Review of the revised manuscript by Riechers&Boers

To begin with, let me express my appreciation to the authors for taking the time and making an effort to completely revise their manuscript following the first round of reviews. The revised and streamlined methods section is now much easier to follow and more clearly conveys the authors approaches. With the expanded results section, the manuscript now also aims to not merely provide a statistical toolset but also paleo-climate relevant results. With this the authors have addressed the major concerns voiced by the reviewers. Judging from their thorough replies the authors have also addressed most of the concerns that the reviewers had in a satisfactory way.

Nevertheless, I still have major concerns regarding the methods the authors present which I unfortunately missed in the first round of reviews.

Main remarks

In their Eq. 29 and the Appendix C the authors derive the posterior distribution of the lag, averaged over all DO-events (even though the they choose to not call it that). This distribution essentially encodes the entire knowledge about the mean lag over the DO-events, given the model and the data, the quantity the statistical tests the authors make are concerned with. The fact that these tests than yield "non-significant" results for the proxy-pairs that put low posterior probability on a positive average lag is difficult to reconcile and merits a more detailed discussion by the authors.

One reason might be that the notion of significance levels and decision thresholds in the context of distributions of p-values is not as straight forward as the authors convey. As the authors point out because of the non scalar nature of the p-value distributions that they employ, they need a way to map them to a binary decision about the null hypothesis (L 284f). The authors list three different decision criteria, all employing the same significance level $\alpha = 0.05$:

1. the p-value value for the expected lag, averaged over the different events: $p(E(\mathbf{\Delta t}))$ not exceeding α;
2. the expected p-value of, averaging over the MCMC samples of $\mathbf{\Delta t}$: $E(p)$ not exceeding α; and finally
3. the fraction MCMC samples of $\mathbf{\Delta t}$ for which the p-values do not exceeded α: $P(p < \alpha) < 0.90$.

Of these three criteria the authors postulate that only the latter two provide the propagation of all uncertainties to the decision about the significance of the lag. Hence the authors base their main argument of the results of the tests (2) and (3).

In the textbook meaning of the significance level α, α denotes the probability that the significance test falsely rejects the null-hypothesis given that it is in fact true. Which is the meaning that almost all readers of the study will be familiar with. However, the significance criteria favoured by the authors to make their argument do not possess this property.

To test the implications of the decision levels proposed by the authors, one can run a simulation of the null-hypotheses repeatedly and assess the rate of false rejections for the different criteria. To do so, consider the following simulation, mimicking the null-hypothesis in the setting of the t-test: The lag for each of the n events is assumed to be normally distributed as

$$\Delta t_n \sim N(\Delta T_n, 1^2) \tag{1}$$

representing the draws from the MCMC sampler for the vector of delays $\mathbf{\Delta t} = (\Delta t_1, \Delta t_2, ...\Delta t_n)$. For each of the events, ΔT_n is drawn i.i.d from

$$\Delta T_n \sim N(0, 1^2) \tag{2}$$

representing the sample of n DO-events drawn from a normal distribution centred around zero.

Doing so repeatedly for 16 ΔT_n and than 6000 Δt_i each, the effect of the different criteria by the authors can be tested given the null-hypothesis is true. Admittedly, this simulation uses an extremely idealised case for the null-hypothesis, but one could argue that drawing from standard Normal distribution, especially for the t-test, resembles an ideal case scenario. The code to generate the simulation results will of course be provided to the authors if desired. The same type of simulation could be done for the other two tests the authors use.

Figure 1 shows histograms of the outcomes of all three decision criteria after simulating 50000 tests under the null-hypothesis using the setup outlined above. In each of the panels, the red dashed line indicate the decision criterium proposed by the authors.

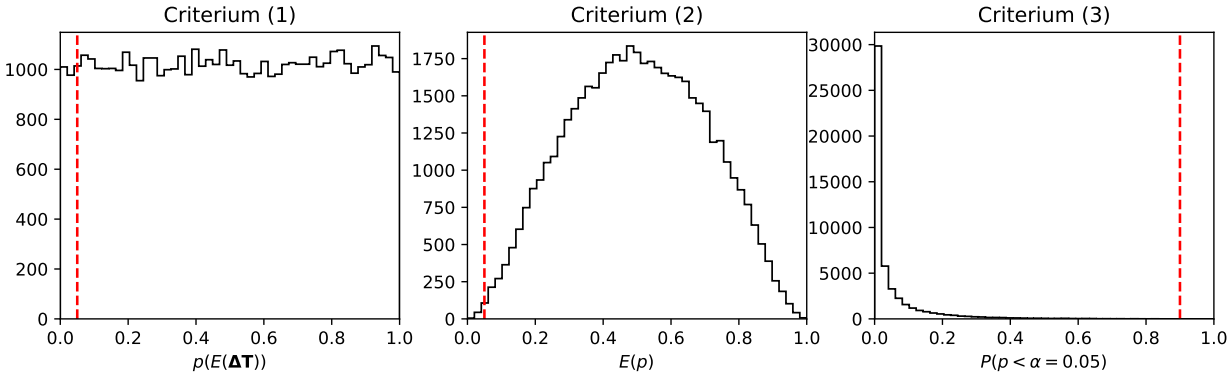

Figure 1: Results of simulation 50000 t-tests using the sampling scheme for the null-hypothesis outlined in the text. The red dashed lines indicate the decision criteria for significance at $\alpha = 0.05$, used by the authors.

As mentioned earlier, a significance level of α indicates that, given the null-hypothesis is true, in the long run exactly this fraction of the tests will have a p-value that is below this level. For a regular test statistic this is due to the fact, that under the null-hypothesis the p-values are distributed uniformly between 0 and 1. For the significance criteria the authors present, this is only the case for $p(E(\mathbf{\Delta T}))$ (1), as shown in the leftmost panel of Figure 1. Both other tests fail to exhibit this behaviour. This means that the significance criteria used by authors are not consistent with each other. In fact, both of the other tests (2, 3) use a much higher bar for significance than α would lead the reader to believe as indicated by the two other panels of the Figure. For (2), $E(p)$,

the probability of $E(p) < \alpha = 0.05$, given the null-hypothesis, is only 0.0017, almost a factor of 30 lower than the implied value of 0.05. The same holds true for the third proposed criterion, where out of the 50000 simulations, only five exceed the $P(p < \alpha) > 0.9$ threshold. Accordingly, by using these two criteria, the authors set the bar for a "systematic lag" a lot higher than the significance level of $\alpha = 0.05$ would make it appear. Consequently, both of these test do yield a non-rejection of the null hypothesis for all studied cases in turn leading the authors to their main conclusion: that there is no evidence for a lead-lag relation in any of the studied proxy pairs.

To conclude, I am not convinced that the testing criteria used by the authors to draw their main conclusions are fit for the purpose as stated in the manuscript. By implicitly using much more stringent significance criteria for their two preferred tests, without clearly communicating this to the reader, the authors unfortunately seem to have (accidentally) moved the goal post for what they are trying to investigate. In my eyes this necessitates either a complete re-evaluation of the results of the statistical tests and the conclusions drawn from these results by the authors or a removal of the tests in question from the manuscript.

As an interesting side note: the results from the first decision criterium, deemed "simplistic" by the authors (L 265), are in general agreement with both the "combined evidence" of Erhardt et al. (2019) as well as with the "uncertain sample mean" as derived by the authors. Maybe unsurprisingly, it indicates exactly the opposite of the two other tests: significant evidence at $\alpha = 0.05$ for a less than zero lag across all studied proxy pairs. Obviously, this decision criterion is most closely related to these probabilistic quantities, as it tests wether the average delay between a pair of proxies, taken over all observed events is significantly different from zero. Which is, as far as I understand, what the authors try to investigate.

Additionally, prompted by the authors statement about the novelty of their "uncertainty propagation to p-values" in their response: A cursory literature search (keywords *fuzzy p-values*, *bayesian p-values*) brought up a range of papers that seem to be dealing with p-values in settings similar to the setting the authors deal with here. It would be good if the authors set their approach into the context of aforementioned literature or highlight its conceptual differences should the they decide to further employ it in the next iteration of the study.

Other remarks

L 129ff It is not true that Erhardt et al. (2019) only considered time series free from data-gaps. In fact, looking at Figure 3 in the presented manuscript, both the Ca^{2+} as well as the Na^{+} data series do in fact exhibit at least one section of missing data. Please reformulate.

L 198ff The choice of terms for the different type of distributions is a little bit misleading. Posterior distributions as generate by Bayesian inference are probability distributions. Yes, posterior distributions carry the uncertainty about an inferred parameter conditional on the data and the model, but they are probability distributions nontheless. By using the neologism "uncertainty distributions" the authors implicate that they are more uncertain than their "probability distributions" generated by random experiments. This sets the tone for the discussion that follows in a very odd way. The authors should use the correct term "posterior probabilities". Changing this would also avoid such contortions as the "high uncertainty probabilities" (L 305)

L 208ff Prescribing a "fixed pattern of causes and effects" is an overly strong interpretation. It is quite easy to imagine a range of mechanisms that have an indistinguishable imprint in the

proxy record but trigger a transition from stadial to interstadial conditions. For example the ocean processes alone, that the authors list in the introduction are probably very difficult to distinguish using the proxies presented here as they partly invoke very similar feedback mechanisms. I would suggest to reformulate "similar pattern of cause and effects" to convey the possible ambiguity as a discussion of the imprint in the proxy records by the different causes clearly goes beyond the focus of this manuscript. The same holds true for later implications of the one trigger of DO-Events that the authors make throughout the manuscript.

L 201 The imperative "should" in regards to the setup of the frequentist analysis that follows should be replaced with "could" or "can".

L 220 Either "be" or "bear".

L 224 The footnote should either be added to and discussed in the paper or removed. As it is right now it is just a clever remark that does not contribute to the overall manuscript.

L 263 I think it would be better to say that the sample no longer *only* carries the randomness of the population. As would be the case for a regular statistical test with certain values.

L 284 In null-hypothesis significance testing the null-hypothesis can only be rejected.