Review of Baudouin et al (2024)

For questions regarding this review, please contact bryan.lougheed@geo.uu.se (until 31/08/2024, thereafter @outlook.com).

Summary of review

This paper further investigates an existing algorithm for global mean surface temperature (GMST) that was developed by Snyder (2016) using existing palaeoclimate data. Here, the authors compare the data-based GMST reconstruction by Snyder to a model-based reconstruction of GMST. Specifically, the authors have taken climate model output and passed it through proxy and sediment models, thus producing a type of synthetic palaeoclimate data, which can then be compared to the Snyder analysis of palaeoclimate data.

In general I agree that processes in the sediment archive itself (such as bioturbation) were underestimated by Snyder, so an approach using synthetic palaeoclimate data is useful. In general I find the results to be interesting and the sty to be a useful addition to the literature, but I think the paper could due with some re-writing so that it states what is being done in a more straightforward and accessible way. This would especially benefit readers who are less well familiar with the subject matter. I found the reading quite heavy in places and not so logically organised. There are also very many acronyms.

I also think that the study can be better framed in the abstract and introduction. In particular much of the paper is framed from a point of view of evaluating the Snyder et al GMST method by comparing it to the authors' synthetic palaeoclimate data, thus assuming the authors' data represents the truth. But in reality both methods represent an approximation, and we don't know which is the "truth". I agree that the method of the authors attempts to include more processes, but a more accurate framing of the paper would be to compare the two methods, rather than to use one method to evaluate the other.

Also, throughout the paper I found that the citations skew to the very recent past, with many older papers being forgotten.

Specific comments (authors' text is in blue)

Lines 1-3: Reconstructing past variations of the global mean surface temperature is used to characterise the Earth system response to perturbations as well as validate Earth system simulations. Reconstructing GMST beyond the instrumental period relies on algorithms aggregating local proxy temperature records.

Perhaps this can be written more clearer, because GMST from the instrumental period also requires an aggregation algorithm due to the non-uniform spatial distribution of thermometers on the planet.

Lines 4-5: Here, we propose to establish standards for the evaluation of the performance of such reconstruction algorithms.

In my opinion the above text does not correctly characterise the work. Firstly, in this study you are essentially comparing palaeoclimate data (from the Snyder approach) to palaeoclimate model output, so it is not strictly possible to independently evaluate either the data or the model output. We don't know if either the data or the model is correct, or if both are correct, or if neither are correct. Perhaps a more accurate statement that characterises the work would be that you "investigate the the level of agreement between data and model", which is of course a valuable exercise. As for

"establishing standards", I don't agree with this statement. The authors put forth an interesting approach for a data-model comparison, but I don't know why it should be a standard. Other authors may use different approaches, would their approach then be non-standard?

Lines 10-11: We find the algorithm to be able to reconstruct timescales longer than 4 kyr over the last 25 kyr. However, beyond 40 kyr BP,age uncertainty limits the algorithm capability to timescales longer than 15 kyr.

Do you mean: "temporal resolution of the algorithm is limited to 4 kyr for the last 25 kyr"? (This assertion is of course based on the assumption that the synthetic palaeoclimate data represents the truth, or at least a more complete effort to quantify of the truth).

In Lines 75 to 85 the authors explain why they use the PalMod database of proxy data as opposed to the original dataset of Snyder. I guess the main reason is to us a newer dataset with more data than the Snyder one (2016 was 8 years ago now). I don't agree with some of the reasoning given regarding superior chronological control in PalMod. Yes, Bacon was used to construct 14C + d18O tuning age models, but Bacon is only as good as what is put into it and is known to underestimate the total age population contained in multi-specimen discrete-depth sediment samples from deep-sea sediment (Bacon was originally developed for lacustrine sediment). Age-depth points based on d18O here are base based on visual matching benthic d18O data to regional benthic d18O curves (Lisiecki and Stern, 2016), which are themselves dated based on visual matching benthic d18O data to Greenland and speleothems, so we can say the age is "double tuned", with double potential interpretation error based on visual matching. Furthermore, there is a certain assumption of global synchronicity between all these records when tuning. Such approaches seems to be pretty standard in many palaeoclimate papers, so I don't wish to single out the authors in this case, but in a paper that seeks to quantify all sources of error, I believe these potential pitfalls should be pointed out clearly.

As for 14C, 14C in age-depth models in bioturbated archives can display large age-depth artefacts during periods of highly dynamic D14C (such as the last deglaciation) due to forams with very different 14C activities being combined into the same sample (Lougheed et al, 2022; https://doi.org/10.5194/gchron-2-17-2020). This uncertainty could possibly be also mentioned and considered...

Greenhouse gas concentrations are prescribed following the measurements from the EPICA DOME C ice core (Lüthi et al., 2008),

Last glacial EPICA Dome C is largely dated by making assumptions about local temperature and obliquity (Parrenin, et al., 2007; doi: 10.5194/cp-3-485-2007). Therefore, the greenhouse gas data from EPICA Dome C is unfortunately not independent of palaeoclimate assumptions, with consequences for LOVECLIM runs forced by EPICA Dome C greenhouse gases.

Comments on section **3.2.1 Construction of pseudo-proxies with sedproxy**

Please mention and cite the bioturbation model that is used within sedproxy (Berger and Heath)

One of the major issues with deep-sea sediment records based on multispecimen foram samples is the interaction between bioturbation and temporal abundance (e.g. Löwemark et al 2008; 10.1016/j.margeo.2008.10.005), meaning that centuries and/or millennia of high abundance are overrepresented in the sediment archive. Does the approach using FAME correctly account for granular changes in foram species abundance as a fraction of the total sediment flux?

Comments on section **3.2.2 Specification of the pseudo-proxy experiments**

I think it is good that the authors try to do many experimental setups, but at the same time, the tradeoff is that I as a reader I start to lose the ability here to keep track of them all. Admittedly, I have not slept properly in recent years.

Comments on section **4 Specification of the pseudo-proxy experiments**
It is unclear to me here what is being discussed here and shown in Fig 3.

If I am correct, the aim of this study is to compare the Snyder algorithm GMST to the pseudoproxy GMST, right?

Fig 3 just shows the ensemble results from the pseudoproxy GMST, right? Not the Snyder algorithm. Yet the text in the paragraph around line 300 refers to Figure 3 in the context of the "ability of the GMST reconstruction alogrithm".

I think I figured out now what is in Fig 3… for example in Fig 3C, is the black line the Snyder approach and the red line the ensemble mean from pseudoproxy approach? (Or is it the other way around?) So essentially the the text is referring to the agreement between the red and the black line (for Fig 3C).

Perhaps clearer and more consistent language should be used here, because my main issue is figuring out what is what ("reconstructed GMST" and "simulated GMST"). Use "PPE GMST" and "S16 GMST" throughout the text and the caption here? Although this risks introducing more acronyms.

Line 316: "One of the most important contributors to the uncertainty estimate is the scaling factor."

Again here, when referring to the "uncertainty estimate" it is unclear at first reading if you are referring to the Snyder approach or to your PPE approach.

Line 372: We also estimate an upper bound to the effect of seasonality bias on the pseudo-proxy reconstruction. If all records were not recording the mean annual temperature, but that of the warmest month, it would generate biases up to 0.75 K. This bias mostly follows the northern hemisphere insolation curve:

What is the northern hemisphere insolation curve? Be more specific.

Line 375: These orbital-scale biases can significantly impact the evaluation of orbital timescale variations

What is orbital-scale and orbital timescale? Obliquity, eccentricity and longitude of perihelion are changing as I am typing this sentence. Perhaps be more specific as to the exact timescales you are referring to. Half an obliquity cycle? (20.5 kyr?)

This warm season bias has, for example, been one of the main hypothesised reasons for the model-data discrepancies during the early Holocene (Liu et al., 2014; Marsicek et al., 2018; Bova et al., 2021), although there is also evidence for cold season bias for other periods and species (Steinke et al., 2008; Timmermann et al., 2014).

Bova et al essentially detrend their data for general precession, which is one of the major drivers of Quaternary global climate.

Line 560: First, the age uncertainty is the limiting factor preventing the reconstruction of multi-millennial timescale beyond 30 kyr BP, and the main focus should be put on reducing this uncertainty (e.g. Waelbroeck et al., 2019; Peeters et al., 2023).

I think all Earth Scientists would like to see reduced age uncertainty in data, I'm not sure if those two papers were the first to point it out. In particular the Waelbroeck paper concentrates on better implementing age uncertainty… in some cases this actually *increased* the age uncertainty over the original datasets. So I would say the main focus should be in quantifying age uncertainty. If it gets reduced then that's a bonus.

Comments on section **5.3 Improvement of the algorithm**

Once again, here the paper is being framed as a way to evaluate an algorithm, by comparing to how it compares to pseudoproxy data developed from climate model runs. This assumes that the latter represents the "truth" and that the algorithm must be evaluated against this truth.