Comment on egusphere-2022-184
Jessica Tierney

Erb et al conduct a new DA of the Holocene using the Temp12K database. As an advance over previous work (i.e. Osman et al), they include terrestrial temperature data. Generally they find a fairly flat temperature trajectory across the Holocene, with less overall variance in global T than previous reconstructions.

I think this is a useful contribution and it’s interesting to compare/contrast this result with Osman et al., 2021 in particular. However it remains unclear to me why the Erb et al reconstruction has less variance than previous work. It would be helpful if the authors could identify why this is the case. I wonder whether it has to do with the technical choices made in the DA. Although the testing in the Appendices suggests that the same result is coming out under different experiment assumptions, I wonder if perhaps the use of interpolation on the proxy data is playing any role (see 2. below).

Here are some general comments:

1) I am not clear on why the authors chose decadal as their base resolution. The mean temporal resolution of Temp12K is 200 years, as stated in Appendix B.5. This is why, in Osman et al., we chose 200 years as our reconstruction bin size (90% of the data have this resolution or better). Perhaps some of the terrestrial data include finer-scale records but for lakes, even if the sampling resolution is decadal is the lake is not varved there is no way that represents the actual age resolution because bioturbation would smooth the signal. Is there really any recoverable information in the proxy database below 200 years? The authors need to justify their 10-yr choice, especially given that Appendix B shows that 200-yr bins give essentially the same result. I’m just concerned that users of this DA will think that the decadal resolution of this product is “real”... when effectively it is not given that most of the underlying data does not have information at this timescale.

2) The proxy data information is interpolated and I wonder if this impacts the DA results. I'm reading correctly, it seems like the data are interpolated annual resolution then re-binned to 10-years in all cases? (Line 206)? Any sort of interpolation in my view is unideal, because you effectively introduce imaginary proxy information (that assumes the climate trajectory between data points is linear). Moreover I see that Line 224 indicates that proxy data is used repeatedly for all time steps that it spans. I think this could potentially contribute to the relatively flat posterior that the authors get, although I'm not
Binning is a more robust approach because it ensures the proxy information is only used in one time interval (hence binning has been used for many of the PAGES reconstructions). I see that binning was testing in Appendix B.5, but it looks like the proxy information was still interpolated in this case, hence the information is still used in multiple bins. Instead, put the proxy data into only one bin, but then resample age models and thus shuffle proxies between the time-bins (as we did in Osman et al) to account for the fact that you don’t know exactly the age of (most) of the data. That would be a better test and would also assess whether the interpolating is biasing the result at all.

3) Age model uncertainty. I don't think it is accounted for? I should be included as all the marine and lake records (except varved lakes perhaps) will be radiocarbon dated with uncertainties of at least several decades. You can simply do ensemble DA experiments to sample the uncertainty (this is what was done in Osman et al.). Age model uncertainty is another reason why a 10-yr bin is too optimistic IMO.

4) Is the multi-timescale DA approach needed? As I read it this simply uses a different covariance based on each proxies' resolution, but I’m not clear how it is appropriate to do that when your base time step is always 10 years. If that's the case, then it seems to me that a 10-yr covariance should always be used. Regardless, I’m not convinced it is actually needed because, model covariance 10-years and longer tends to be very similar (see the experiments I did in my 2020 LGM paper testing the DA with different time-averages). When most of your proxies are low-res, you might as well use just a standard time for both the assimilation and the covariance.

Accordingly, Appendix B.5 suggests that the multi-timescale method isn't adding much. So there needs to be some justification for using it - what is the added value?

5) Regarding the paper's claim of the DA having mid-Holocene “warmth”: I wouldn't call an anomaly of 0.09°C as evidence of warmth. seems like a stretch - and not significant within uncertainties. Considering the uncertainties, this DA shows a flat Holocene trajectory since 7 ka and that's how the results should be described (unless 0.09 can be proven to be statistically significant).

6) Finally, in addition to the lack of incorporation of age uncertainty (which I think can and should be addressed) the authors should be clear in the manuscript that a major caveat of their DA is that they are not using proxy forward models. These models are available and out there, and indeed we used the marine ones in Osman et al. I understand that it is easier to work in temperature space but the reality is that many of these proxies (pollen, marine Mg/Ca and d18O, ice core isotopes) are multivariate and not exclusively sensitive to temperature and working in T space ignores these aspects of the proxies.

Here are some specific comments:

Line 65: I would be more specific and say the main difference is the inclusion of terrestrial proxy records.

Line 80: reiterate what the age control and time-resolution criteria are for inclusion

Line 98: since this is relevant for your choice of base resolution (decadal) show a histogram (like ED Fig. 1 in Osman et al) of the proxy time resolution.

Line 99: this isn't a good assumption -- it is probably not true for most of the data. Researchers typically sample at discrete depth intervals (e.g., 0, 4, 8, 12 cm) and do not sample cores contiguously. You could check this by investigating the depth sampling
intervals in Temp12K. This is another reason that arguably the best approach to doing DA on this timescale is to bin the data rather than interpolate.

Line 126: "minor checkerboard"?? That sounds like a modeling artifact. Can you elaborate, and would this affect the DA? Introducing spatial model artifacts would certainly mess with the covariance.

Line 137: If I'm reading this correctly, you are stating that decadal resolution is equal to or higher than the mean resolution of most of the proxy data? If that's the case, then decadal is not the right choice.

Line 158: Lack of robust PSMs? This is offensive to those of us who have spent a large portion of our research career developing forward models for temperature proxies. It's not just me either. The Mg/Ca community has made major strides in this, as has the pollen community (see for example Parnell et al., 2016, QSR: https://doi.org/10.1016/j.quascirev.2016.09.007), as has your co-author Sylvia Dee. You didn't do the DA in proxy space because it was easier to do it in temperature. That's fine, but this is a caveat of your DA, so just admit that rather than erroneously claiming that the forward models weren't there (they are).

Line 515: mid-Holocene anomalies are usually calculated across some kind of time window (i.e., 5-7 ka minus 0-2 ka). Here is says 6 - 0.5 but is there a window around those centered values? Presumably yes, since mid-Holocene is not a single instance in time. Up to you what window to use but since the other reconstructions are lower res I would do at a minimum 5.5-6.5 (1000 years) minus 0-1 ka or something like that.

Line 518: Is the 0.09 "warmth" statistically significant when compared to the baseline (presumably, last 2K)? It doesn't really look like it is based on Figure 12. Or else marginally. Provide a p-value. Also I would calculate this over a larger window as suggested in the previous comment.

Line 537: Reference here some of the work we did in Osman et al to try and get at the origin of this difference.

Line 576: To some extent you have this metadata in the form of the depth intervals that researchers averaged over for their analyses (in the ideal case, researchers should list both the starting and ending depth of their data point, but they do not always do this).

General comment on localization (Appendix B.2): The results here are consistent with what we have found, which is that some localization improves the reconstruction. While I understand the argument not to localize, on the flip side, localization was designed in the first place to eliminate spurious covariance structures. So why not use some localization in your reconstruction? I think you need to justify, in the main text, why localization is not used.

Line 835: This is incorrect. We have implemented localization in the joint update and that is what is used in both Tierney 2020 and Osman 2021. c.f. the DASH codebase on GitHub if you want to investigate the method.