Reply to Reviewer 2. Review is (selectively) quoted in italics, with our responses interleaved

Thank you for the detailed comments. We agree with some points, but unfortunately the reviewer appears to have misunderstood some aspects of our work, which may indicate a lack of clarity in our manuscript. We will attempt to improve the explanations throughout the manuscript. We address the reviewer's comments in the order presented.

[...]However, numbers like those mentioned above matter — they get passed on in the literature for generating model boundary conditions[...]

-> We believe we have included sufficient caveats about the results such that we cannot reasonably be blamed in advance for the hypothetical argument that others may use our results incorrectly. To quote from our conclusions:

"...the different data sets produce rather different estimates ranging from 1.0 to 4.7○C for the best estimate of global surface air temperature anomaly. All the data sets are sparse with high uncertainty, and therefore our confidence in our result is not very high. We think that the regional scale information in the reconstruction is not likely to be reliable..."

"With such small data sets as we have here, the models also necessarily play an uncomfortably large role. We have investigated the effect of using the PlioMIP models themselves as a prior, versus recentering the ensemble on the data. This choice has a significant influence on the results. While in principle we prefer the data-centred approach, this is not an unquestionable choice to make."

If the reviewer has any specific concerns about the wording in our manuscript, it would be helpful for them to suggest ways that this could be made clearer.

1 [..] Despite having considerably less proxy constraints as well as different models, Annan et al. are largely content to simply follow the empirical methodology presented in AHM22 without providing supporting evidence (such as careful validation testing; see Point 2, below) relevant to the present datasets / context. [...]

-> It is precisely because of the small data set that such validation is not viable. We will explain this more clearly in the revised manuscript. With a prior predictive RMS error of about 2.8 degrees and only a dozen points, the standard error on this estimate is itself around 0.8C (implying a 95% confidence interval of plus or minus twice this value). We have already shown in AHM22 that the reconstruction is not very sensitive to parameter values

used in the method, and in that work we had 400 data points such that more modest differences in outcome could potentially be identified. Furthermore, we have already shown here that the reconstruction is highly sensitive both to the choice of data set and also to the choice of recentering (or not) the model prior. It is hard to imagine that the small changes that could arise from any reasonable changes in parameter values could alter these conclusions. We did of course perform a larger number of sensitivity tests than are presented in the manuscript but the results of these were unremarkable, as expected following those presented in AHM22, and were omitted to keep the manuscript readable. Some will be added as mentioned later in this reply

i. The authors apply a localisation of 2500km as in AHM22.

-> This comment appears to be based on a misunderstanding by the reviewer. The 2500m value is the half-width of the localisation cut-off, as was made clear in AHM22 (and also in the code). Specifically, it's the parameter "c" in the Gaspari and Cohn formulation. Thus, the reviewer's assertions on this point (and also Specific Point e later) are incorrect, as the cut-off is 5000km which while smaller than Tierney et al, still avoids there being large data voids. We will revise the text to include the explanation from AHM22, in order to avoid other readers making this mistake. The recentering process already ensures that each data point has global influence regardless of localisation in the EnKF. However in the case where recentering is not performed, we agree it would be reasonable to use a greater localisation length scale. We will therefore present the results of a test using a greater localisation length scale together with the uncentred approach.

ii Second, given the core importance of error quantification in data assimilation, the authors are rather ad hoc in their treatment of proxy uncertainties in assuming a constant 2°C error (1σ) across their proxies.

-> It is not within the scope of this work or the expertise of the primary authors to explore in detail the modelling of uncertainties in proxy analysis. We also think that it is obvious that the large uncertainty in our result, dominated as it is by the choice of data and ensemble recentering methodology, will not be significantly altered by such second-order effects as the detailed modelling of uncertainties of these data points. Code is of course available for any other researchers who wish to perform such investigations, and we believe that other researchers may be better placed than us to explore these issues.

iii Third, the authors invoke various pre-assimilation steps involving EOF's truncations and bias-adjustments of their multi-model mean. It is noted, however, that this re-centering step is not used nor recommended by the vast majority of paleo-data assimilation practitioners, given that it can precondition

the posterior result significantly (c.f., the authors' Fig. 3a and Fig. 3b vs. 3c), and thus potentially bias the answer further from the "true" state which is unknown.

-> Thank you for noting the originality of our work. There has indeed been very little discussion of the importance of the prior in paleoclimate reconstructions of this type, a regrettable state of affairs that the authors accept some blame for. Indeed our first work in this area (Annan et al, Scientific On-Line Letters on the Atmosphere, 2005) used a single model with varying parameter values in an attempt to represent uncertainties in climate feedbacks. It was only subsequent to this work that we came to more clearly understand the limitations of such an approach. While we have subsequently shown that multi-model ensemble provides a more robust approach, we also demonstrated in AHM22 that any significant biases in such a multi-model prior would pass through in the posterior, even in that scenario where we had 400 data points distributed widely over land and sea. See Sections 5.2 of AHM22 for analysis and discussion of this issue, and also Section 6 of that paper for further comparison with Tierney et al 2020. With only at most two dozen data points, it is inevitable that the choice of prior is a critical factor in this current work and the use of an ensemble of convenience simply because it's what everyone else does is not tenable. We had hoped that this point was already well enough made in AHM22 but perhaps it bears repeating in this manuscript.

We note that the reviewer does not present any scientific arguments in favour of using the uncentred multi-model ensemble, let alone the single model ensembles that are still sometimes used in this area of research. Our method did not appear from a vacuum but rather through a critical analysis of previous work, including our own. We hope that other researchers working in this area will follow our lead in considering more carefully the sensitivity of their results to the priors that they use.

We already emphasise in the manuscript that the reconstruction is strongly dependent on the recentering decision. If other researchers have reason to prefer the uncentered ensemble, that option is available to them.

2. The authors do not validate their assimilated results using withheld or independent data.

-> We return to the point made previously, that with only a dozen or so data points (a maximum of 23), there is not any hope of meaningful validation of the approach in this application, which is why we rely on the validation performed in ANH22. We will explain this point in the revised manuscript. The EnKf itself is of course decades old and well established. We are not presenting further methodological innovations here, merely applying the

approach of AHM22 to a different time period.

Specific points:

a Yet, this is exactly what I feel this manuscript should be exploring.

-> However, this is not the area of expertise of the primary authors. As in our replies to reviewer 1, we can highlight some of the reasons for these discrepancies in a revised manuscript but this remains a highly active area of research. A recent review showed that the absence of multi-proxy single-site analyses has significant impact on addressing this issue (McClymont & Ho et al., 2023). We are aware of ongoing work which is specifically seeking to understand the apparent cold bias in Mg/Ca Pliocene temperatures but this work is early in its development and not available for discussion here. Many of the parameters which could explain this bias (e.g. salinity, seawater chemistry, carbonate dissolution, seasonality, depth habitat) are even less well constrained for the Pliocene, and in many cases can't yet be quantified (McClymont & Ho et al., 2023). We re-emphasise here that our focus is to explore the impact of proxy choice and site distributions on data-model assimilation, and that by showing this impact we hope that this will motivate further work to investigate why.

b

-> The same comment as for (a) above applies.

We agree with the reviewer that by by averaging the two alkenone data sets we do intrinsically reduce the low latitude error (in that BAYSPLINE is more like a 4*C uncertainty) whereas in the high latitudes we're creating a value which sits between two calibrations, even though the difference between them is <0.5*C (with an overall calibration error more like 1.5*C). As per our reply to the previous comment, we can highlight this issue in more detail and explain the rationale for the omission of BAYMAG data.

c

-> The suggestion of ad-hoc movement of data points does not seem entirely consistent with the reviewer's complaints regarding a number of decisions we've already taken. Where models cannot resolve coastal areas adequately, model-data comparison is always going to be challenging. Furthermore, some of the omitted data points are those in the Benguela upwelling area where where there are significant issues.

d

-> We will change the wording

e

-> The 2500km issue rears its head again. The reviewer's comments about a substantial part of the globe being unaffected is incorrect. We will present a test using a longer length scale to demonstrate that this issue does not affect our conclusions.

f.

-> The argument is not circular when we perform the comparison to the model prior. The argument is based on the simple observation that the data errors cannot plausibly be greater than the (prior) model-data difference, since models also have errors when compared to the unknown truth, and the modelling errors can be reasonably assumed independent of data errors. Similarly, pairwise model differences provide some evidence as to the magnitude of model errors, though this can only ever indicate a lower bound on such errors, as we cannot reasonably assume model errors are independent across the model ensemble. Of course sampling errors also limit the precision of what can be reasonably concluded from these analyses, but they are still relevant information.

g

-> We will add pictures to the supplementary information

h

-> We will add the values to this table.