Dear Reviewer,

Thank you very much for reviewing our discussion paper and your constructive comments. Below we respond to your comments set in blue italic font. The author comments are set in black normal font.

Substantial Concerns:

1. **The assumption of model simulation as reality:**

o Although the authors are upfront regarding their underlying assumptions and essentially state that they are taking the model output at face value, I strongly recommend exploring more ways in which the model simulations might be oversimplifying their results. For example, in their Discussion section, the authors discuss the role of the "spatial correlation structure of model simulations" and how these correlations might be overestimated. The authors should add a discussion here about how biases in the simulation of climate variability itself in these transient models can lead to biases in correlation distances.
o In other words, if a (hypothesized) transient simulation from 6 ka to present showed the same, coherent changes across the entire Northern Hemisphere, the calculated SNE, as the authors propose, would be exceedingly low - however, we know that such a transient simulation is an unlikely representation of reality. Thus, I feel that the manuscript would greatly benefit if the authors included text on how typical (and atypical) shortcomings of MPI6k and T21k are influencing their results.

We agree with the reviewer that including a better discussion of the correlation structure, possible shortcomings in the model simulations and their effect of our results would be useful. This was also asked by Reviewer 1.
We suggest to address this point in the revised version by 1.) adding a new section 3.2, discussing the spatial correlation structure in the models vs. the spatial correlation structure in reanalysis data, and 2.) extending the discussion section to include a list of potential model shortcomings that may lead to an overestimation of the spatial coherency in the models.

Concerning 1.)
To check the realism of the correlation structure in the model simulations, we further analysed the correlation structure of the surface temperature field in the 20C3M reanalysis product (Compo et al., 2006) (Fig.R1). Interestingly, analysing the full time-period of 1871-2011 results in a much higher decorrelation length than estimated for the Holocene, likely caused by the coherent anthropogenic forcing. Removing the last decades to minimise the human influence, e.g., analysing 1871-1950 results in a correlation structure resembling the spatial correlation of MPI6k.
As we expect that the climate does not get more localised on longer time scales, but if anything, more spatially coherent (e.g., Jones et al., 1997; Kim and North, 1991) this suggests that the decorrelation lengths used in this study might not be unrealistically large.
Thus, instead of relying on climate model simulations one could even obtain similar results based on the reanalysis correlation structure and assuming that the correlation structure is similar on longer time scales than on the time scales sampled by the instrumental data. To make this point, we suggest adding the reanalysis correlation structure estimated from the proxy positions in the manuscript Figure 3 (Fig.R2).

One could still argue that fine-scale structures (e.g., at the coast or at shelves) not resolved by the models (as well as by the reanalysis) might lead to localised variations as we already discuss in Section 5.1, but we do not see a clear evidence for this on inter-annual and longer time scales from analysing high-resolution model simulations (e.g., the AWI-FESOM simulation in an eddy-permitting resolution). However, as this latter work is still preliminary we would not include it and just discuss this possibility.

Concerning 2.)
There are several shortcomings in present climate model simulations such as the two simulations used here that may lead to an overestimation of the coherency in the models. Possibilities include that models underestimate internal climate variability that is generally more localised than externally forced climate variability (e.g., Laepple and Huybers, PNAS 2014). One possibility (Laepple and Huybers, GRL 2014) is that the model effective horizontal diffusivity may be too large which would reduce internal variability and lead to larger correlation structures. Further, the low, non-eddy permitting resolution of the model simulations used here might suppress small scale features and the role of persistent coastal currents.

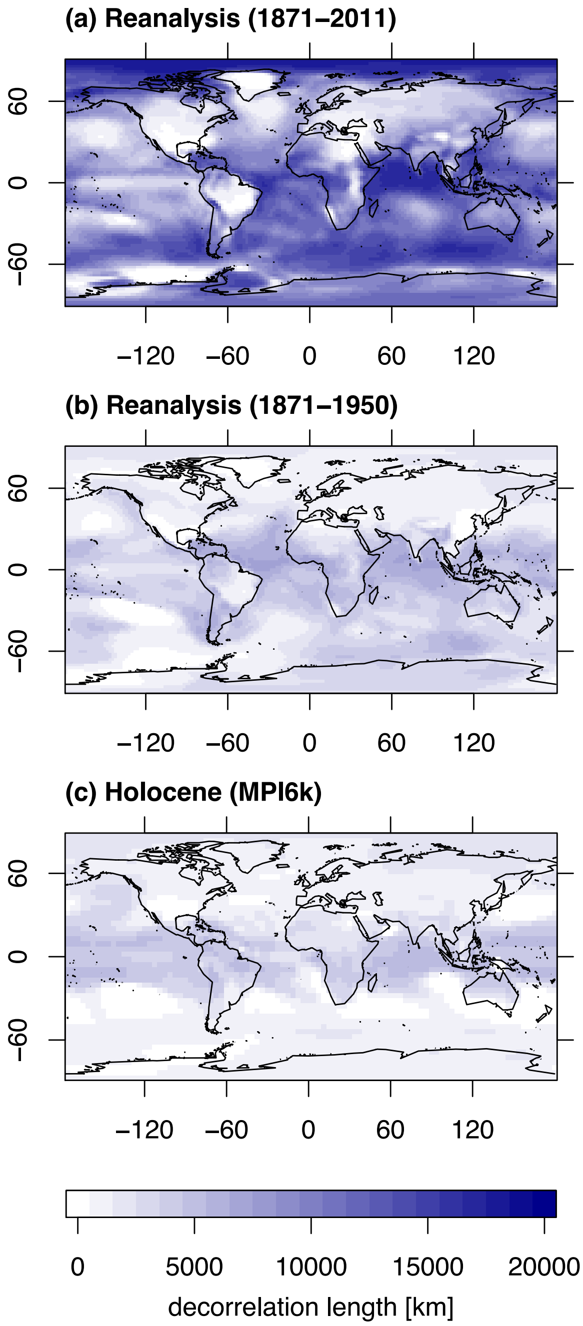

Fig.R1: Decorrelation length of reanalysis data and the 6ky simulation of MPI6k. The decorrelation length is similar for the Holocene and reanalysis data from 1871 to 1950 indicating that the Holocene spatial correlations are realistic.

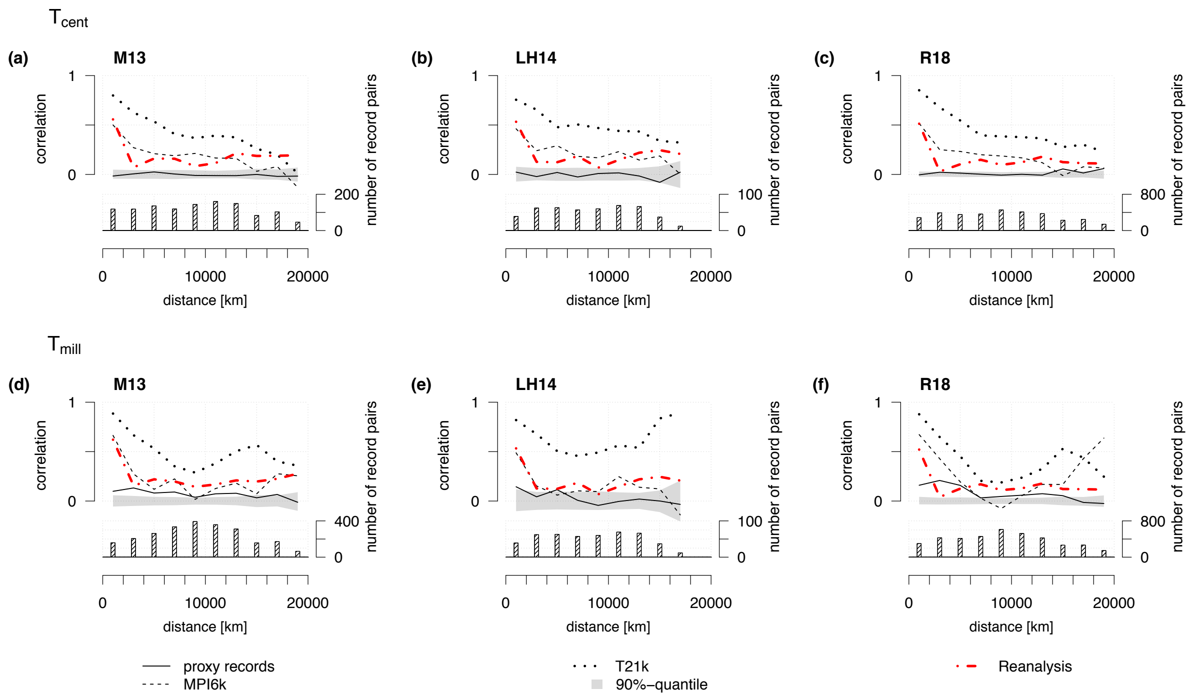

Fig.R2: Spatial correlation of reanalysis data for the time window from 1871 to 1950 (red lines). As the correlation over distance plots for the reanalysis data are very similar to the ones of the MPI6k this indicates that the spatial correlations of the model data are realistic.

2. Clarity on the separation of "multi-proxy syntheses" versus individual paleoclimate datasets and suggestions for improvement:

o In their abstract, the authors state that "The estimated low signal content of Holocene temperature records should caution against over-interpretation of these kinds of datasets until further studies are able to facilitate a better characterisation of the signal content in paleoclimate records." Here (and later on in their manuscript) the authors need to be very clear about what "these kind of datasets" mean. If they are implying that a broad-brush collation of datasets such as R18 or M13 is over-interpreted, I might agree with them that their analysis tends to demonstrate this aspect. However, this is untrue for a myriad of individual paleoclimate datasets (many of which are subsamples of aforementioned synthesis data sets) that are carefully vetted with high sensitivity to temperature and/or other variables such as precipitation, vegetation, salinity, productivity, etc. and more so, to seasonality - both aspects put together which are not addressed in this paper at all. I strongly recommend rewriting the above statement in the abstract as well as the final statement in the introduction ("more reliable interpretations of proxy records"; amongst other places) as it unnecessarily detracts from what the authors are proposing.

We agree with the reviewer and will be clearer in the abstract and conclusions to separate between multi-proxy datasets and individual paleoclimate datasets. Specifically, we will precise that 'these kind of data' are large multi-proxy and multi-site data compilations.

Such statements are also arguably misleading (e.g., modern monitoring and culturing will lead to far better interpretations of proxy datasets compared to estimates of SNR with a climate model) especially considering the point above that their analyses hinge on taking model output at face value.

We would argue that both methods, case studies such as modern monitoring, culturing, sediment traps, etc. as well as global statistical approaches such as used in this study, will lead to complementary information about the signal contained in the proxy records.
Case studies will be much more precise on the aspects they are analysing, but might omit other effects which are also present in the down-core record. Global statistical approaches include all effects influencing the down-core record, but suffer from the need to make (strong) assumptions.
Ideally, both methods converge to the same results giving credibility that the proxy system and its limitations are completely understood.

o The authors' work is a significant advance concerning model-data comparison. In its current version, suggestions on how model simulations or proxy development or the comparison of the two might be improved for better comparative metrics are lacking. I feel that some discussion on how their analyses might be developed further could be helpful.

We agree that this discussion would be useful and will add a new discussion section on implications and future steps forward.
Recent progress in computing power has enabled climate models to perform high resolution, often eddy permitting model simulations (e.g., HighResMIP project) and long simulations (>1000 year) are getting in reach. This is an important step to resolve the spatial scales and regions (mainly shelf areas and coasts) sampled by the proxies. Confronting these results with (replicated) sediment records, ideally accounting for seasonal/depth habitat using heuristic (Jonkers and Kucera, 2017) or complex ecological models (PLAFOM) would allow to better constrain the centennial spatial structures and climate variability as well as to refine the estimates of the proxy signal content shown in this study.
While our assumption of ignoring variations in the seasonal and depth habitat of the proxy recorders and the potential shortcoming in the current model correlation structure might have led to pessimistic SNR estimates, our results still underline the challenge of resolving the small Holocene temperature variations with current marine proxy records. Further improving our understanding of the proxy systems using modern monitoring, culturing and sediment traps and implementing this knowledge into ecological models (Jonkers and Kucera, 2017; PLAFOM) and proxy system models (Dolman and Laepple, 2018) is needed. Forward modelling the proxy records allows to better estimate the signal content and to optimise the sampling (e.g., replication of cores) and measurement process (e.g., sample size, number of foraminiferal tests). Although labour intensive, replicate records would allow to separate local climate variability from non-climate variability and thus provide an important step forward in understanding the proxy and climate variability.

Minor questions and comments:

- Perhaps I missed it, but why are there no counterpart plots to the T-cent in Fig. 1d-e shown in the main text for T-mill?

We agree that the naming of the figures was misleading. The aim of these figures is to provide a visual impression of the decorrelation lengths based on MPI6k and T21k. To only show one set of maps, we combined here both time scales (time scales larger than 400y, no detrending). The effect of the two time scales can be seen in Figure 3. We will change the nomenclature and describe this more clearly.

- Why does the correlation in T-mill with T21k (Fig. 3e) as well as with Uk'37 and Mg/Ca (Fig. 4b) show an uptick after 15000 KM distance?

This uptick is likely the result of the orbital forcing that is partly symmetric (effect of obliquity) and antisymmetric (precession) between the hemispheres. For the LH14 dataset (manuscript Fig.3e), the positive correlations at distances >15000km are between the tropics and the northern or southern hemisphere temperate zone as well as between sites of the northern and southern hemisphere temperate zone. There is only one time series pair with negative correlation.

- What are spatially important regions for proxy record development? Considering that the authors' work is specifically geared towards correlation distances, do their analyses pinpoint which regions are particularly data-deficient (e.g., Indian Ocean, South Atlantic, etc.) and would assist in their comparative metric?

There are some regions with a low number of sites such as the Southern Oceans. However, for this kind of study, more important than reducing the lack of single site data in these regions would be enhancing the number of replicate cores (= cores from nearby deployments that were subject to the same climate signal). This would allow to improve estimations of the signal content of proxy records and to test our understanding of proxy formation processes. This is shortly mentioned in 5.1, but we suggest to add this in the new section on implications and future steps forward.

- Is there any particular reason that the authors have not performed a similar analysis with the combined multiproxy datasets of R18, LH14, and M13?

The datasets were collected with a different focus (M13: reconstruction global temperature; LH14 and R18: temperature variability analysis) and currently use self-consistent, but different calibration and age-modelling approaches. Thus, we use them to test the sensitivity of the results on the choice of the dataset, but combining all datasets would necessitate recalibrations which is beyond our study.

- Again, I would suggest adding up front in the discussion that their analysis explicitly discounts the seasonality of proxies.

We will add that we neglected in our study the proxy-specific recording and especially seasonality. We will further discuss the effect of ignoring seasonality in more detail as suggested by Reviewer 1.

- Section 5.1: Is there a reference for anthropogenic forcing strongly increasing correlation decay length? Why necessarily, should this be the case? I feel there ought to be a statement explaining this here.

The correlation decay length observed in instrumental data (ignoring the last decades) and unforced models is largely consistent with a diffusive energy balance model (Kim and North, 1991) with increasing correlation lengths related to longer time scales (= more time to diffuse). In contrast, forced variability has a correlation length dominated by the spatial pattern of the forcing. For example, for a global forcing such as increasing greenhouse gases this leads to a globally coherent signal overlaying the internal climate variability.

This has been noted by Jones et al. (1997) and to some extend by Sutton et al. (2015), but to our knowledge there is no separate publication on this. However, it is clearly visible when analysing the decorrelation length of the surface temperature field in the reanalysis data (Compo et al., 2006) (Fig.R1). Focussing on the entire reanalysis time period results in a mean

decorrelation length of ~9150km. Contrary, analysing the time window from 1871-1950 the mean decorrelation length is ~3020km (Fig.R1), a finding consistent to the role of anthropogenic forcing.
We will add a short explanation and reference to Jones et al. (1997) in the revised manuscript.

- Although the Reschke et al. in review citation is provided, is there any reason for the 1/400y cut-off for the centennial time scale as opposed to something else?
Given a set of time series and their sampling resolution, the optimal cut-off frequency is the highest frequency that can be still resolved by the sampling without introducing a strong bias in the metric of interest, here the correlation.
Simulating surrogate records with the same sampling properties as the true records, Reschke et al. (2019) found that 1/400y is the optimal cut-off for a reasonably large subset of the data used in this study. Due to the Nyquist theorem, one needs at least 2 observations per period, and for typical non-equidistant paleo-data, four times the mean sampling frequency seems to be a rule of thumb appearing from several studies (Laepple and Huybers, PNAS 2014; Reschke et al., 2019) although this will depend on the sampling properties and thus testing this individually using Monte Carlo experiments is the safest option.

Once again, thank you for your comments,
Maria Reschke

References:

Compo, G. P., Whitaker, J. S., and Sardeshmukh, P. D.: Feasibility of a 100-Year Reanalysis Using Only Surface Pressure Data, Bull. Amer. Met. Soc., 87, 175-190, 2006.

Dolman, A. M., and Laepple, T.: Sedproxy: a forward model for sediment-archived climate proxies, Clim. Past, 14, 1851-1868, doi:10.5194/cp-14-1851-2018, 2018.

Jones, P. D., Osborn, T. J., and Briffa, K. R.: Estimating Sampling Errors in Large-Scale Temperature Averages, J. Climate, 10, 2548-2568, 1997.

Jonkers, L., and Kucera, M.: Quantifying the effect of seasonal and vertical habitat tracking on planktonic foraminifera proxies, Clim. Past, 13, 573-586, doi:10.5194/cp-13-573-2017, 2017.

Kim, K.-Y., and North, G. R.: Surface Temperature Fluctuations in a Stochastic Climate Model, J. Geophys. Res., 96(D10), 18573-18580, doi:10.1029/91JD01959, 1991.

Laepple, T., and Huybers, P.: Global and Regional Variability in Marine Surface Temperatures, Geophys. Res. Lett., 41(7), 2528-2534, doi:10.1002/2014GL059345, 2014.

Laepple, T., and Huybers, P.: Ocean Surface Temperature Variability: Large Model-Data Differences at Decadal and Longer Periods, P. Natl. Acad. Sci. USA, 111(47), 16682-16687, doi:10.1073/pnas.1412077111, 2014.

Reschke, M., Kunz, T., and Laepple, T.: Comparing methods for analysing time scale dependent correlations in irregularly sampled time series data, Comp. Geosci., 123, 65-72, doi:10.1016/j.cageo.2018.11.009, 2019.

Sutton, R., Suckling, E., and Hawkins, E.: What Does Global Mean Temperature Tell Us about Local Climate?, Phil. Trans. R. Soc. A, 373(2054), 20140426, doi:10.1098/rsta.2014.0426, 2015.