Review of Slattery et al.

In their manuscript Slattery et al. present an analysis of the performance of a modified ramp-fitting method for the determination of the onset of stadial to interstadial transitions and apply this modified method to decadal resolution model data.

Overall, the topic of the manuscript is of interest to the wider paleo-climate community and especially the detailed analysis of interstadial onsets in model output is interesting and a good fit for CP However, there are some larger issues with the manuscript concerning the conceptual background as well as to the presentation of the results that need to be worked out before the manuscript should be considered for publication.

General remarks

The authors extent the ramp-fitting method of Erhardt et al. (2019) (E19) by including two extra parameters for the pre- and post-transition slope of the data. This is necessary to better describe the model-output that the authors want to apply the method to, especially the AMOC strength. The authors do not show, if the inclusion of the slopes into the model is necessary for all their data and also do not provide any evidence that the inclusion of the slopes decreases the uncertainty of the estimates for the relevant parameters. One would also hope, that the addition of two extra parameters to the model would make the model consistently outperform the original E19 version. Judging from the figures in the appendix this does not necessarily be the case. However, due to presentation of the results that is difficult to judge.

Before coming continuing, a note here on terminology: The authors chose to coin a new term for the average value of the posterior samples. This is unfortunate, as there already is a term that is used for this: posterior mean or more specifically the marginal posterior mean. This is the term that I will use and that I strongly urge the authors to use, too.

In an extensive sensitivity study, the authors test if the posterior mean of the transition onset is on average an unbiased estimator of the onset of the transition. Most, if not all, of the content of the paper hinges on the results of this sensitivity study. Despite that, the authors never explicitly state the exact focus of the sensitivity study and the implications that come with this choice. However, that exact focus of the tests is a big potential issue: In the sensitivity tests, the marginal posterior mean does not need to equal to the known parameter value and same is true for any other point estimate from the full marginal posterior. If the posterior is skewed or multi-modal then any point estimate will grossly misrepresent its true location and disregard its shape. Especially if skewness of the marginal posterior is systematically induced due to the noise or due to trends around the transition, it can introduce a systematic bias into the marginal posterior

mean. This can for example be seen in Figure 2 of E19, where the two marginal posterior distributions for the start and the end of the transition show a clear indication of skewness and in the case of the onset maybe even bi-modality. The same argument is of course true for the length of the transition and also extends to all other estimated parameters. Sune Olander Rasmussen (SOR) raised a not dissimilar issue in his review: By using the marginal posterior mean the authors disregard the sometimes large uncertainty in the estimated parameters that is carried in the complete marginal posterior distribution. As pointed out by SOR, the width of the marginal posterior distributions can easily be larger than the reported bias of the posterior mean. From my point of view, this sole, narrow focus on the marginal posterior means is in stark contrast to the strong general statements on applicability of the method throughout the paper, especially because this tight focus is not explicitly communicated to the reader.

In addition to the conceptual shortcomings of the sensitivity study, the presentation of its results in Figure 3 and in the appendices is also not ideal. If I guess correctly (there is nothing written in the manuscript), the authors average over the variables that they do now show to present the data in a reduced number of dimensions. The main issue with this approach is, that it hides any possible inter-dependencies between variables and does not allow for clear separation of the effect of the single variables. I think this is a missed opportunity and might be potentially misleading, depending on the parameter interdependencies.

The authors apply their modified ramp model to climate model output of DO-type transitions that is available in decadal averaged output. For this the authors chose to stay with an AR(1) noise term for the application of the ramp model. I am, like SOR, a bit puzzled by the long autocorrelation times in Fig 3e. Of all the parameters only the ocean circulation would likely have long-term memory, of all the other parameters, I would assume that decadal averages would mostly eliminate the autocorrelation. I would ask the authors to provide a more detailed explanation of this phenomenon and a discussion of the influence of the long autocorrelation times on the timing estimates. Generally the autocorrelation of the residuals in the model is in competition with the transition ramp itself, for lack of a better term. If the model fails to describe the noise of the data accurately then the model will give unreliable results. Long autocorrelation times can smear out the transition for sure leading to larger uncertainties and maybe even "biases" towards earlier or later transitions. This can make the choice of the noise model and the choice of priors for its parameters very important. In the case of the presented study, the authors explicitly excludes the case for non-autocorrelated data and put, in comparison to E19, larger prior probability on long autocorrelation times. Especially the former might be a serious issue in case of the decadal averaged data from the model and merits a thorough discussion.

To overcome the bias that the authors discovered, they use "analogous" data to estimate the bias and later correct for that. Generally the authors do not provide much detail about the exact parameters that they use to generate the "analogous" datasets that they use to make the bias correction for the different parameters. It would be great if the authors could provide more detailed information about the parameters that they chose such that the results are later reproducible. Their statements about the fact that they cannot reliably estimate the onset or the length of the transition given their methods makes me also wonder how they estimated all the parameters that they need for the "analogue" data. Did they use a different method or can they use the ramp-fitting method to estimate all other parameters needed for the surrogate data reliably? Especially in light of the statement by the authors that the relationship of the bias to the ramp parameters is complicated it would seem to be necessary to estimate this parameters

very precisely or carry out the bias calculations for a range of possible parameters. I strongly urge the authors to investigate this in more detail before making definitive claims about a bias. The proposed bias correction also carries additional uncertainty with it, how do the authors propagate that to the estimates later?

In the analysis of the modelled precipitation data, the authors state, that the noise characteristics of the data change over the transition. The authors somewhat disregard this observation with the notion that they failed to include this in the model. However, the results of the analysis are dependent on the noise model adequately describing the data. This is especially worrisome as the precipitation time series exhibits the biggest lead over the temperature in the model data. The fact, that the authors state that inclusion of changing noise characteristic over transition negatively impacts the reliability of the method might hint at a deeper problem. This makes me question the results that they present and the applicability of the estimates for the generation of the "analogous" data and thus the bias correction that the authors apply

Specific remarks

58–63 Order of publications is wrong.

Table 1 This suggests a connection between Ca and the NAO, please justify or remove

125ff Focus of the work: If that is the case, please rephrase the introduction and the abstract accordingly. In this case you can also remove all the model data. Or spend some time discussing the data and its interpretation. There is no need for statistics without a clear question to answer.

138ff This is a wrong statement. If that were true then OLS-regression would be probabilistic, too.

143ff Please show proof of this.

149 Remove reference to Erhardt et al., as they do not use any frequentist methods.

169ff Nomenclature: There are some already established terms that the authors should use as much as possible instead of inventing new ones: the average of the posterior distribution is the posterior average or posterior mean.

Fig 3 The presentation of the results is unclear. The authors do not state what they do with the other parameters that were varied. I assume that they just averaged over all the other parameters that are not shown in the figure. However this approach will hide any co-dependencies of the bias between different parameters.

213ff : Please specify how you set these two sets of data up. I would imaging that the exact difference between the two resolutions depends a lot on the parameters as well as the timing of the onset, depending if it happens at a boundary of an averaging period. The results here are also contradicted a little by Fig C1, which shows clear differences between annual and decadal resolution data

Table 2 The bias of what? Specify in the table header.

Sect 3.3 Please specify the exact parameters used to generate the analogous transitions. You say that you cannot reliably estimate the transition length, but you do not show that you can actually estimated the other parameters without issues. I think this is minor but a few sentences in this regard would be good.

230ff It is unclear how the posterior mean and event-means are set up here. Are you generating 19 events or are you subsampling 19 events out of your 1000 realisations to calculate the event-averaged timing?

239ff Please elaborate on your the heteroscedacity issue: If the method is unreliable with the inclusion of different noise levels over the transition, than how is it reliable enough to estimate analogous parameters. It is also unclear if you generated the analogues with or without heteroscedacity which makes the last sentence of the paragraph quite confusing. Please rework. It also needs to be clearly shown and discussed whether the bias correction for the modelled precipitation is valid and if the model can be applied at all to the heteroscedastic data. The issue with these models is that their results are conditional on the model being a good description of the data, if that is not the case the results are to be taken only with a grain of salt and proper justification.

Tab 3 : Please do not format the table in such a way that Ca seems to be associated with the NAO. This is potentially misleading especially as you shy away from discussing the proxy interpretation at all. Also, Even though temperature and precipitation over the North Atlantic that you use is probably not unrelated to the values at NGRIP, they are not the same. Please redo the analysis with the relevant variables or rework the manuscript in such a way that it is clear that you are not looking at model output for Greenland at all. This included the Table where these are stated next to each other as if they are the same.

255ff Please discuss why the lags are so different between the different variables. Is that because of the shape of the transitions or the noise levels or both?

Sect. 3.5 Again there is no clear description on how the analogous transitions were generated. Without that information the results cannot be reproduced or verified by anyone.

281ff I think the leading paragraph is a bit misleading without the actual information under which circumstances the bias arises.

285ff This paragraph probably should state somewhere that you look at this from a frequentist point of view. Erhardt et al. never made any claims about statistical significance, they provide credible intervals which are the parameter ranges that are consistent with the data under the assumptions of the model. This is not the same as the assumptions that go into the significance test, namely that under repeated observations and calculation of the significance interval of 95% the true value of the parameter will fall into the significance interval 95% of the time. These two views are fundamentally different and are not necessarily compatible with each other.

318ff The paper mostly focuses on the adjusted method that you are proposing, and not the original method of E19. Please rephrase as all results with regard to the original method are only in the appendix of the paper.

327f The conclusions of Capron et al. (2021) are based on fundamentally different reasons than the once here. Your's are purely related to the method and the interpretation of the results whereas those of Capron et al. (2021) are based on considerations with respect to climate variability and the tight coupling between different parts of the climate system. Please remove or rephrase.

329ff The conclusion is inconsistent. The method is either good or not for the investigation of DO events. Please decide and rewrite. Please also state why you think that the method would be suited if the lags are larger than 20yrs. This is not entirely clear given the discussion beforehand.

Supplement

Choice of prior distributions

- Either Equation 10 or your code is not correct, please check.
- The prior for tau in your analysis puts a lot more emphasis on much longer and much shorter autocorrelation times than the one in the original method. This is a bit problematic as we do not expect to see decades long autocorrelation in climate data as far as I know. Please elaborate.

Sampler setup Please clearly describe the setup of the ensemble sampler. I have noted that you chose to keep the setup of E19 despite increasing the number of parameters. This should also go hand in hand in hand with a change in the sampling strategy i.e. burnin and thinning and maybe even the number of ensemble walkers. These need to be carefully chosen to not get nonsensical results in the end.

Figures C1-C3 Judging from a closer comparison of the results from your method to the results of the original E19 method in Figures C1 and C2 it would seem that the addition of the slopes generally decreases the performance of the method. This seems to be the case for the sensitivity of the bias towards noise, the transition length and very, very worryingly towards the inclusion of an interstadial slope. Please elaborate and discuss this especially in light of the analysis that you base upon the result from the fit including the slopes. Also it would seem, that at zero slopes prior and after the transition the original method overall outperformed yours, which would mean that part of the bias is because of the inclusion of the slopes into the model not despite.