Overview

This is an interesting and well-presented paper which I enjoyed reading. I would like to thank the authors for their work. They provide a statistically-rigorous approach to combine information between various sediment cores when all these cores provide observations of the same (fairly smooth) underlying function. This is known as *stacking* the records.

I presume that the model builds on earlier work (called HMM-Stack, Ahn *et al.,* 2017) in its HMM aspect for each sediment core. At its heart, the method assumes that each core j records the same underlying function, providing paired observations (y_i^j, d_i^j) where:

$$y_i^j = f\big(\theta_j(d_i^j)\big) + \epsilon_i^j$$

Here $\theta_j(d_i^j)$ represents the age-depth relationship in each core (which can be based upon radiocarbon dating or any another technique). The methods uses MCMC to iterate between updating the age-depth models $\theta_j(\cdot)$ for each core; and the shared function $f(\cdot)$. Within this MCMC, $f(\cdot)$ is modelled as a Gaussian Process (GP), and the age-depth model in a more complex manner (presumably based upon an approach laid out in HMM-Stack). The method does some initial particle filtering but then seems to actually ditch that approach (using it only for initialisation) to use a Metropolis-Hastings As such the particle filtering appears somewhat redundant and could therefore be de-emphasized.

Overall, the paper is nicely written with sufficient technical detail to allow reproduction. The authors also give useful practical examples for $\delta^{18}O$ reconstructions from several marine cores. The method is potentially adaptable to a considerable range of scenarios and will provide a significant contribution to the community (although my expectation/experience is that for records which are not as smooth, or as shared, as $\delta^{18}O$ some bespoke modifications might be required to get the model to fit – which the authors also state).

Statistical Comments (mainly regarding the SI):

1) My main statistical comment is that, as a new reader, I do not sufficiently understand where the specific three state HMM age-depth model comes from. I am presuming this specific age-depth model builds on previous work. In the model, there are considered to be three states for a core. Given a particular state then the sedimentation rate follows a mixture of three log-normals restricted to being within a certain range.

   This particular sedimentation model seems extremely specific, yet its justification is not really provided in either the main paper or S1. It is not clear, to a new reader, where this model comes from: either in terms of three states (with seemingly arbitrary sedimentation rate bounds) or the mixture of three log-normals within its permitted ranges (are these fixed or also estimated). What is the benefit of such a three state model, why were the boundaries chosen, and how are the parameters for each state selected? Is it somewhat arbitrary or is there geoscientific insight as to why there are three distinct states with these values?

   I presume this model, and its explanation, comes from the earlier HMM-Stack work of Ahn et. al (2017). If so that is fine – it does not need to be re-justified here in detail. However I feel there does need to be an intuitive lay-person explanation in the Intro about how it builds on this earlier work and what is specifically new here. Currently the HMM model appears somewhat out of nowhere. Also S1 needs much stronger referencing to that work (to clarify a reader should look there for the justification.

2) I do not quite understand how the particle filtering is used to initialise the MCMC. How do you choose which particular particles to use (after you have run the particle filtering step) as the initialisation of your later MCMC? Do you run the MCMC many times with lots of initial starting points? How have you checked actual MCMC convergence and ensured you have explored the space fully from your initialisation?

3) Outlier model – I may have misunderstood but, formally, it seems you have chosen $g()$ to depend upon μ. If so, I think you probably cannot entirely ignore those observations classed as outliers in the MCMC updating. When you update the GP $\mu|O, Y, A$ I would presume that the outliers will still inform as they come from a distribution that depends upon the parameter you wish to update. Consequently, I'm not sure that formally you can ignore all the values with $O = 1$ and just fit a GP to the others.

   This is unlikely to make much difference in practice **so I am not saying that you need to change it** (but you should perhaps mention this is an approximation). Perhaps you could get around this by keep the same mean for the outliers as the stack but just altering/increasing the variance for the outlier component $g()$. If you do this then one would presumably still include the observations in updating the GP stack but the outliers would have less weight.

4) The section on length/complexity does not really tell me anything practically useful, e.g. the DNEA stack has a run time of 1.8 hours. That's partially interesting, but how many MH iterations actually is that (bearing in mind you have ditched the particle filtering by that point)? You could presumably make it arbitrarily faster/slower entirely dependent on how many iterations you run everything (optimisation, particle filters, MCMC, …). Please tell me how many MCMC iterations you performed.

More General Applied/Presentation Comments:

1) I think it would be worth mentioning how your work links with/alongside broader *errors-in-variables* regression analysis. Fundamentally, that is rather analogous to what you are doing here – if the primary interest is in the stack rather than the age-depth model of each core which it seems to be. In errors-in-variables analysis, one has a series of observations y where $y_i = f(\theta) + \epsilon_i$ but you do not know θ_i precisely (you only observe $T_i = \theta_i + \eta_i$). This is effectively your situation - where your sediment cores provide a specific type/structure of calendar age uncertainty η_i and is some cases the T_i's are not observed at all.

   In a geoscience setting, using Bayesian techniques similar to you, this is basically what we do to make the IntCal curves (e.g. Heaton et al. 2020) but there is also quite a lot of general statistical methodological literature (e.g. Bayesian approach of Cook and Stefanski, 1994) on the topic. Additionally, there is quite a lot of literature on *registration* in functional data analysis which could briefly be mentioned (e.g. book by J Ramsay and B Silverman).

   I also did some work with a similar (but identical) goal – aiming to sharing age information between records using tie-points and a GP – in Heaton *et al.* (2013). This was used to create calendar ages for the Pakistan and Iberian Margin (Bard et al. 2013) , and Cariaco Basin

(Hughen & Heaton 2020) data which then went into IntCal13 and IntCal20. This work was somewhat different in that we only tried to transfer dating information from one record to another and only used the tie-point ages. However it does provide a previous context where tie-points are used in a method that aims for statistical rigour rather than eye-balled tuning (with uncertainties on the contemporaneity of the ties rather like your model). Your work is however more in depth and generalisable than ours (we needed fairly simple age-depth models with multivariate covariances so owe could input then into the main IntCal curve creation)

Suggest that all this only needs a brief line or two in the Intro – just to add more detail/context about how your work fits within the wider statistical research literature.

2) My colleagues (when I tried to suggest a similar approach to them to map all features across records for other proxies) were very cautious. They felt that, for many records, the entirety of the proxy could not be mapped between cores. They rather believed that, for many proxies, it was often only the sharp/main transitions that were shared between records and they did not want to match everything.

   I feel this point, that users must consider if trying to match every feature is something that will work for their proxy/data, should be made very explicitly. You do mention this in the manuscript but it is somewhat hidden and only appears towards the end (in the middle of the section on Strengths/Weaknesses on lines 520-525). I feel this caveat needs to be made considerably more prominent in the Intro/Conclusion when discussing GPs so readers will not misunderstand.

   I am not a sufficient expert here, but it may be that benthic $\delta^{18}O$ is more globally homogeneous than many other proxies (and the method must be used with considerable caution for some other proxies where responses can be antithetic).

3) Your Marine sites are very spread out and will not be expected to have the same regional offset ΔR from one another. You have chosen a mean of $\Delta R = 0$ for all the cores but then quite a large uncertainty ($\sigma = 200$) on ΔR to account for uncertainty. Again this is probably fine, as you have chosen a fasirly large value (and I think everything will be somewhat led by the fitting of the many $\delta^{18}O$ measurements anyway). However, I would suggest that you might advise users to initialise a different ΔR for each core using the Reimer and Reimer (2017) database.

   We do not advise people to choose $\Delta R = 0$ if they have other information available. The belief is that, at least during the Holocene, any regional ΔR will remain roughly constant over time and so will be applicable along the core (as regional upwelling/ocean depth might remain relatively constant). If you choose an independent ΔR from one observation to the next then you do not model dependence in

   Note: This is a fairly minor point that I doubt will affect your results due to the volume of $\delta^{18}O$ observations. If it is a lot of work (or the marine core sites you use do not have ΔR estimates) then I suggest you just add a caveat/explanation for the paper (rather than redo everything).

4) Is there a reason as to why the sedimentation rates of Lin et al. (2014) are applicable elsewhere? This seems like a considerable assumption. Hence while it is potentially a strength of your method to provide automated selections of sedimentation rates it is also a considerable danger if other use it as a black box when it is not appropriate. At the very least, you must ensure that any user inputs their data on the same measurement scale (i.e. m or cm) as the analysis you did for Lin et al. (2014).

Smaller Specific Points:

Main Document:

1) Line 66 – it is not only ^{14}C production rate changes which cause variations in past atmospheric ^{14}C/^{12}C levels but also rearrangements of the carbon cycle (see e.g. Heaton *et al.* 2021). Suggest minor rewording to acknowledge this.

2) Line 370 are your stack estimates smoother because you use a GP which is fundamentally a significant smoother? Or due to other factors such as averaging over calendar ages? Also does the smoothed version lose genuine features - are the features you say you smooth/lose thought to be genuine phenomena?

3) Figure 3 and Figure 4 – in the panel As showing the final stack, can you overlay the posterior mean estimate on top of the observations (rather than underneath where currently it can't be seen)

4) Line 473-474 - *Users should be aware that the age uncertainties returned by BIGMACS for age models generated by multiproxy alignment or stacking do not include the age uncertainty of the alignment target.* I do not understand this comment about an alignment target – based upon your SI you suggest you can use your method on records where there is no a priori alignment target (i.e. when you just have a selection of cores each with their own 14C dates). Have I misunderstood?

Suppl. Information

1. There are repeated uses of sigma to mean many things – unclear what the values that are updated in S4 refer to. Equally what are the h's – need to be made somewhat clearer?

2. More detail is needed on the parameter choices for the age-depth model – can refer to other work if this is suitable.

3. Minor point – the likelihoods are not probabilities (the densities are continuous)

4. S5 – There is some referencing to other sections that has gone wrong: "The stack construction algorithm first iterates steps in subsections S4.2, S4.3 and S4.4 until convergence and then update the new one by the method in S4.1."

   There is no S4.4. Also, do you mean S5.1 at the end rather than S4.1?

General Questions (as I'm interested – not requiring further work):

1) I tried work on a similar topic a few years ago. I found that the lack of homogeneity in the underlying functions we considered (and that it was only some features that were shared) made the method hard to implement in practice. I didn't get it to work very well (hence it remains unpublished).

   Do you think that there is something special about the $\delta^{18}O$ signals you use that mean the features are highly shared between cores? Do you expect it to work as well for more challenging/variable functions/proxies? Do you think there is a danger that you get into highly multi-modal fits in some cases which the MCMC will not fully explore – or is your age-depth prior sufficiently strong to avoid that?

2) How much of a difference do the ^{14}C measurements really make a difference when you have to match 2000 $\delta^{18}O$ observations? Do these swamp the independent calendar age information? Might there be use in having a dependency in the proxy measurements you wish to construct (from one observation to the next)?

3) As a statistician, I think it is a bit of a shame that all of the material on the methods itself has been moved to the SI. I appreciate I am biased and that many readers will be much more interested in the results than technical details.

References

Ahn, S., Khider, D., Lisiecki, L. E., and Lawrence, C. E.: A probabilistic Pliocene–Pleistocene stack of benthic δ18O using a profile hidden Markov model, 2, https://doi.org/10.1093/climsys/dzx002, 2017.

Bard E., Ménot G., Rostek F., Licari L., Böning P., Edwards R.L., Cheng H., Wang Y., Heaton T.J., 2013. Radiocarbon calibration/comparison records based on marine sediments from the Pakistan and Iberian margins. Radiocarbon 55,1999-2019.

Heaton, T., Bard, E., Hughen, K., 2013. Elastic Tie-Pointing—Transferring Chronologies between Records via a Gaussian Process. *Radiocarbon, 55*(4), 1975-1997. Doi:10.2458/azu_js_rc.55.17777

Heaton T.J., Blaauw M., Blackwell P.G., Ramsey C.B., Reimer P.J., Scott E.M., 2020. The IntCal20 approach to radiocarbon calibration curve construction: a new methodology using Bayesian splines and errors-in-variables. Radiocarbon 62,821-63.

Heaton T.J., Bard E., Ramsey C.B., Butzin M., Köhler P., Muscheler R., Reimer P.J., Wacker L., 2021. Radiocarbon: A key tracer for studying Earth's dynamo, climate system, carbon cycle, and Sun. Science 374: eabd7096.

Hughen K.A., Heaton T.J., 2020. Updated Cariaco Basin ^{14}C calibration dataset from 0–60 cal kyr BP. Radiocarbon 62,1001-43.

Reimer, R. W. and Reimer, P. J.: An Online Application for ΔR Calculation, Radiocarbon, 59, 1623–1627, https://doi.org/DOI: 10.1017/RDC.2016.117, 2017.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. Journal of the American Statistical Association, 89(428):1314{1328.