Clim. Past Discuss.,
https://doi.org/10.5194/cp-2020-136-AC1, 2021

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

Keno Riechers and Niklas Boers

riechers@pik-potsdam.de

Received and published: 1 February 2021

We thank the referees for their very thorough and careful review. Following the constructive criticism and valuable feedback, we would like to propose several changes to the manuscript. We are convinced that these changes will substantially improve the quality and clarity of our manuscript and that they will address the referees' objections, questions and suggestions. Since both reviews share several general and important

aspects, we would first like to reply to these in a combined way here, and propose a substantial restructuring of the manuscript before we address the specific comments of this referee further below.

1. Most importantly, both referees urge us to expand the analysis to the full possible set of pairs of proxy variables to make the manuscript more appealing for CP-readers. After careful reconsideration, we agree with this criticism and would integrate the analysis of all pairs of proxies that are reported to show a clear lag-behaviour by Erhardt et al. in a revised version of the manuscript. We would not include those proxies, which Erhardt et al. find to transition simultaneously.
2. We acknowledge the comment made by referee 1, that the manuscript appears as a method collection and agree that there is a lack of guidance and motivation throughout the method section in the current version of the manuscript. In particular, both referee reports indicate that the role of the hypothesis tests hasn't been made sufficiently clear. In response to this, we would restructure the manuscript and focus on the key question the manuscript aims to answer: Can we rule out that the lag-tendency observed from a sample of 16 DO events arises by chance, with an underlying population mean of the lag equal to zero? This question can be directly answered in terms of the extended significance tests we carry out, and we therefore plan to put considerably more focus on this in a revised manuscript.

In the revised manuscript, we would then take the following, natural line of inference that should be immediately obvious to the reader:

1. Establishment of the statistical framework in terms of random experiments: The DO transition onset lags constitute an n=16 sample (for other proxy-pairs this number can deviate) drawn from an underlying population. Here, we would integrate a detailed comparison with the assumptions made by Erhardt et al. to derive what they call 'combined evidence', as requested by both referees.

2. Introduction of uncertainty propagation in the statistical inference in the case of uncertain samples.
3. Testing whether or not the observed sample contradicts the null-hypothesis of a population mean equal to (or greater than) zero, given the uncertainty of the sample. A population mean significantly different from zero can be interpreted as a systematic lag. Here, ‘systematic’ does not directly imply a causal relation but is a necessary condition for the inference of causality.
4. Comparison of the uncertain sample mean with the combined evidence reported by Erhardt et al.

We would introduce this streamlined methodological framework in the ‘Methods’ and show the results of its application to all proxy-pairs under study in the ‘Results’ section.

We would like to propose to remove the following from our manuscript in order to increase its stringency and readability:

- Remove the derivation of distributions for the population mean, because it is not required to answer the stated question.
- To simplify the presentation, if you agree, we would refrain from carrying out all the derivations for empirical densities and use continuous probability density functions instead. The equivalence between empirical densities (as provided by the MCMC) and continuous probability density functions in all our derivations would be shown in the appendix. While the formulation in terms of empirical densities has not explicitly been criticized by the referees, we noticed that it does not contribute to the understanding of the physical and statistical reasoning of the manuscript.
- Remove the discussion on the probability for n events to be lead by calcium. Originally, we used this derivation to reject a causal relation, but the objection of

referee 2 made us reconsider our method and we found argumentative inconsistencies.

These changes will result in a streamlined version of the manuscript and make it significantly easier for the reader to follow the main line of thought without getting distracted by a lot of technical details. We think that tightening of the methods together with the incorporation of the additional proxy pairs will yield a convenient balance for publication in CP.

As both referees proposed to treat all 16 * 6000 (6000 MCMC-samples for each of the 16 DO events) values as equal observations of the same quantity and gather all values to one single empirical probability distribution, we would add a section to explains why this cannot be done (for an explanation see General comment 2 of review 2 and similarly L.243 of review 1).

Point by point answer to referee 1:

’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 importance 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.'

We thank the referee for this positive feedback. A clear focus on the main line of reasoning will facilitate the readers' understanding in a revised manuscript.

'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.'

We thank the referee for this comment, which motivated us to substantially reconsider the way in which we present our results. We addressed this in our introductory statement.

'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.'

We will follow this advice and reach out to our collaborators in this regard. In the revised manuscript we will correct these inaccuracies.

Specific Remarks

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.'*

We retrieved the data directly from the supplement to Erhardt et al. (2019) who were the first to publish the shown data for Ca2+ and Na+ concentrations. We will also add the original reference for the different datasets in a revised manuscript. . We would change the color of vertical lines (and tilted connecting lines) to light gray. We would add the time scale to those Figure captions where it is missing.

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.'*

We would add total numbers to clarify what is meant by 'high'.

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.'

The reviewer is right, the caption is indeed ambiguous. We would clarify in a revision that only data from events printed in bold is further investigated in the manuscript. Further, we would change the sentence 'Bold print indicates those events for which the stochastic, MCMC-based method successfully detected empirical density distributions for transition onset.' to 'Bold print indicates those events for which application of the probabilistic MCMC-based method yields a convenient sample from the posterior probability distribution of the ramp-fit parameters. For other events, the rejection rate in the MCMC sampling procedure exceeded a critical threshold of 70

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.'*

We used the ramp-fit algorithm as provided by Erhardt et al. 2019 and we will add a sentence to unambiguously clarify this.

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

A revised version of the manuscript would not comprise the computation of the probability for n_{Ca2+} events to be lead by calcium.

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.'

We would change all references (Erhardt et al., 2018) which refer to the data stored on the Pangea website to (Erhardt et al., 2019) which refers to the article.

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.'*

Thank you, 'compression' will be replaced by 'thinning' and 'deposition rates' will be corrected to 'concentrations'

L.106: 'stochastic' would be replaced by 'probabilistic' and accordingly everywhere in the manuscript, when regarding the ramp-fit algorithm.

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.'*

We would change the sentence 'First, we introduce the stochastic transition onset detection algorithm that by design returns an uncertain transition onset t_0 in form of a posterior probability density distribution.' to 'First, we introduce the probabilistic transition onset detection algorithm that by design returns an uncertain transition onset t_0 in form of an empirical posterior probability density distribution.'

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?'*

It is true that the given example is of no relevance for the readership of CP and hence we would delete the paragraph. We were surprised ourselves that we could not find any literature on this specific issue of hypothesis testing with a sample comprised of uncertain individual measurement. We suspect that in most relevant cases, the uncertainty associated with the individuals of the sample is small compared to the uncertainty that arises from the spread of the individuals within the sample.

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.'*

Computing the distribution of Δt from the two individual distributions for the transition onsets of calcium and sodium corresponds to a convolution of the latter two. Convoluting the bimodal distribution for the calcium transition onset with a kernel as broad as the distribution for the sodium transition onset merges the two peaks of the bimodal distribution into a unimodal distribution.

L123: *'See comment above about the product of MCMC.'*

Again, we would change ‘posterior probability distributions to ‘empirical posterior probability distributions’.

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

We would introduce the notion of the rejection rate from the MCMC-sampling procedure here.

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.'

In a revised manuscript we would add an investigation of how the auto-correlation of the noise influences the performance of the algorithm, thus making our performance test two-dimensional as proposed by referee 2 (L.150). The white noise used up to now corresponds to auto-correlated noise in the limit of the auto-correlation tending to zero and hence still constitutes a valid test case for the algorithm. What is shown in Erhardt et al. (2019) is not a systematic investigation of the influence of the signal to noise ratio on the quality of the returned empirical posterior probability density. First, the signal-to-noise ratio itself is not controlled but only estimated from the data in a way not further defined in the publication. Second, there is no direct way to control how the inferred empirical posterior probability density relates to the true value. In our investigation, both are guaranteed.

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

We in fact think that the problem could, in principle, be tackled with hierarchical distributional models, and plan to investigate this in future work. We'll change this accordingly.

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

The sum is required for the empirical density distribution - the sum is technically not executed. In order to propagate the uncertainty inherent to the distribution in Equation (6) would indeed make it necessary to store all 6000^16 values. We would add a clarifying comment in a revised manuscript.

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.'*

The uncertainty arising from subsampling is investigated in Appendix C. Table C1 provides an overview of results obtained from randomly generated alternative subsamples. The results are robust, which shows that 6000 samples are sufficient to represent the density in 16 dimensions.

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 section 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.'

The referee correctly pointed out the differences between our approach and the one chosen by Erhardt et al. As already mentioned above, we will emphasize and clarify these differences in a revised manuscript. However, regarding the DO time lags as an outcome of a random experiment draws on the notion of a population. In this framework, it is not our main priority to compute the uncertain sample mean, but instead test whether the given sample of 16 uncertain lags contradicts a population mean equal to zero. A population mean equal to zero is identified with the absence of a systematic lags, which serves as our null hypothesis. The explanatory power of the sample mean U (termed average lag by the referee) lies in the fact that it is the best estimate (point estimate in case of a sample without uncertainty) of the population mean. Since the distributions of the population mean are not required for the line of inference proposed for a revised manuscript, we would exclude them and build the analysis solely on the hypothesis tests. We will test whether the chosen discretization has any effect on the kernel density estimates.

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 distribution 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.'*

We would not present the probability distribution of the population mean anymore in a revised manuscript.

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 (μ) and bs (μ) 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.'*

We would refrain from presenting the probability distribution for the population that is based on the bootstrapping approach. The referee seems to be right that both methods (bootstrapping and using the t-distribution) produce similar (although not the same) results regardless of the process that generated the data. The fact that the

results of the two approaches converge for large sample sizes n can be explained as follows: For large samples, the probability distribution for the sample mean tends to a Gaussian distribution with standard deviation of sigma / sqrt(n) and mean μ, with sigma and μ being the standard deviation and the mean of the population regardless of the population's shape (central limit theorem). In this case, as we are bootstrapping from the cdf induced by the observed sample, μ is given by u and sigma is given by s the sample mean and standard deviation. Hence the bootstrapped distribution of sample means converges to N(u, s/sqrt(n)) for large sample sizes. The t-distribution with n-1 degrees of freedom with (z= u-μ/(s/sqrt(n)) converges to the same gaussian distribution as n increases. In our case, n might still be considered small enough for the approaches to yield different results, but the uncertainties in u and s finally blur these differences. Both approaches rely on strong assumptions – the first that the original population's cdf is approximated reasonably well by the cdf induced by the sample and the second that the original population is Gaussian.

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.'*

The key question we want to address is whether the measured lags between the different proxy variables significantly contradict a population mean equal to (or greater than) zero. If a population mean equal to zero cannot be ruled out, a systemic lead-lag relation cannot be evidenced. For this aim, hypothesis tests constitute the convenient and scientifically well-established tool. Given that we do not present the probability distribution of the population mean, the tests will be key in the analysis. If the lags of the individual DO events were free of any uncertainty, then a single p-value could be computed for the sample (for each test). The uncertainty of the sample immediately translates into an uncertain p-value. Given that we could not find any literature on this specific problem, we propose two possible interpretations of this uncertain p-value. Under different assumptions on the shape of the population (and no assumptions in case of the bootstrap test) the corresponding tests fail to reject the null hypothesis of a population mean equal or greater than zero, for both interpretations. It is the fact that even under different assumptions we achieve the same result that makes our analysis robust. We would keep the section on hypothesis tests but would make an effort to better explain the role of these tests within the framework of our analysis.

'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.'

In our case, the distribution of p-values arises from the uncertainty in the individual lag measurements within the n=16 sample. The distribution indicates the probability that the observed sample corresponds to a certain p-value with respect to the null-hypothesis. The uniform distribution of p-values mentioned by the referee indicates that the probability to realize an n-sample with some p-value p from the population is uniform, given the null hypothesis holds true. Therefore, we do not think comparing the derived p-value distributions to a uniform distribution is meaningful.

'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.'

We agree: The uncertain n=16 sample of DO time lags is represented by 6000 vectors in 16 dimensions (as generated by the MCMC). For each of these vectors a p-value is calculated, for the test whether this vector contradicts a population mean equal to zero. Since the 6000 vectors represent the uncertainty of the sample, the 6000 p-values represent the corresponding uncertainty of the p-value. This enables us to deduce the probability for the sample to significantly contradict a population mean equal to 0.

*'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.'*

We are not sure if we fully understand this comment. We assume that the referee proposes to treat all 6000 * 16 values acquired from the MCMC-sampling as equally meaningful observations of the same quantity and put all of them together into one pot. As already mentioned, we would incorporate an explanation why this is not valid in a revised manuscript. Our objection to this approach draws on the following: There is no physical quantity that would be represented by such a lumped / gathered distribution.

Physically, 16 observations have been made for 16 different DO-events, each of which is uncertain and hence represented by 6000 MCMC-samples. Together, these 16 * 6000 values must be regarded as an empirical probability density in 16 dimensions and not in one dimension. Disregarding this would have severe consequences for further inference. For example: Assume that one tries to observe the outcome of a repeated random experiment. In the first attempt to observe it one is uncertain whether the observation was either 1 or 2. In the second runone observes 2 or 3. Gathering these possible observations together results in a set of observations 1,2,2,3 which corresponds to mean u=2 and a standard deviation of 2/3. However, equation 14 yields four possible vectors (u,s) which are (u=1.5, s=0.5), (u=2,s=2), (u=2, s=0) and again (u=1.5, s=0.5). All four vectors carry the same probability weight. From this, one may compute the expectations <u> = 2 and <s> = $\frac{3}{4}$ of the uncertain quantities u and s .

L396ff (the referee mistakenly wrote 496) : *'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?'*

As mentioned above, we would not present the probability distributions for the population mean in a revised manuscript. However, for sake of clarity: Consider a certain sample with mean u and standard deviation s. According to Equation (12) the u and s induce a probability distribution for the population mean μ centered around u. If the sample is now taken to be uncertain, uncountable many combinations (u,s) induce distributions for μ centered around the corresponding u. According to equation (15) they all contribute to the population mean distribution under uncertainty. Hence, this distribution must be broader than the distribution for the sample mean. Or the other way around, any sample mean is associated with a broad range of possible population means that define a population which potentially has generated the sample.

L406f: *'I think this point deserves a moment of attention: Despite the added layer of uncertainty for the posterior distribution of U Δt and the additional assumptions going into μ Δt 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.'*

The referee mentions a very relevant point here. However, we disagree with the inference that the referee proposes. We have inferred a probability of âĚŸ for the population mean to be less than zero. As far as we understand, the referee interprets this results as an indication that ‘it is likely that the transitions are led by a transition in Ca’. In this sentence ‘the transitions’ apparently refers to all transitions. However, the population mean itself does not allow to make any statement on the probability of a single DO time lag randomly generated from the population to be less than zero - nor about the probability that the true values of individual observed time lags were in fact less than zero. In a revised manuscript we would not present the probability distribution of the population mean but we would still present the probability distribution of the sample mean. Also from the latter one cannot deduce the probability that the true values of all observed time lags were negative. We have tried to make such statements in the last part of our analysis, but the given arguments were inconsistent and would therefore not be shown in a revised manuscript.

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.'*

After thorough review of this section we found inconsistencies in the reasoning. A stringent discussion of this issue is beyond the scope of this answer. We refrain from presenting this approach in the revised manuscript and will return to this in future work.

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?'*

Sorry, we think there's a misunderstanding here. The question is which are the specific DO events that potentially are led by sodium. In the previous paragraph we argued that the most likely configuration is one where 10 events are led by calcium while 6 are led by sodium. However, since these are all probabilistic statements, it makes no sense to indicate which specific events these are.

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).'*

In our opinion, the chosen wording does not conflict with the demands of the referee. We state that the results 'cannot serve as evidence for atmospheric changes to trigger sea ice retreat during DO events' – and this statement does not deny that a review of the investigated data in combination with additional data or other methods will support the hypothesis of an atmospheric trigger. We will nevertheless clarify this point in a revised manuscript.

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.'

We thank the referee for this comment. It is true that the manuscript has a strong emphasize on the methods, while the possible physical mechanisms at work during DO-events are treated only to a limited extent. We would elaborate on this aspect and include this possibility as a potential explanation of the data already in the introduction. There we would add a sentence like: 'Previous studies have found a tendency for calcium to transition before sodium. This may be interpreted as an indication for the atmosphere to trigger a change in the sea ice extent. However, it may also be that both – atmosphere and sea ice – respond to some other trigger, with the atmosphere simply responding faster. If a change in the calcium records was to be the trigger for the change in the sodium concentrations, a lag between the transitions should consistently be detected. We show that this is not the case and therefore argue that the second interpretation is the more plausible one. However, we confirm the tendency of a delayed sodium transition which we interpret as a slower reaction of the sea ice to the original trigger.'

References: Erhardt, T. et al. Decadal-scale progression of the onset of Dansgaard-Oeschger warming events. Clim. Past 15, 811–825 (2019).

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