Reply to Referee #1

Authors

May 2022

Thank you for the detailed and interesting questions and comments. To aid readability of our reply, specific changes proposed for our manuscript will be *highlighted in red*. Reviewer text quoted below is *highlighted in blue*.

One major change in response to major point 2 that should be highlighted at the outset is that certainly *we will make our code and results available*. Therefore interested researchers will be able to check our result and further test it with their own calculations and diagnostics in any way they wish. This should also help with the various comments about the details of our approach, the interpretation of results, and possible further tests.

1. There are too few plots and the discrete color scales with often wide temperature steps make the current plots not very informative. [. . .]

The number of plots will increase significantly in our revision, with some specific additions mentioned here and also in our response to the other reviewer. However, we believe that making our code and output available should satisfy most requests more effectively than producing a large number of figures that we believe will be of little interest. Perhaps we have misunderstood your requests but there would be 378 figures of pairwise model differences and 405 plots of leave-one-out validation tests. Both of these numbers would double if we included both SST and SAT. We are thinking that this is a rare case of a few words being worth a thousand pictures.

As for the colour scheme, it was originally used in AH2013 and was based on the draft of IPCC AR4 which was circulating at that time. We actually thought our nonlinear scale with 1 degree bins close to 0C added useful clarity compared to the 4 degree bins (+2 – -2C, -2 – -6C etc) that the IPCC used. 11 colour bins doesn't seem particularly out of line with other similar papers, eg Fig 2d of Tierney et al and Figure 4e of Cleator et al. We agree that the figures are challenging to interpret precisely and spent some time working on them but we are not experts in graphical design and since numerical output will be provided, the figures are only needed as a summary for those who do not require high precision. We hope that supplying the code and data will satisfy most of your criticisms as this will enable all researchers to perform further calculations and re-plot the data in their preferred style.

2. A statement on data and code availability is missing.

Yes, we apologise about this and *will include a statement to the effect that data and code will be available as supplementary information*. We were originally planning to make these available at the revision stage, but clearly we should have done so for the initial submission.

3. The used error metrics in the validation and sensitivity sections are strongly focused on the posterior mean and the global mean temperature. As this is a Bayesian reconstruction of climate fields, a stronger focus on the spatial structures of the reconstruction (and how different choices discussed in the sensitivity tests influence it) and of the full posterior probability distribution would be more informative (e.g. maps of cross-validation results, coverage frequencies to study the meaningfulness of the posterior uncertainties, probabilistic score functions). The included proxy data tends to be spatially clustered. Therefore, I wonder how much leave-one-out cross-validation is influenced by spatial autocorrelation and whether leaving our more data at once would be a better choice (e.g. h-block or leave-N-out cross-validation).

We have already undertaken 'extensive sensitivity tests' (in your own words below). We have specifically tested omitting large subsets based on TEA, MARGO and Bartlien et al which we consider to be a particularly strong challenge as these different researchers have collated and calibrated their data in different ways. *We will add a figure showing a rank histogram of the leave-one-out validation results.*

4 The abstract is too short and not very informative outside of the stated global mean temperature anomaly. I recommend to specify the used proxy data, simulations, and methodology. Actually describing the results from investigating the differences compared to Tierney et al. (2020) would be more informative for the interested readers than just writing "We discuss the reasons for this discrepancy".

We will improve and extend the abstract, including a more explicit outline of our method and comparison of our results with other work.

5. The explanations on how statistical and data processing choices influence different features of the reconstruction are very valuable and should be a useful guide for future applications of data assimilation in paleoclimatology. However, they are often very general(e.g. p. 1l. 2-3,7;p. 2l. 8-9;p. 3l. 10,26;p. 5l. 16;p. 6l. 12-13;p. 8 l. 9-10; p. 11 l. 12-14; p. 14 l. 4-5; p. 15 l. 5, 7-8, 19-20; see also the specific comments below) such that readers are forced to believe the authors and cannot trace the explanations in the results. Therefore, I recommend to specify these explanations.

We will improve the text, and code will also be made available which clarifies the details. Our results are rather insensitive to the choice of various parameters in the scheme and the final choices made are somewhat subjective.

Specific comments

Replacing "climate" by "temperature" in the title would be more precise.
Agree.

p. 2, l. 8-10: What are examples of absent localized features and suspected small-scale artifacts? Specifying the magnitude and spatial-scale of these features would be valuable to better interpret the results.
We will describe in more detail along with the comparison to previous work.
For example, there was (what appeared to be) a notable artefact in the Southern Ocean of the AH2013 reconstruction, remote from data.

p. 1, l. 18-19 (and several subsequent instances): In which sense are the resulting fields 'physically consistent'? While the individual ensemble members are physically consistent, it is not clear to me how physically consistent the fields are after de-biasing and applying the ensemble Kalman filter.
Consistent in the linear sense, an approximation which is intrinsic to the Kalman Filter (along with the assumption of gaussianity). The contrast we are drawing is with simpler assimilation schemes (and ordinary statistical interpolation approaches) which don't achieve this. *We will clarifiy the wording*.

p. 3, l. 10: What does "unusual ocean boundaries" mean? Has this been reported elsewhere? Given the numerous papers employing the PMIP3 ensemble, this finding of potential bugs in these simulations sounds relevant for others.
We doubt that this is a bug in the implementation of the PMIP protocol or the model simulation itself. What we have is model output downloaded from ESGF on an ocean grid that does not interpolate well onto our 5 degree grid, losing a lot of coastal area in the process. Researchers who are not using similar gridded (primarily coastal) data probably won't notice a problem. It may be that more sophisticated processing (we used standard routines in CDO) could resolve the issue but, as we already had alternative models from these centres, this did not seem worth pursuing.

p. 3, l. 26: What was used as cutoff values for RMS / pattern correlation? Giving more details seems necessary to interpret / reproduce the present details.

We did not use a fixed cutoff, but selected the models as described. We approached this decision (and many others regarding the detailed implementation of the algorithm) as a fundamentally subjective one, with tests and calculations used as a guide rather than a strict but arbitrary rule. Making code available will allow others to test different choices if they wish.

What is the GMST anomaly of this outlier? What is the difference to the

closest ensemble member? Given how much the GMSTs are used throughout the manuscript, plotting the GMST anomalies of the ensemble members (selected and removed ones) would be useful for the readers.

We will include histograms of the GMST of the 28 and 19 member ensembles, and the posterior. That model has an anomaly of about -11C, with the others ranging from around -3C to -7C.

How much of the variance is explained by the first four EOFs? How close is q to the final posterior mean?

Values will be included.

Cleator et al. (2020) augmented the Bartlein et al. (2011) data by Australian records from Prentice et al. (2017). These ones could be added to the dataset employed here. Given the new temperature reconstructions from ice cores that have been produced over the last decade, are the values from Schmittner et al. (2011) still up-to-date?

We are reluctant to go down the route of chasing up small updates and changes, especially as we are not expert in such data analyses and compilation ourselves. While we do seek to use recent, large, and openly available datasets, such that our results are credible, our main focus is on the assimilation of data and models, which we believe is an important and sometimes under-recognised skill in itself. We do hope that others with greater expertise in climate proxy data may be able to implement our code incorporating their own data.

Is the selection of TEA over MARGO based on the grid box averages or on the level of individual records? Given that TEA uses different archives than MARGO (inclusion of d18O in TEA, use of foraminiferal assemblages in MARGO) there might be some inconsistencies when the selection of MARGO over TWA is performed on the grid box level, in case there are systematic differences between different proxy types.

All of our work is at the grid box level, and indeed these authors quite possibly (probably) interpreted the same data in different ways. Where the data sets coincided, we used TEA in preference to MARGO, which avoids any possibility of double counting.

Our tests of the TEA and MARGO data sets, both in the assimilation and also via a direct comparison at the gridpoint level was specifically aimed at checking for inconsistencies. We did find that TEA represents very slightly cooler conditions, but the discrepancy appears small. However the RMS difference between the data sets was significant, a point of evidence that we use and discuss in our estimation of observational uncertainties.

How is the specific implementation of the ensemble Kalman filter selected? How does it compare with the ensemble square-root Kalman filter employed in Tierney et al. (2020)? Are there previous examples of applying ensemble Kalman filters to multi-model ensembles? Are the assumption of it still satisfied

for multi-model ensembles?

We are not sure if the EnKF has been applied to multi-model ensembles, but it certainly has with perturbed physics ensembles. The linear and gaussian approximations are not strictly true for any complex nonlinear situation, including numerical weather prediction where this approach has a long and successful history.

Our specific choice of algorithm was made for convenience and familiarity. TEA contains references for their algorithm. It is a slightly different implementation of the same fundamental equations (i.e. the Kalman equations).

p. 14, l. 4-5: How is "no sign of over-fitting" measured? What would be considered over-fitting?

Over-fitting was checked by looking for a worsening of the fit to withheld data as the number of EOFs was increased (and this did not happen). See AH2013 for previous analysis, though this used the model fields for the fitting rather than an EOF decomposition. The choice of number of EOFs to use is subjective, as is the case for a number of other parametric choices we made. Since we do not believe that there is an objectively correct way to perform the analysis, we tested here and elsewhere that the results were robust to the choices made.

It would help me if the rank histograms were shown and not just described. We will include these rank histograms.

p. 14, l. 27-28: Which numbers of EOFs were tested? Assuming that the general cooling pattern is mostly contained in the first EOF (as they are not shown, I cannot determine that for sure), it would likely be more interesting to compare not just the GMST but the spatial patterns.

All values from 1 to 8, and also the full set. With the uncentred approach, the first EOF is close to the ensemble mean and the subsequent ones represent variability at increasingly finer scales. There is no obvious physical interpretation however.

p. 15, l. 5: How is the worsened fit to the data quantified?

By the RMS of the residuals.

Given the extensive sensitivity tests, are there general recommendations for the future usage of PMIP ensembles in climate field reconstructions and potentially for the design of future PMIP cycles? How could/should multi-model ensembles be designed for effective usage in climate field reconstructions?

We will add some discussion about this. It has been a frequent topic of discussion within PMIP meetings but may be useful to a wider audience. Design of experiments needs to sample the relevant uncertainties, which in the case of paleoclimate will normally include forcings and other boundary conditions, as

well as uncertainties in physics including feedbacks which are sampled across the ensemble of models.