Reply to Referee #2

Atsushi Okazaki, Diego S. Carrio, Quentin Dalaiden, Jarrah Harrison-Lofthouse, Shunji Kotsuki, Kei Yoshimura

We sincerely thank the reviewer for the positive assessment and the helpful comments and excellent suggestions. Below, we provide a point-to-point responses to all the reviewers' comments. The reviewers' comments are in *blue and italic*, and the replies are in black.

This paper provides an important contribution to our understanding of the uncertainty parameter (R) in paleoclimate applications of DA. As the authors note, this is an often-minor consideration of previous researchers, but nonetheless an important parameter. The innovation statistic application that is tested provides a useful alternative to the linear regression methods which requires significant overlap with 19th-21st century climate observations, but future tests will eventually be required to ensure that the proposed method is skillful for deep time reconstructions. The manuscript is well-structured and methodologically sound, and I recommend minor revisions prior to publication to further strengthen the clarity and accessibility of the work, as well as to understand impacts to the posterior ensemble.

Specifically, the text could be improved by additional plain language description of the innovation statistic method (see comment for line 126) and impact of age uncertainty (see comment for Line 554). Furthermore, an important role of R in paleoclimate studies is to properly quantify changes to the ensemble spread, and it would be beneficially to include additional analysis or commentary on how the innovation statistic impacts the ensemble range rather than just the mean.

Thank you very much for your detailed review of our manuscript. The responses to the points raised here are written below.

Line 126: This paragraph would benefit from expansion to provide a plain language summary of innovation statistics. The methods section is quite technical and will be difficult to follow for readers who are unfamiliar with the method.

Thank you for the valuable suggestion. We will include a brief explanation of the innovation statistics in the revised manuscript.

Line 144: Could you provide additional clarification on the difference between the LETKF and an EnKF implemented with a localization radius.

Thank you for your comments. Simply put, they are a totally different idea. LETKF solves the update equation locally to enhance the computational efficiency, while the localization is a technique to mitigate the detrimental impact of the limited ensemble size.

Line 203: Please explicitly state whether the OSSE is equivalent to a pseudoproxy experiment, or explain the differences if not.

They are the same in the context of paleo-DA. We include this in the revised manuscript.

Line 215: Does "MIROC5" refer to "MIROC5-iso" or a different simulation?

They are different. "MIROC5-iso" is a name of the model we ran for this study. "MIROC5" referes to the model which participates in CMIP5 and provides data.

Line 216: Why was only r1i1p1 used to create the prior? If this is the CMIP5 MIROC5 simulations, wouldn't more ensemble members be available?

Yes, there are five ensemble runs in total. The purpose of using the two runs is to create an idealized situation, where the nature run and the prior model simulation are similar statistically. For that purpose, we used one for the nature run and the other for the prior. Therefore, we do not need more ensemble runs even though they are available.

Line 221: Were proxies records filtered to span a certain amount of the 1870-2000 study period?

Yes, the proxies and the model simulation have to overlap longer than 30 years. This treatment is necessary to calculate the anomaly of each. We will add the explanation to the revised manuscript.

Line 224: What do you mean by "complementary"?

The purpose of the paper is to estimate the observation errors of the climate proxies. The temperature data is instrumental data and not a climate proxy in this regard. This is the reason why we used the word. We will rephrase it in the revised manuscript.

Line 224: It's unclear why just the documentary data were used for temperature.

Probably, there is a misunderstanding. We used surface temperature data recorded in the historical documents. The other proxies are also used for temperature reconstruction.

Line 229: Could you describe the linear interpolation method. Is this an interpolation between two grid center points? How does this work in two-dimensional space?

Thank you for raising this point. We calculated the weighted mean of the adjacent 4 grid points for the 2D interpolation. Specifically, we used bilinear interpolation. We will specify the interpolation method in the revised manuscript.

Line 248: What metric(s) was optimized that resulted in a half-localization scale of 8,000 km?

Thank you for pointing it out. We optimized the localization scale with the correlation. We will add the explanation to the revised manuscript.

Line 295: Do these skill metrics consider the ensemble spread or just the ensemble mean?

Thank you for the comment. We do not evaluate the ensemble spread in the manuscript. But, the discussion on RV is tightly connected with the ensemble spread. When RV is small, the observation error is large as discussed in Sect. 3. This means that DA weighs the simulation more, implying relatively small ensemble spread. In the revised manuscript, we will add more discussion about the ensemble spread.

Line 295: Do these skill metrics consider the spatial correlation or interannual variability? If the former, how does the innovation statistic impact interannual variability in the posterior?

Thank you for your comments. The metrics are calculated at all the model grid points, and then averaged spatially. Therefore, the correlation evaluates the interannual variability. We will add the explanation to the revised manuscript.

The innovation statistics change the observation error and consequently change the interannual variability. The reason is that DA combines a simulation and observations based on the corresponding errors. With larger observation errors, DA weighs the prior simulation more, and vice versa. The interannual variability of the analysis changes depending on the weights (errors), unless the correlation between the observation and the simulation is 1.0.

Line 315: Please clarify the units within Figure 2.

Thank you for pointing it out. The units are Celsius degree or permil. The units will be added in the revised manuscript.

Line 378: Is this because no PSM was applied?

Thank you for raising the important point. It is possible that the SNR for ice cores are low because no PSM was applied to them. However, even with PSM, the estimated SNR should be smaller for ice cores because the skill of the PSM is still relatively low compared to the ones for coral and tree-ring cellulose (Okazaki and Yoshimura, 2019). We will add a discussion on that in the revised paper.

Line 380: Also important to note that the small sample size of 3 records.

We believe the small number of proxy points is not the cause of the small SNR. For each proxy point, more than 100 samples are used to estimate the SNR for each proxy point.

Line 523: Please clarify what the 5%-30% improvement is measured against.

They are measured against the reconstruction without observation error estimation.

Line 554: Age uncertainty is a very important consideration for deep time. Not only is the exact date uncertain, but also the amount of time that each measurement represents, which will impact the variance and therefore the estimation of R. Given the authors highlight deep-time applications as a key motivation, a more extensive discussion — or a small pilot analysis (e.g., assimilating non-annual records (i.e., speleothem) with age uncertainty) — would strengthen the case for broader applicability.

We appreciate your important suggestion. Although we strongly agree with the idea that an additional analysis would enhance the value of the study, we would like to keep this as a future work because it is far beyond our scope; we need to develop another method to deal with the temporal uncertainty. Instead, we will expand the discussion in the revised manuscript.