never the sole main factor explaining the drop in coherence. It is nevertheless worth noting that, in the case of the last 25 kyr (Panel B and C), bioturbation does not seem to play a relevant role in the smoothing of the signal any more when the mixing depth is reduced to 5cm. We will add this information to the manuscript.

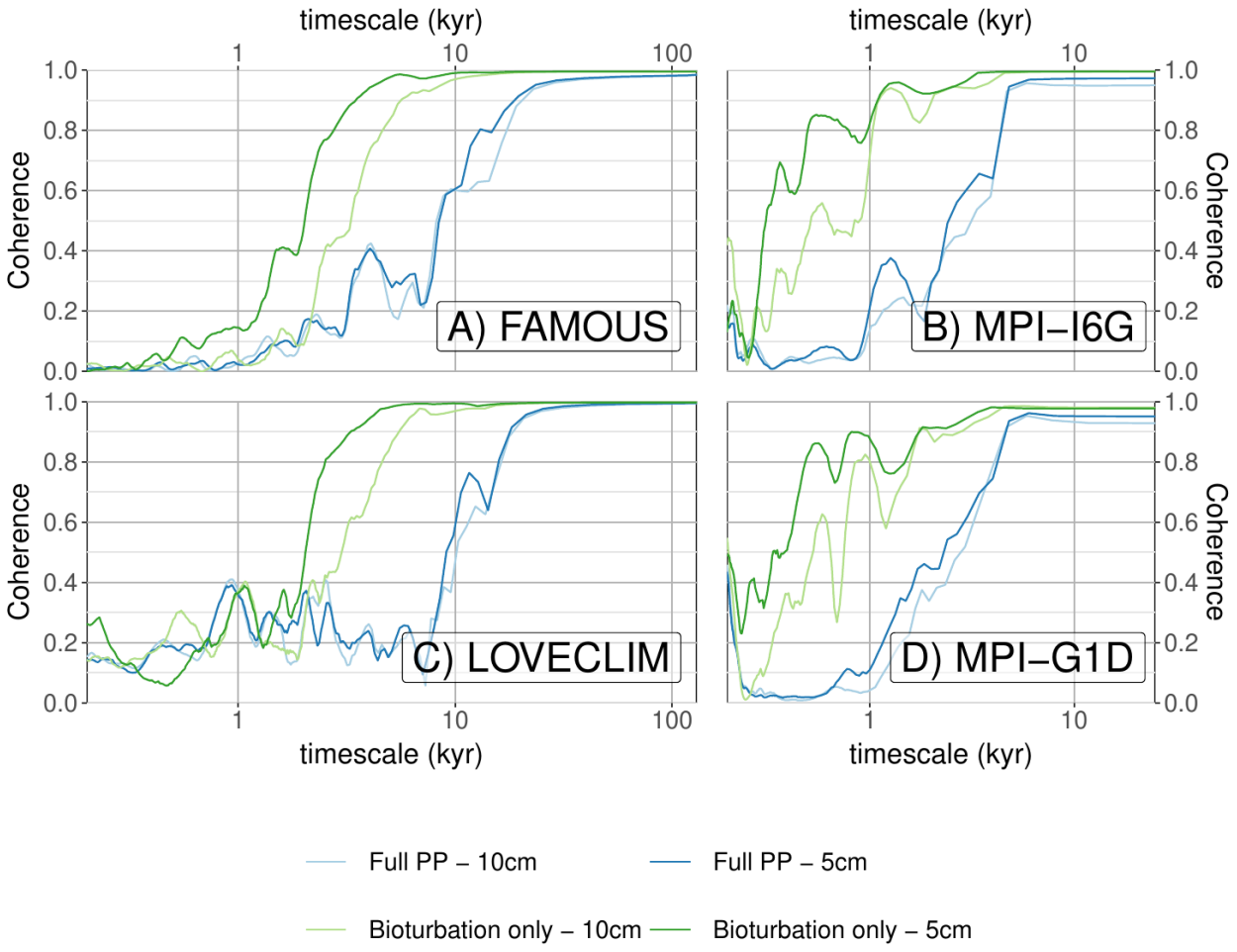

Figure 1: Impact of bioturbation rates on the coherence of the signal. We test two mixed-layer depths: 10 cm (light colours, same data as in the manuscript in Fig.8) and 5 cm (dark colours).

Major Comment 2

Clarity regarding the use+utility of J20*: Unless I am mistaken (which is highly likely), the authors do not actually use or show the GMST calculated (using any algorithm) from the reconstructions collated in J20. They merely use the metadata of proxy parameters within J20 as a framework for their pseudo-proxy experiments. The clarity of the manuscript would be improved if the authors were more upfront about this aspect in their abstract and introduction. On the other hand, this also leads me to question this omission as potential comparisons between reconstructed (from data), simulated, and resampled GMST calculations would be highly interesting. However, I recognize that this may be beyond the scope of this work—accordingly, I feel that the authors should preface this aspect and consider using the term metadata in their abstract and text. I think that the title of the manuscript should include/reflect 'sensitivity/uncertainty experiments' and/or 'pseudo-proxy experiments' because in its current form, without comparing simulated and actually reconstructed GMST, I do not think the authors can claim to test the 'reliability the reliability of global surface temperature reconstructions of the last glacial cycle'. Rather, they are testing the reliability of the methodologies used to create global mean reconstructions… hence I feel that a title revision is needed.*

- You are correct, we only use the metadata of the J20 dataset, not the actual measurements. We will clarify this on the manuscript. However, producing and analysing the GMST reconstruction based on the J20 data is out of scope for this manuscript.
- We argue that the pseudo-proxy experiments provide a more robust methodology to test the reliability of the reconstructions than model-data comparisons. In particular, directly comparing model outputs and proxy data suffers from the same limitations as the one we face when extrapolating our quantitative results from model data to real reconstructions (e.g. the LGM bias). However, our evaluation of the reconstruction algorithm still provides further insights, in particular regarding which timescale can be reliably reconstructed. We will add this argument in the introduction. In addition, given that all our study revolves around PPE, it is worth adding it to the title.

Major Comment 3

Inclusion of 'Full PP at random locations': *It appears that the authors do not show any results from this experiment, yet it plays an important role in their analysis (see, e.g., Lines 335–340: "In addition, location resampling and latitude band configurations, which aim to account for it, are not large enough to cover the bias. Yet, the pseudo-proxy experiments using random proxy locations can reproduce the simulated MSST.") I recommend that the authors make a plot showing results from this experiment and to be more quantitative/precise regarding the ability or lack thereof of these simulations to reproduce simulated mean SSTs.*

We designed the experiment "Full PP at random locations" mostly to investigate the impact of proxy numbers on the reconstructions (L. 350-356). We realised later that they could also be used to study the effect of locations biases. The respective timeseries are shown below in Fig.2 . We can see that the increased number of records has only a small impact on the mean reconstructed values, and that these are very close to a smoothed version of the simulated GMST. However, the increase in the number of proxy time series reduces noticeably the confidence interval.

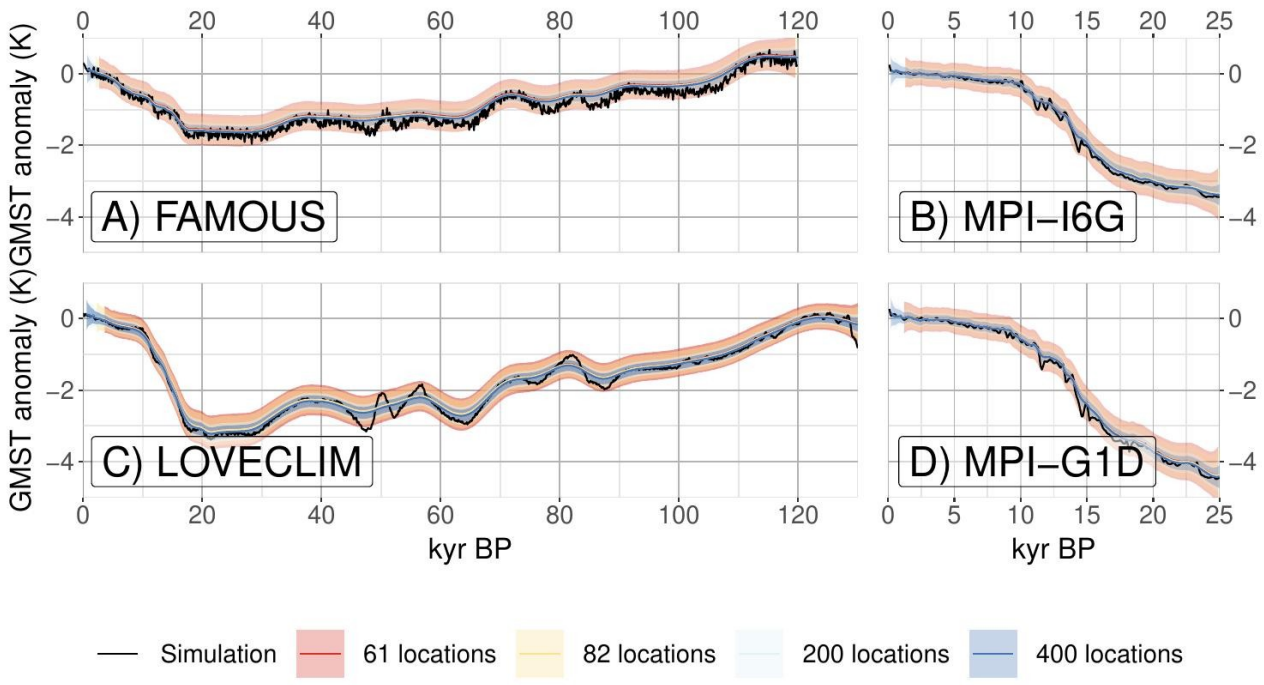

Figure 2: Reconstructions based on the experiments "Full PP at random locations", for various number of records. The reconstructions (colours) are compared to the simulated GMST (black). The shading correspond to the 90% CI.

Major Comment 4

Attribution of a specific set of J20 locations as a significant bias: *Based on the last point, the authors state that "Therefore, the bias is caused by the specific set of locations in the J20 dataset: there is an over-representation of regions with strong LGM cooling". Whereas this assessment may be accurate, I do not find the regions that the authors identify to be a convincing explanation (e.g., NW Atlantic/Kurushio extension)—instead, it seems to me that this is a bias related to the relative proximity of core locations to continents—where land-based cooling strongly impacts these sites as opposed to open-ocean marine conditions. Is it possible for the authors to combine inferences from the 'Full PP at random locations' or another sensitivity experiment to test this possibility?*

We show below the map of LGM (19-23kyr) temperature anomaly as used to compute the pseudo-proxies for the 5 simulations, with the location of records with data in both the last 5kyr and the LGM. There is no clear specific cooling pattern at the proximity of the continents in the simulations. By contrast, there are large differences in and between the North Pacific and the North Atlantic, where the sampling of location is quite heterogeneous. In the text, we point in particular to undersampled regions where the cooling is weaker, hence explaining the cold bias in the reconstructions. Note that we do not rule out a cooling bias due to the proximity of the continents in the proxy reconstructions. The simulations we use simply do not show such an effect.

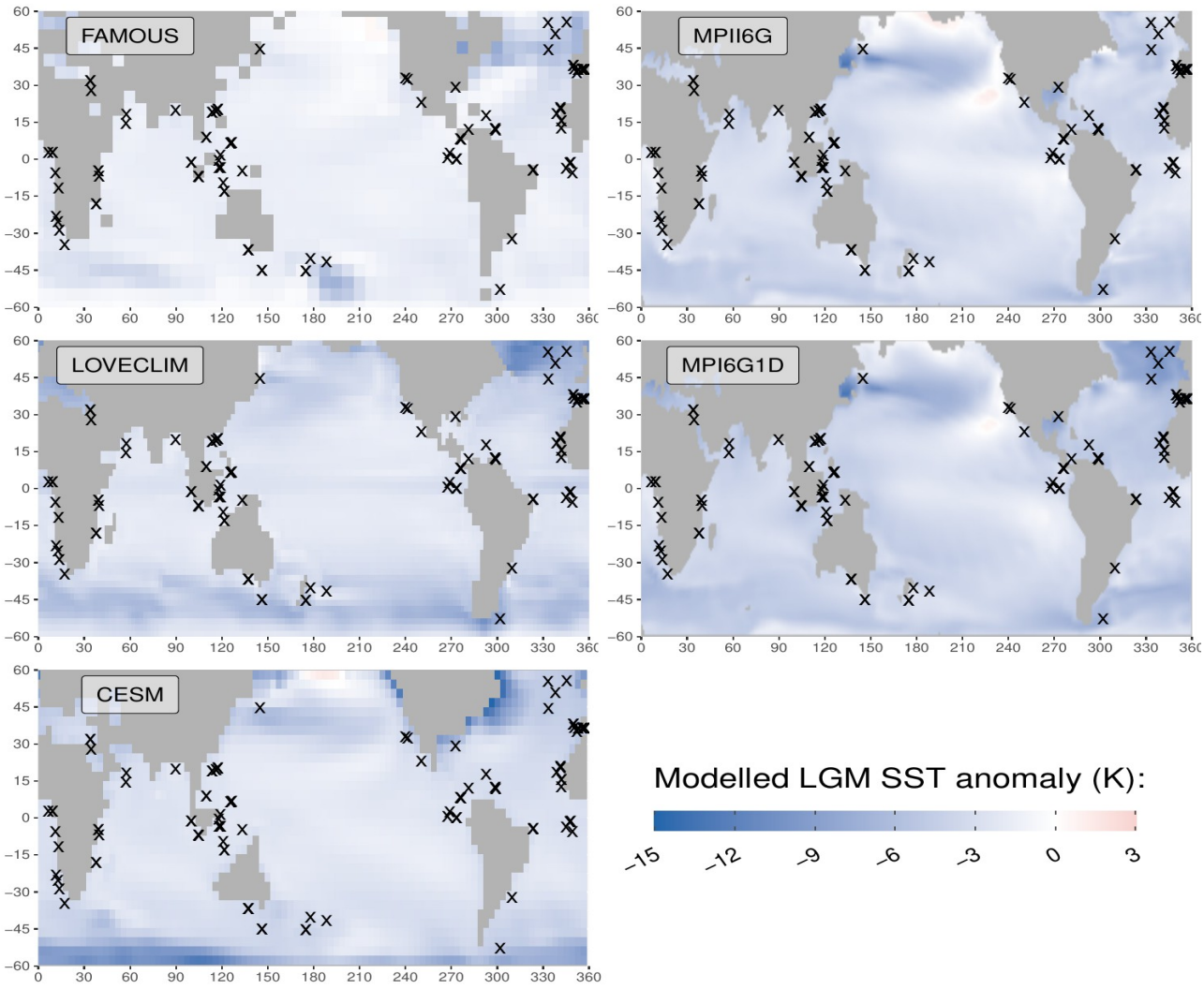

Figure 3: LGM temperature anomaly for the 5 simulations. These are the temperatures used in the pseudo-proxies (SST or ts with the proper land mask). The crosses correspond to the record locations with at least one value during the LGM (19-23 kyr) and one during the reference time period (0-5kyr).

Minor Suggestions

Line 81: "…needed to develop our evaluation standards." Please rephrase.

- This sub-sentence is not needed and indeed a bit confusing. We will remove it.

Lines 91–93: Are there only 7 (112–105) unspecified datasets? Figure 1 says otherwise—please clarify.

- No there are 112. 105(annual)+112(unspecified)+48(pseudo-annual) = 265 (total)

Lines 162–163: Please consider adding more information to contextualize why the following steps are being performed. This would be a great spot to clarify the involvement/extent of usage (or lack thereof) of the actual records within J20.

- What is used from J20 is more relevant for the construction of the pseudo-proxy, and we prefer to add it in the data section. As for here, we will change the sentence to be more explicit: "We decide to adapt the S16 algorithm to make full use of the J20 dataset (age ensembles, records with no data in the last 5kyr), and improve the uncertainty quantification by the algorithm."

Lines 236–237: 0.26 K and 0.23 K seem to be exceedingly low values for analytical uncertainty. Does this take into account sampling uncertainty (see e.g. Thirumalai et al. 2013) which is the uncertainty that foraminiferal shells (with lifespans of a ~month) would have grown at different points of time within the sampling interval, and thus will have uncertainty in reconstructing the 'interval mean'? If sedproxy takes this into account, it would be good to mention this aspect.

- Sampling uncertainty is applied for Mg/Ca as explained in the paragraph just above (L.228-235). The values are otherwise justified in Dolman and Laepple 2018.

Lines 245–246: Have you considered performing a depth-sensitivity test? i.e. instead of the uppermost grid location, what about the integration of the top 50 m—which is a more realistic scenario for the proxy integration of temperature signals for the chosen sensors. Perhaps this also might explain the cool bias?

- The main limitation to analyse the sensitivity to depth is data availability. Using a constant depth would most likely lead to slightly smaller amplitude of variation and to warmer LGM anomaly. This does not constitute an explanation to the cool bias however, but instead add a layer of uncertainty. Change in habitat depth through time can also be even more impactful, as we mention in the discussion / outlook of the paper.

General comment on Figures: Where there are many lines depicted in figures, it is very difficult to parse the colors of each line (especially on the spectral plots) and to associate them with the legend. I would strongly consider using a different backdrop color or thicker lines with a different subplot layout to more cleary delineate results from various experiments.

- Thank you for the feedback, we will increase the thickness of the lines, particularly in the legend, and test if the plots can be split in two.

Discussion and Lines 545–555: The authors should consider discussing the values of ***scaling utilized in Clark et al. (2024)*** *and how they fit within this portion of the discussion.*

-> We present below a scatter plot of GMST VS GSST in the simulation we used, including Snyder 2016 and Clark et al. 2024, with uncertainty. In short, the two are very similar, but the uncertainty range behave differently. Hence, it doesn't seem that the polynomial fit of Clark et al 2024 brings a better agreement. In addition, we would argue that neither a purely additive noise as in Clark et al 2024 or a purely multiplicative noise as in Snyder 2016 are able to capture the uncertainty correctly. Part of the issue in Clark et al. 2024 could be that they only use CESM-based simulations for the colder than PI part of the fit, which undersample the uncertainty. We will add a couple of sentences in the paragraph about the new fit from Clark et al. 2024, thank you for pointing it out.

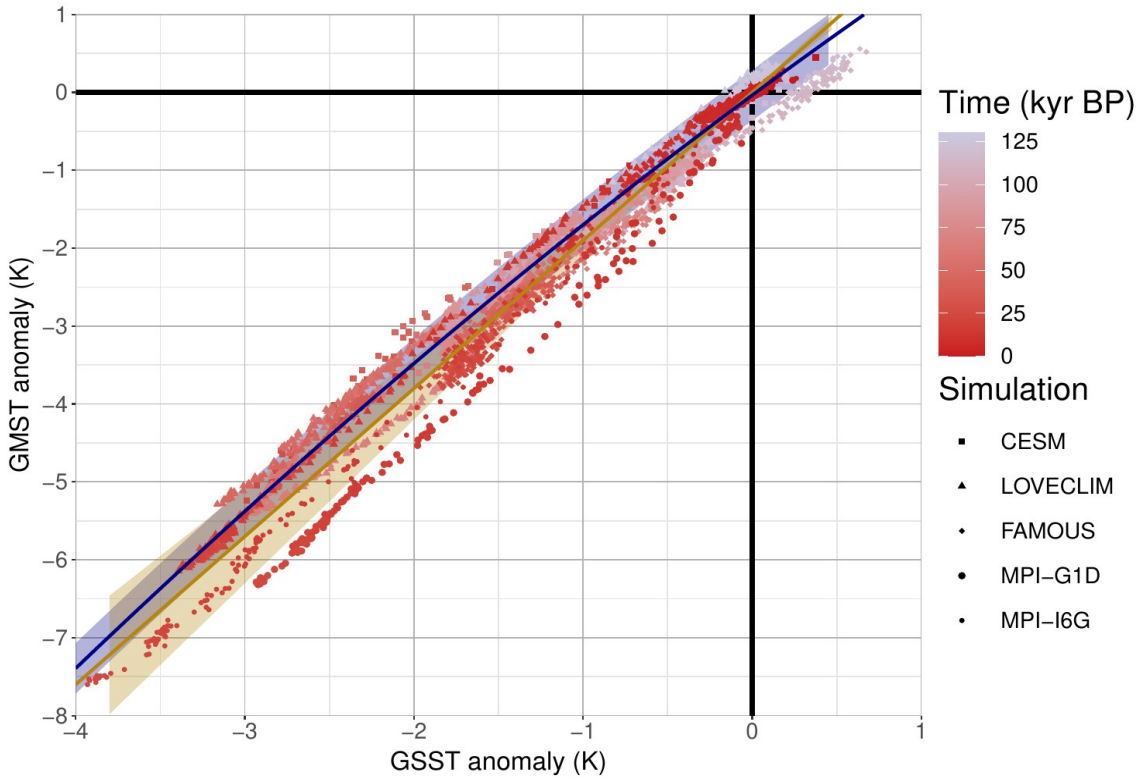

Figure 4: Comparison of the scaling between GMST and GSST anomaly. The dots correspond to the various simulations used in the study, with the colour indicating the time period. The dark orange line corresponds to the scaling used in Snyder 2016, and the shading to 2 standard deviations. The blue line is the scaling used in Clark et al. 2024, with the shading also corresponding to 2 times the standard deviation.