Clim. Past Discuss.,
https://doi.org/10.5194/cp-2020-136-AC2, 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

General Points

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 in-

tegrate 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 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.'

We agree that this should have been described more clearly and would discuss the differences in the underlying assumptions explicitly in the Methods section of a revised manuscript. Additionally, in Figure 4 we would include the 'combined evidence' according to Erhardt et al. instead of the probability distribution of the population mean, and provide a detailed explanation of the differences. 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 16 observed time lags as the result of a random experiment that was performed 16 times. 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: 'They [probability density estimates] clearly show that, on average, both the reduction in terrestrial aerosol concentration and the increase in annual layer thickness precede the reduction in sea-salt aerosol for all stages of the transition'. Our tests indicate that one cannot differentiate the observed tendency for a sodium lag from the result of a random experiment with population mean equal to zero. In this way, our results from the hypothesis tests add valuable information to the debate on the succession of events at DO onsets.

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 sum-*

ming 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.'

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 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 . 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 (3). 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.

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?'*

We hope this was clarified in the explanation of our proposed restructuring. In brief, we think that the hypothesis tests provide the most straightforward way to answer the question whether the observed sample of lags is inconsistent with an underlying population mean lag larger than or equal to 0. Our results show that, under consideration of the uncertainties present, it is not inconsistent with that null hypothesis, which adds an important layer of information to the debate on the trigger of DO events.

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'll add 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 adjust 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 would change 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?'*

The study only takes into account those events for which the rejection rate during the MCMC-sampling process was below 70

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 must admit that we did not consider this. We would add the same investigation as carried out for the influence of the noise amplitude for the correlation coefficient of the noise. This would also meet the demands formulated by referee 1 regarding this issue.

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 analysis in Appendix C shows that 6000 vectors from the 16 dimensional probability density distribution suffice to capture the distributions features. Randomizing the choice of these vectors does not influence the results. Note that this would not appear in the manuscript's main text as we would refrain from using the empirical density notation in the revised methods section.

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 would delete 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 calcium onset and the sodium onset only receive meaning when compared to each other. Further, a measurement of the timing of the Ca2+ transition from the i-th DO event cannot be compared to the timing of the Na+ transition of 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 distribution of the test statistic under the null-hypothesis has a very lengthy expression and therefore was not explicitly given. It is true that Eq.(29) is missing some explanation here. We would add the functional form of P(w'$_i$).

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 would 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 would therefore not present this at all. 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?'*

In Figure 1 we used the data provided by Erhardt et al. 2019 who use BP (before 1950) as age reference. This is the only data source we are aware of that provides d18O data with 10y resolution. In Table 1, however, our indication is fact wrong. These are ages

in b2k (before 2000).

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

Will be corrected.

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

Will be corrected.

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.'*

This is true and will be adjusted.

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. However, we would change this, since the formulation including $\sum(y_i)/n$ seems more instructive.

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

In view of the proposed changes, this is obsolete. Here the term 'expected' was chosen on purpose, since we consider the distribution $\rho(\mu)$ as a function of the two variables u and s. Since u and s are uncertain, we can compute an expectation of $\rho(\mu, u, s)$ by averaging over (u,s) with the weight $\rho(u, s)$.

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

They are missing an ^{emp} – definition is then given in the previous line.

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 corresponds to the 2-sided test. Equation (20) is the mathematical formulation of what the text states: 'Given a statistic of a certain sample realization $\phi(x - 0) = \phi_0$, the p-value is indicates the cumulative probability for obtaining a more extreme $\phi_1(X_1)$ from a second sample X_1 , provided H_0 was true.'

L387:*' missing delta in the sum.'*

Will be corrected.

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.'*

Given the already extensive mathematical formulation of our methods, we tried to keep it short at this point. Given the comment we would now add more explanation to this formula.

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