General Points

We thank the referees for their very thorough and careful review. Following the constructive criticism and valuable feedback, we substantially changed the manuscript. We are convinced that these changes improved the quality and clarity of our manuscript and that they address the referees' objections, questions and suggestions. We do not provide a mark-up version of the changes, because there apart from the introduction all sections have been rewritten at least in parts.

Since both reviews share several general and important aspects, we would first like to elaborate on the changes that have been implemented in response to these general points in a combined way here, 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 agreed with this criticism and integrated the analysis of all pairs of proxies that are reported to show a clear lag-behaviour by Erhardt et al. in the revised version of the manuscript. We did 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 was a lack of guidance and motivation throughout the method section in the original 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 restructured the manuscript and focussed on the key question the manuscript aims to answer: Can we attribute the observed lag-tendencies to a systematic mechanism or can we not distinguish them from randomly occurred lag-tendencies? This question is directly answered in terms of the extended significance tests. Hence, we left aside all additional statistical considerations presented in the original manuscript and put the focus solely on the hypothesis tests in the presence of uncertainty.

In the revised manuscript, we now 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 samples drawn from underlying populations.
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.
4. detailed comparison with the statistical perspective assumed by Erhardt et al. as requested by both referees.

We now 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 removed the following from our manuscript in order to increase its stringency and readability:

- derivation of distributions for the population mean, because it is not required to answer the stated key question.

- we refrained from carrying out all the derivations for empirical densities and used continuous probability density functions instead. The equivalence between empirical densities (as provided by the MCMC) and continuous probability density functions in all our derivations is now 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.
- We removed 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 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 now yields 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 added a paragraph in the discussion section that explains why this cannot be done (l.571).

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. The clear focus on the main line of reasoning in the revised manuscript will facilitate the readers’ understanding.

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 paid special attention to accurately handle dates and units of the climate proxies under study. Also, we carefully reviewed that all quantities are correctly indicated.

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.

Figure1: 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 also added the original reference for the d18O data and the GICC05 age scale in the revised manuscript. We changed the color of vertical lines (and tilted connecting lines) to light gray. We added the time scale to those Figure captions where it was missing. The time scale in Fig1. is indeed BP (before 1950), the ages of the DO events as indicated by Rasmussen (2014) have been converted accordingly.

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.

l.73: We added total numbers (7 years of higher) or 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.

Tab.1 was removed from the manuscript. Fig. 2 in the revised manuscript now takes a similar role. The caption for Fig. 2 was rewritten entirely to avoid the ambiguity pointed out be the referee.

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. This is now clarified in l.72.

('We use the same data and the same probabilistic transition onset detection method as provided by Erhardt et al. (2019).')

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

The revised version of the manuscript does 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 changed 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 in the main text. In the data availability statement we still cite the Pangea reference but we are absolutely willing to change this, if requested.

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' was replaced by 'thinning' and 'deposition rates' were corrected 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.

We introduce the transition onset detection method designed by Erhardt et al. under 3.1 (l.149). In our understanding, the formulation of a linear ramp fit is part of this method and the MCMC is used to handle the posterior probability distributions, which are induced by the stochastic linear ramp model.

We hope this is conveniently explained in Sec. 3.1 and in Appendix B, accordingly.

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 removed 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 \Delta t 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.

This was clarified in Sec. 3.1 and in particular in l. 191 where we explain, that throughout the main text we refer to continuous probability densities instead of empirical probability densities. The derivation of the methods in terms of empirical densities as induced by the MCMC sampler is given in Appendix A.

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

For sake of comparability we included the same DO events previously investigated by Erhardt et al. - all of these events are successfully fitted by the ramp-fit method. We removed the comment on the success criterion of the MCMC sampler, but included an explanation on the data selection (l.129)

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.

We revisited the robustness test carried out by Erhardt et al. and agree with the referee, that this legitimates the application of the Bayesian transition detection method. Furthermore, the focus of this study is the application of hypothesis tests to uncertain data samples. Hence, we removed all considerations regarding the performance of the ramp-fit method and rely on the assessment provided by Erhardt et al.

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

After reconsideration, we believe that hierarchical models could in fact be applied to the data set. We corrected the manuscript accordingly (l.601).

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

Working with empirical density distributions requires to store all values comprised in the representative set of the distribution. The sum is not executed in the sense, that is returns a single value. The computational problems are now explained in more detail in Appendix A and B.

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 B. Tables B1 and B2 provide an overview of results obtained from randomly generated alternative subsamples. The results are robust, which shows that 6000 samples are sufficient to represent the respective densities.

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 emphasized and clarify these differences in the revised manuscript in Sec. 3.4 and in the discussion.

Also, the referee is correct in the sense that the computation of the uncertain sample mean (termed average lag by the referee) adds valuable information without invoking any additional assumptions. Additionally, the study aims to judge whether or not this sample mean can be attributed to a systematic mechanism. Therefore, the hypothesis testing is key to the revised manuscript.

The other methods have been removed from the manuscript to improve the stringency of the argumentation.

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 do not present the probability distribution of the population mean anymore in the 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 ρ(μ) and ρ 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.

Since the revised manuscript focuses on the hypothesis test, we refrained 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. This is – at least partially – due to the central limit theorem.

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 starting point of this study was the observation by Erhadt et al, that transitions in Ca2+ and the annual layer thickness start on average about one decade earlier than their counterparts in Na+ and d18O. We set out to answer the question, if this observed lag can be attributed to an underlying mechanism or whether it may have arisen simply by chance. For this aim, we regard the observed lags (for a given pair of proxies) as observations of independent, identically distributed random variables that have been generated in series of a repeated random experiment. As a next step, we identify the potential systematic mechanism with a bias in the population of this experiment, which finally allows us to test the null-hypothesis of an unbiased population (or even reversely biased with respect to the observation). We have structured the revised version of the manuscript around this natural line of inference and hope, that the revised version now explains well the role and the meaning of the hypothesis tests. We have made an effort to clarify the null hypothesis and the assumptions underlying the applied tests in Sec. 3.3.

Technically, the main obstacle for this line of inference was the uncertainty comprised in samples that were meant to be tested. Hence, in Sec. 3.2. we establish a general framework for the application of hypothesis test to uncertain samples, since we could not find any such framework in the literature. The referee is right – in the case of the t-test and WSR test, this boils down to applying the tests to the individual members of the MCMC sample. However, this is not exactly the case for the bootstrap test. Since the propagation of uncertainty is not straight forward, we have treated this aspect with caution and provide detailed derivations. An interpretation of the uncertain p-value is provided in Sec. 3.2.

‘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 samples. 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.

‘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.’

Remark: In the original version of the manuscript only the lag between calcium and sodium was discussed. For this case, uncertain lags could be derived for 16 events.

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 incorporated an explanation why this is not valid in the revised manuscript. Our objection to this approach draws on the following:

There is no physical quantity that would be represented by such a pooled / 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> = ¾ of the uncertain quantities u and s .

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?

As mentioned above, we do not present the probability distributions for the population mean in the revised manuscript. However, for sake of clarity: Consider a certain sample with mean u and standard deviation s. According to Equation (12) [original manuscript] 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, uncountably many combinations (u,s) induce distributions for μ centered around the corresponding u. According to equation (15)[original manuscript] 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. Due to the substantial changes, this question is not discussed any more in the 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 refrained 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 posed in the original manuscript was *which* are the specific DO events that potentially are led by sodium. This question can in fact not be answered unambiguously, since all statements on leads and lags for individual events are probabilistic. However, the paragraph this comment referred to was removed from the manuscript.

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).

The revised manuscript uses a slightly different wording (l.592). We agree, that at a later stage, the data and the analysis presented by Erhardt et al. may certainly be used in combination with more data ore other methods to draw conclusions on the sequence of events at the onset of DO events. However, in our opinion, the chosen wording does not conflict with the demands of the referee. Our statement 'However, if the common proxy interpretations hold true, our findings suggest that the hypothesis of an atmospheric trigger - either of hemispheric or synoptic scale - for the DO events should not be favoured over the hypothesis that a change in the North Atlantic sea-ice cover initiates the DO events.' refers to the data as presented by Erhardt at al. and additional information may certainly falsify this statement.

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. Given the restructuring of the manuscript and the strong focus on the hypothesis testing the potential conclusions that can be drawn from the analysis are somewhat narrower compared to the original manuscript. That is why we do not elaborate on potential triggers which are not reflected by the investigated proxies in the revised manuscript. Instead, we restrict our conclusions to the key finding, that the observed lag tendency cannot be discriminated from one that has arisen purely by chance and that hence, the hypothesis of an atmospheric trigger should not be favored over the hypothesis of a sea ice trigger at this stage.

Point by point answer to referee 2:

General Comments

1. The authors do not specifically discuss how their approach is different to Erhardt et al 2019. While the manuscript gives the impression that previous studies completely disregarded uncertainties and used expectation values, this is not true for Erhardt et al, who included the uncertainties in a rigorous manner in their own right. The difference lies in the interpretation of the estimated onsets as random

variables. Erhardt et al consider the uncertain samples as measurements of the same fixed quantity (a fixed time lag of Ca and Na at DO onsets), which allows them to simply multiply the individual MCMC posteriors to obtain a single posterior distribution that represents the measurement uncertainty of the fixed time lag. In contrast, the present study interprets the time lag to be a random variable, varying in between DO events. Thus, they cannot multiply the individual posteriors. In the discussion, the authors contrast their approach with the results one would obtain when completely discarding the uncertainties in the onset determination. It is not very surprising that including uncertainty yields hypothesis tests which are no longer significant. Instead, it would be more relevant to highlight the contrasting results of their approach to Erhardt et al.

1. We agree that this should have been described more clearly and discuss the differences in the underlying assumptions explicitly in the Methods section (3.4) and the Discussion section of the revised manuscript. Additionally, in Figure 5 compare the ‘combined evidence’ according to Erhardt et al. with results obtained for the uncertain sample mean (this study). While Erhardt et al. assume that the time lag between the calcium and the sodium transition was a physical constant, such as the speed of light, we argue that climate variability will cause this time lag to differ from event to event. Assuming that the DO mechanism remains unchanged between events, it is very convenient to treat the observed time lags as the result of a random experiment that was performed repeatedly. Even if the results of the hypothesis tests were not surprising given the uncertainties in the observations as the referee states, they still contradict the conclusions of Erhardt et al. who state: ‘Taken at face value, this sequence of events suggests that the collapse of the North Atlantic sea-ice cover may not be the initial trigger for the DO events and indicates that synoptic and hemispheric at- mospheric circulation changes started before the reduction of the high-latitude sea-ice cover that ultimately coincided with the Greenland warming’. Our tests indicate that one cannot discriminate the observed tendency for a sodium / d18O lag with respect to Ca or the annual thickness from a lag that has occurred purely by chance and in the absence of a systematic mechanism. In this way, our results from the hypothesis tests add valuable information to the debate on the succession of events at DO onsets and we conclude contrarily: ’However, if the common proxy interpretations hold true, our findings suggest that the hypothesis of an atmospheric trigger - either of hemispheric or synoptic scale - for the DO events should not be favoured over the hypothesis that a change in the North Atlantic sea-ice cover initiates the DO events.’ (l.592)

2. The authors speak of a rigorous propagation of uncertainties to p-values, among other things. To achieve this, they introduce probability distributions of the mean and other test statistics, which are formally represented using delta distributions arising from the empirical sample. In practice, these distributions are all computed by summing a random sample from the individual MCMC posteriors, which is essentially bootstrapping if I understand correctly. Can the authors comment on how the results of this work would be different if they simply summed the individual MCMC posteriors (which would be equivalent to bootstrapping as well), and then looked at the arising distribution of the lag, determining the probability of a lag >=0? This would be much simpler and the obvious alternative approach to the multiplication of the posteriors by Erhardt et al. It would probably also give a non-significant result regarding a Ca2+ lead.

Remark: In the original version of the manuscript only the lag between calcium and sodium was discussed. For this case, uncertain lags could be derived for 16 events.

2. We are not sure if we fully understand this comment. The referee proposes to ‘sum the individual MCMC posteriors and look at the arising distribution of the lag’. We interpret this as follows: For each DO-event the MCMC samples 6000 lags from the corresponding posterior distribution. The referee proposes to put all these 16*6000 values together in one pot and regard them as observations of the

same quantity. Our objection to this approach draws on the following:

There is no physical quantity that would be represented by such a pooled / 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> = ¾ of the uncertain quantities u and s .

We have included a paragraph on this issue in the discussion (l.571).

The referee states: ‘In practice, these distributions are all computed by summing a random sample from the individual MCMC posteriors, which is essentially bootstrapping if I understand correctly.’ Here, as well we are not sure if we understand this statement correctly. In our understanding a random sample from the posterior lag distribution for a single DO event corresponds to the set of 6000 values, sampled by the MCMC from the corresponding posterior probability distribution. The empirical density distribution that is induced by this MCMC-sample approximates the true posterior distribution for the DO-event in the sense of equation (4) (l. 170). The MCMC-sampling procedure that gives rise to the sample of 6000 values is not equivalent to bootstrapping. The MCMC algorithm is used to sample from a continuous probability density, because the distribution is too complex to use it directly for inference. Bootstrapping is used to generate synthetic samples from a finite size sample by ‘drawing with replacement’, not to sample from a continuous probability distribution.

Appendices A and B are dedicated to the empirical density distribution and how they arise from the MCMC sampling procedure.

3. I am wondering in what way the hypothesis tests introduced in Sec. 3.6-3.7 are necessary and add to the results? When reading the manuscript, I was surprised that hypothesis tests were introduced after distributions for the sample or population mean had already been given, which would directly allow to test the hypothesis of a mean >=0?

The starting point of this study was the observation by Erhadt et al, that transitions in Ca2+ and the annual layer thickness start on average about one decade earlier than their counterparts in Na+ and d18O. The current version of the manuscript sets out to answer the question, if this observed lag can be attributed to an underlying mechanism or whether it may have arisen simply by chance. For this aim, hypothesis tests constitute the right tool. We have preferred this line of inference over the derivation of an uncertainty distribution for the population mean, because it is appears natural at this stage and gives a clear answer to a well posed question.

4. Why did the authors restrict the analysis to Na and Ca, and omitted an analysis of the offsets of d18O and lambda? This would be an important consistency test, and might even be more relevant since lambda has a very direct meaning as a proxy, and d18O is the most important of all proxies.

This point was addressed in our first statement, we added the investigation of the other variables in the revised manuscript

Specific comments:

L15ff: I think the conclusions should be stated differently. In my interpretation, the analysis does not contradict a lead or lag. Rather, as a result of the large uncertainties in estimating the individual onset timings, it cannot be excluded that there are in fact no leads or lags. Similarly, on the grounds of the uncertainties in the individual event timings, there is not enough evidence to conclude that atmospheric reorganization systematically preceded sea ice retreat for all events.

We agree and would adjusted the abstract in the sense of the comment by the referee.

L34-35: The interstadials lasted up to 10 millennia, with 1.5 millennia being the average.

Thank you, we changed this to: 'The abrupt warming is followed by gradual cooling over centuries to millennia, before the climate abruptly transitions back to cold conditions'.

L130: Why does the transition detection fail for some events?

For sake of comparability we included the same DO events previously investigated by Erhardt et al. - all of these events are successfully fitted by the ramp-fit method. We removed the comment on the success criterion of the MCMC sampler, but included an explanation on the data selection (l.129)

L150: How would this depend on the autocorrelation of the noise? The Na and Ca records might have different noise structure and thus there is the potential for systematic biases indeed.

We revisited the robustness test carried out by Erhardt et al. and agree with the referee 1, that this legitimates the application of the Bayesian transition detection method. Furthermore, the focus of this study is the application of hypothesis tests to uncertain data samples. Hence, we removed all considerations regarding the performance of the ramp-fit method and rely on the assessment provided by Erhardt et al..

Eq. 7: Maybe the authors can elaborate specifically in the text what this approximation does. Since the order of MCMC samples for each event is arbitrary, by associating the i-th MCMC sample for every event, it seems like it is just a random sampling of m=6000 points in the joint space. If this is the case, why not simply sample randomly in the first place, and why not choose many more than m=6000 points? Or is this rather done in order to simplify the notation?

The referee is correct. Associating the i-th individual member of every event's MCMC-sample is a random choice and is indeed the simplest in terms of notation. The impact of this choice is discussed in detail in the Appendices A and B. In short, the sample is large enough such that the final results are robust against any sort of randomization of this choice.

L278ff: I am not sure why the authors say that they are only given "relative" data, since they estimate the onsets in the two proxies. Furthermore, since until this point the data was already given exclusively as onset timing differences, I wonder why it is necessary to introduce "paired samples" now? This is a bit confusing to the reader.

We agree that the sentence is not very helpful and removed it. In fact, the lags constitute relative data in the sense that there is no absolute time for the onset of the DO event. The timings of two transition onsets detected in different proxies in this study receive meaning when being compared to each other. Further, a measurement of the timing of the transition in one proxy from the i-th DO event cannot be compared to the timing of the transition in another proxy from the j-th DO event, that's why the data is paired.

L348: I might have missed this somewhere, but how is the distribution of the test statistic under the null hypothesis constructed?

The original manuscript did not properly introduce the null distribution of the WSR test. It is mostly a combinatorical problem but must be computed explicitly. In the revised manuscript, in line 334 we provide guidance for the construction of the WSR null distribution, which can as well be found in lookup tables.

L396: Could the authors explain more explicitly why the sample and population mean distributions are so different, and what this means for the interpretation of the results? Which distribution should be preferred?

Note that we do not show the population mean's distribution anymore in the revised manuscript. For sake of clarity, we nevertheless respond here: Assuming that all DO events followed the same physical mechanism, we regard the observed time lags as the outcome of a repeated random experiment with the randomness being due to climate variability. A random experiment is fully characterized by its population. The population mean indicates systematic leads or lags and hence is the decisive quantity. The explanatory power of the sample mean U (termed average lag by the referee) is limited to the fact that it is the best estimate (point estimate in case of a sample without uncertainty) of the population mean; conclusions based on the sample lag hence need to be accompanied with information on how likely it would be to obtain this sample mean from a population characterized by a null hypothesis

To give a simple example: If you want to judge whether a coin is biased, it is not sufficient to toss the coin 16 times (with head assigned a value of 0 and tail assigned 1) and report on a mean of, say, 0.3, arguing that it is different from 0.5, the *expected outcome* if the coin is unbiased. Rather, one would have to compute the probability of obtaining a mean of 0.3 or smaller under the assumption that the coin is unbiased. If this probability is small, one can (at a reported significance level) argue that the coin is biased.

Also for the reasons given in the introductory statement, we hence prefer applying hypothesis tests to the uncertain sample over deriving the distribution for the population mean.

The population mean distribution must necessarily be broader than the sample mean distribution. If the observation was free of any uncertainty, the sample mean would be a scalar. The t-distribution would induce a probability distribution for the potential population mean around this sample mean. Now, with the sample mean being uncertain itself, the certain sample mean probability distribution for the population is convoluted with the sample mean distribution and the convolution of two functions is always broader than either of the functions being convoluted with one another.

L441ff: Here the authors introduce another simple method to address the likelihood of a systematic Ca2+ lead. I think it would be more coherent if this were moved to the Results Section. Furthermore, I find the conclusions from this simple calculation too confident, and in spirit contradictory to the interpretation of the other results. The probability of obtaining exactly n=16 events with a Ca lead will always be very small when there is a relatively large measurement uncertainty of the individual lags (spanning both positive and negative values), even if all individual posteriors would be clearly centered at negative values. Just like the probability of flipping 16 out of 16 heads is still very low for a strongly biased coin. This does not allow one to contradict the hypothesis that all events would follow the same pattern with a Ca2+ lead. For the data at hand, all but two events show posterior distributions centered at a negative value. Maybe the authors could write instead that from the MCMC posteriors it seems unlikely that all DO events occurred with a preceding abrupt change in Ca2+. However, this might merely reflect the fact that due to the large uncertainty in estimating the onsets, the individual MCMC posteriors have significant support for positive lags as well. Arguing like this would also be much more in line with the authors' earlier statements that they cannot infer an absence of causality from their non-significant tests.

The comment of the referee made us review the reasoning invoked in this section. In fact, we find inconsistencies and therefore not present this at all in the revised manuscript. We will try to reconcile the inconsistencies and derive further statements on a potential causal relationship between the two transitions at a later stage.

Figure 1 and Table 1: Just to be sure, can the authors confirm that the time scale they use really is years BP (before the year 1950 AD), and not years b2k (before the year 2000), which is what is commonly used in GICC05?

According to the data source (Erhardt et al. 2019), the age scale in Figure 1 is in fact given in years BP. The ages of the DO events, which are due to Rasmussen 2014 have been converted accordingly.

L18: ...holds true, the we conclude...

the sentence was removed

L98: Instead of "compression" rather say: ...due to the thinning of the annual layers in the core.

was corrected (l.117)

Figure 3: Maybe it would be good to choose a different color for the null hypothesis in panel b. Otherwise it gives the impression that it corresponds in some way to the blue distribution in panel a.

Thank you - this is true. This was adjusted – now Figure 4.

Eq. 14 and Eq. 16: I am wondering whether "u" in the second delta function should be replaced by the empirical mean sum(y_i)/n?

Mathematically this should be the same, since the first 'delta function' forces sum(y_i)/n to assume the value u. The equations now only appear in Appendix A. For sake of readability, we did not change the notation.

L235: Maybe "marginal distribution" would correspond better to the nature of this distribution?

The sentence was removed from the manuscript.

Eq. 17: What is the notation u_j and s_j?

The corresponding equation was removed from the manuscript.

Eq. 20: I am unfamiliar with this definition of a p-value. Is the integration not simply over all phi < phi_0 (for a one-sided test)? The definition here could lead to rather strange results for very asymmetric and long-tailed distributions.

The definition provided in Equation (20)[original manuscript] corresponds to the 2-sided test but in fact only holds true under certain conditions. We now present a more common equation (11) for the one-sided left-tailed p-value (l.251).

L387: missing delta in the sum.
Was removed.

Figure 4: Larger fonts and panels would be nice for better visibility.
We will adjust the figure to improve readability.

Eq. 35: It would be good if the authors could point to where the individual probabilities in the product come from.
This section was removed from the manuscript.

References:

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