Clim. Past Discuss.,
https://doi.org/10.5194/cp-2020-136-RC1, 2020

Climate of the Past Discussions

Interactive comment on "A statistical approach to the phasing of atmospheric reorganization and sea ice retreat at the onset of Dansgaard-Oeschger events under rigorous treatment of uncertainties" *by* Keno Riechers and Niklas Boers

Anonymous Referee #1

Received and published: 14 December 2020

In their manuscript Riechers and Boers present a method for the statistical analysis of the results presented in Erhardt et al. (2019) with the goal to determine the average phasing of the onset of Greenland Interstadials (GIS). To do so, they present three different approaches with the same goal: To account for between-event variability. This additional layer of variability, as pointed out by the Authors (as well as more implicitly by Erhardt et al., 2019), was assumed to be non-existent to derive the averaged estimates in the original study. The exemplary application of their method highlights the impor-

tance of accounting for this additional layer of variability and underlines the difficulties of investigating changes in a very tightly coupled system under uncertainty.

Overall, the paper is very well structured and written and the presented methods provide an interesting toolset for a range of questions. The derivations of the methods are presented very thoroughly, and all necessary technical information is present. For the broader paleo-climate community (which is the audience of CP) the method descriptions could be supplemented with more explanations to aid intuition and prolonged use of the innovative approaches.

Unfortunately, and this is my biggest concern, the authors focus the manuscript entirely on their methods and only provide one brief albeit very intriguing example at the end. This is exceptionally disappointing as the analysis could easily be extended to the other records presented in Erhardt et al (2019), especially given the simple calculations needed for the presented methods. The answer of the authors to this concern will likely be that this analysis will follow in a later paper. However, this raises the question whether the manuscript in its current form qualifies for a journal that is not focused on statistical methods but the climate of the past. In the present form, the paper is basically a method collection and is better suited for a different i.e., method focused journal. I thus strongly urge the authors to also add the results for the other records presented in the original study. This would not only demonstrate the viability of their method but would also further our understanding of the climate dynamics during the onsets of an GIS. Additionally, this would allow the authors to put their results into perspective when compared to the vast amount of research (other than Erhardt et al .2019) that exists on the records and mechanisms DO events which presently is sorely lacking from both discussion and conclusions. With these additions, the resulting paper would be made much more valuable to the broader paleo-climate community.

Throughout the text there are a number of inaccuracies such as wrongly stated ages, confusion of fluxes and concentrations, the nature of ice core records and MCMC. Even though each on their own might seem minor, I will warn the authors that they could be

interpreted as negligence. I thus strongly encourage the authors to seek out the input of experts in ice-core records (of which there are plenty in the TiPES project) to give the manuscript a thorough once-over to avoid potential pitfalls.

Specific remarks

L18: additional the

Figure 2: The vertical lines are colored blue. This is probably an accident. Please also add the full list of references for the datasets shown in the Figure as well as the age-scale to either the caption of the text.

L64ff: What is high? Both for the statements about the record resolution as well as about the variability choice of a relative term is only useful if it is clear what the resolution or frequency is high in relation to. It would be much better to state at least orders of magnitude for these instead.

Table 1: The caption is a bit misleading: In the data that you use in your study only the DO events given in bold are contained. Furthermore, the statement "the stochastic, MCMC-based method successfully detected empirical density distributions for transition onset" is inaccurate on multiple levels: To begin with the method is probabilistic, not stochastic, secondly, the method does not detect empirical density distributions but provides those for the transitions. Following the description of the model in Erhardt et al., 2019, it seems likely, that the investigated transitions where chosen because they could be described well enough with the ramp model. This seems especially likely as the very short sub-events are sometimes poorly defined in the ice core record and/or often exhibit too short stable levels before or after. This is also stated in the text in multiple occasions. The ages stated in the Table are numerically identical with the ones provided in Rasmussen et al. (2014), that means that the age reference is than in fact not 1950 but the year 2000. Please check this throughout the manuscript to make sure

that the correct ages are used at all times. It is also advisable to avoid the use "BP" to avoid confusion with Radiocarbon ages.

L78ff: Please elaborate if you extended or changed the original approach by Erhardt et al. or used it as is. Judging from the code/data availability statement the latter seems to be true. Should that be the case this needs a clear statement to distinguish prior from original work.

L85: Consider not using the variable name here at is it only fully introduced and used much later in the manuscript.

L95: Please consider to not cite the pangea reference separately as it is technically only a supplement to the 2019 study by Erhardt et al. Having both mentioned separately seems contrary to the notion of a supplement.

L98: Ice core record is technically not compressed in greater depth but rather extended in the horizontal due to glacial flow which in turn leads to a thinning and has nothing to do with compression due to hydrostatic pressure. Please use the correct term "thinning". Deposition rates and concentrations are two fundamentally different things, from what I gather you are using concentration records only. Please correct "deposition rates" to concentrations.

L107f: Because the algorithm is based in MCMC it by design (and necessity, as the problem has no analytical solution) returns samples from the posterior distribution, not a probability density function. The probability density functions are later only approximated using kernel density estimates.

L114ff: I appreciate the authors desire to advertise their approaches to a wider audience. However, the provided example is completely irrelevant in the context of the readership of this journal. Furthermore, it is somewhat contradictory because if it were true, then the statistical approaches outlined in the manuscript are hopefully in fact not new but long solved in the medical context. Please find a better suited example and

elaborate why the problem of inferring a population mean from an uncertain measurement is not yet solved? And if it is solved, under which assumptions is it solved and how do your assumptions differ?

Figure 2: How is it possible, that the pdf for Δt is unimodal if one of the pdfs for the transition onsets is bimodal? Please elaborate.

L123: See comment above about the product of MCMC.

L129: See comment above about the statement of failure of the MCMC algorithm. Please either remove the statement or elaborate.

L147f: The investigation has basically been already performed already in the original publication (see Figures A1 and A2 in Erhardt et al. 2019). Furthermore, the test data the Authors use here violates an important and explicit assumption of the algorithm: autocorrelation of the noise (i.e. an autocorrelation time larger than zero) and are thus rendering the tests invalid. I appreciate the intention of the authors here, but the oversight of the white-nose vs red-noise assumption is disappointing at best. I suggest the authors remove this section entirely.

L184: Please elaborate on the statement why hierarchical distributional models cannot be invoked here and better define the term.

L196: As a side note: Summing does not require to keep all summands in memory.

L195ff: This paragraph makes an interesting point as the samples of Δt for each of the transitions are interchangeable, they could technically be reshuffled to simulate more samples. Could you elaborate on the uncertainty that this sub-sampling adds to your methods and how much the results are dependent on the individual realization? If the results are not stable it might hint at the fact that the 6000 samples from what is a 16-dimensional distribution might not be enough to fully capture all of the uncertainty.

L205-213: In this short section, the authors provide the arguably most elegant way of drawing inference on the population from the underlying data. Even though this sec-

tion is a little bit hidden and Eq. (9) is not straight forward to understand, the resulting convolution of the individual posteriors for Δt_i provides a posterior for the average lag of the DO events. My judgment of this section being the most elegant stems from the fact that it is not dependent on any additional assumption such as the presence of an infinite number of DO events or normality of any of the distributions – It rather only answers the question which average lags are consistent with the observed 16 DO events, which is arguably the question the authors set out to answer. Comparing this to the estimates that Erhardt et al. call the "combined evidence" the difference stems from the fact that Erhardt et al. assume that all DO events exhibit an archetypical lag, i.e. one that does not vary between events or verbatim: "[...] this implicitly assumes that the timing differences for all interstadial onsets in the parameters investigated here are the result of the same underlying process or, in other words, are similar between the interstadial onsets" The method presented here relaxes this assumption by realizing that the averaging can be expressed as summation and thus as a convolution of the posterior densities. In comparison to the convolution described here, it is very important to note, that both the t-distribution approach as well as the bootstrap approach described later aim at something subtle, yet fundamentally different: The convolution provides the probability of a mean lag given the observations. The other two however assess the distribution of this mean under their respective assumptions! I will take the liberty to encourage the authors to treat this section entirely separately from the other approaches t and to extent it a little bit to emphasize its difference to the other methods. As a side note/word of caution: The accuracy of the numerical convolution of the kernel density estimates is very much dependent on the chosen discretization and especially range of values that it is performed for. This can easily be tested when comparing distributions where the convolution is known to their numerical convolution (such as the Normal distribution). I suggest the authors do some experiments and present these in the appendix.

L215ff: The "refinement" that the authors present by using the definition of the t-distribution and a change in variables to estimate the mean of the underlying distri-

bution hinges on the assumption that this is a normal distribution. Even though this assumption seems inconspicuous at first sight, the authors should provide evidence that this assumption is both justified as well as not violated by the samples from the ramp-fit. Depending on the justification of the assumption or consistency with the data the results that are based on the assumptions will not be valid or should at least be interpreted with care. I suggest the authors spent some time in this section (and all other sections) to clearly (and in words) state the underlying assumptions, their implications and justifications.

L243ff: The authors try to justify their choice in the t-distribution based estimation by a presenting a bootstrapping version of the same thing. However, I do not think that this can be used to do so. Looking at the results the agreement between $\rho(\mu)$ and $\rho_{bs}(\mu)$ made me wonder why that might be the case: In fact, no matter what randomly generated data the approaches are use on, the results always very closely agree with each other. This is also seemingly independent from the data being normally distributed or not. This seems slightly odd to me however the reason might be, that both methods fundamentally do the same thing: they aim to estimate the mean and the standard error of the mean from the sample and provide a distribution about mean given this standard error. To assure the reader of the validity of their approach, the authors should spend some time elaborating on this and clarify the rationale of why the bootstrap of the mean should be different than the t-distribution and what we can actually learn from having both if they yield seemingly identical results. Furthermore, the authors should add how many bootstrap samples where generated as this is an important information for the reproducibility of their study.

L261ff: After going through a lot of trouble to derive ways to estimate the distribution of the mean from uncertain observations the authors opt to throw all of this overboard and to start from scratch to come up with a way to test whether this mean is different from zero. I am a little bit puzzled why the authors present, what basically amounts to calculating a p-value for each of the MCMC samples of 16 Δt_i rather than investigating

the distribution of the mean that they just derived. I am sure that the results would likely be not much different. It also remains unclear to me, whether this “propagation of uncertainty to the p-value” is actually a valid approach: Essentially each of the 6000 p-values that is calculated constitutes the p-value of the test of that one sample is from a distribution different to zero. The meaning of the distribution of these p-values over many repeated samples is not straight forward. This starts with the observation that for a non-significant distance the resulting p-values will be uniformly distributed, so the distribution of p-values that the authors present needs to be interpreted within that context. Though the approach the authors present might seem convenient and maybe even clever it leaves me with more questions than answers. And with this I am not arguing against the conclusions the authors arrive with using this method, as anything but a non-significant lead would be surprising given the assumptions of the calculations. What is also missing from the otherwise extensive presentation is the alternative view, that the 6000 MCMC samples each present 16 observations of the mean and could be tested accordingly as 6000*16 observations. I am sure that there is a good reason to not do this, but this alternative should at least be mentioned and discussed in the context of the other methods. I suggest the authors either rework this section entirely to better justify and clarify their approach and to include an investigation of the derived distributions of the mean or complete refrain from presenting hypothesis tests in this context. Should the authors decide to keep this section, they need to make sure to state the Null Hypothesis (and the alternative hypothesis, depending how closely they follow Fisher) correctly and explicitly.

L496ff: The authors observe that the distribution that results from the convolution is narrower than the one obtained by the other methods. This is interesting albeit not surprising, given the conceptual difference of the convolution to the other methods. The brief explanation that the authors give here is quite difficult to follow, could the authors elaborate?

L406f: I think this point deserves a moment of attention: Despite the added layer of

uncertainty for the posterior distribution of $U_{\Delta t}$ and the additional assumptions going into $\mu_{\Delta t}$ both still put around 4/5 of the probability on lead of Ca over Na. Yes, this is not 90/100, but in IPCC parlance it is still likely that the transitions are led by a transition in Ca.

L443ff: The calculation of the number of events being consistent with the lead of Ca over Na again is a very good addition to the discussion and a great extension of the order statistics shown in Fig 5 of Erhardt et al. (2019). The authors interpretation of the analysis is however somewhat strongly formulated given the large uncertainties of the estimates: The postulate that if an atmospheric circulation change (effecting only Ca) would trigger the sea ice retreat of the DO events (in turn effecting Na) than all of the events should show a lead of Ca over Na at their onset. This is not wrong but would only ever occur in a scenario where we would observe these atmospheric and sea ice changes directly and without error, not through a set of proxy records and could reasonably exclude any influence of internal climate variability. All in all, that seems to comprise quite a high bar. I suggest the authors to tone down the interpretation of this otherwise very enlightening analysis.

L468f: How do the authors arrive at the conclusion that the observations cannot be used to investigate the transitions with Na leading? To put it sarcastically: If that is not possible, then why is the reverse, investigating the transitions with a Ca lead?

471ff: In the conclusions the authors very carefully and thoroughly present and interpret their results. Overall, I tend to agree with their conclusions based on the methods and the result that they presented.

480ff: The statement on the ability of the presented results to serve as evidence is somewhat unjustified. I do agree that on the base of the presented data the Null Hypothesis of a zero or larger lead cannot be rejected but in reverse that does not mean that the same evidence cannot be used at a later stage (combined with prior knowledge and more evidence).

486ff: The possible existence of a process other than the processes that directly influence Ca or Na is an important note here and is likely the best explanation for what is visible in the data. The authors could spend a little more time on this point.

Interactive comment on Clim. Past Discuss., https://doi.org/10.5194/cp-2020-136, 2020.