Short summary on anticipated changes to the manuscript:
- We will clarify in the abstract, introduction and methodology, that we focus on evaluating a reconstruction algorithm, and for that only use the metadata from the J20 dataset to compute pseudo-proxies.
- We will enhance the discussion with the points the reviewer suggested and correct the various imprecisions that were noticed.
- We will add the flowchart as suggested in the second reply.

Point by point response to Bryan Lougheed:

Lines 1-3: 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.
- "Beyond the instrumental period, reconstructions rely on local proxy temperature records and algorithms aggregating these records."

Lines 4-5: 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?
- The misunderstanding has been clarified in the previous short reply (https://doi.org/10.5194/egusphere-2024-1387-AC1). In short, we use model data to evaluate the reconstruction algorithm. Given the lack of evaluation of reconstruction algorithms, we propose the methodology we introduce here as a standard.

Lines 10-11: 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).
- We will rephrase to: "The reconstruction based on the J20 database and the S16 algorithm is reliable for timescales above 4kyr during the last 25kyr.", We prefer the term "timescale" to that of "temporal resolution", because the later could refer to the timestep of the reconstruction, which is fixed to 100 years here.

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 deepsea 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...*

- We did not mean that the age models in J20 are superior to those in the S16 datasets. However, the availability of age ensembles for J20 facilitates a more sophisticated treatment of the effect of age uncertainty on the reconstruction, in comparison to the standard deviation used in S16.
- We agree with the reviewer that the age ensembles in J20 do not consider the errors arising from the visual matching, and that a synchronicity is indeed assumed, albeit only regionally. Age uncertainty is already the largest source of signal smoothing for the period where the matching is the most relevant. Hence, considering these errors does not qualitatively change our result, but would increase the timescale below which the reconstruction is not reliable. We will add a sentence in the discussion.

Finally, your point on bioturbation affecting age uncertainty is very interesting, and we will add it to the relevant discussion. With sedproxy, we only consider the smoothing effect of bioturbation but not how it impacts radiocarbon ages. To do this, a more complex pseudo-proxy algorithm would be needed, in particular, one that re-computes the age ensembles.

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.

- The previous short reply should partially clarify this. We compare data from the same simulations (simulated GMST VS reconstructed GMST), so the comparison is not affected by the assumptions made on the data forcing the simulations.

Section 3.2.1

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

- We will add it (Berger and Heath, 1968).

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?

- You are correct, this is exactly what the approach does. The pseudo-proxy temperature is weighted towards the time and temperatures favouring a given species, not only seasonally but also through the bioturbated layer. We will correct the text accordingly.

Section 3.2.2

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.

- We understand this could be difficult for the reader. We tried to make the text as readable as possible, while the table offers a quick overview of each experiment.

Section 4

It is unclear to me here what is being discussed here and shown in Fig 3.

- We clarified this in the short answer, and will add to the caption that the reconstructed GMST is based on the S16 algorithm, and that the simulated GMST is that of the respective climate simulations.

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.
- These comments have been addressed in the short answer. We do not use the reconstruction in S16, only the algorithm, which we applied to the pseudo-proxy data computed from model data ("*reconstructed GMST*"). The *simulated GMST* is simply the area-weighted mean of the surface temperature field produced by the simulation.

Line 316: 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.
- Here, the uncertainty is the one estimated by the Snyder algorithm. It can be split into the 5 components for either PPE reconstructions or real reconstructions (not performed here). However, the validity of the uncertainty estimate can only be determined in a PPE experiment, where the truth is known.

Line 372: What is the northern hemisphere insolation curve? Be more specific.
- Summer solstice insolation at 65°N, we will add it to the text.

Line 375: 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?)
- We mean timescales > 10kyr., i.e., at least half a precession cycle. We will change this in the text.

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

Line 560: 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.
- We agree, the sentence is not very well formulated. We will rather discuss the need to better understand the age uncertainty in order to better constraint it. That will better fit the work by Waelbroeck et al., 2019, and we can include one of the paper you suggested here (Lougheed et al, 2022).

Section 5.3

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.

- This has been clarified in the short answer. We do evaluate the algorithm, and thanks to the PPE framework, we can compare the reconstructed GMST to the simulated GMST from the same simulation, which in that case represents the target.