Author's Response

July 1, 2021

Contents

1 Answer to Referee 1 1

1.1 General Remarks . . . 1

1.2 Statistical Setting . . . 3

1.3 Argument by the referee in favour of criterion (1) . . . 5

1.4 Arguments against criterion (1) . . . 6

1.4.1 Illustrative example . . . 6

1.4.2 Setup designed by referee 1 . . . 7

1.4.3 χ^2-test . . . 10

1.5 Point-by-Point answer to the referee . . . 11

2 Answer to Referee 2 17

2.1 General Remarks . . . 17

2.2 Point by Point answer to the referee . . . 18

1 Answer to Referee 1

1.1 General Remarks

First of all, we would like to thank the referee for the careful second review of our manuscript. Before we address the comments point by point, we will discuss the main criticism, namely the decision criteria we use to reject or not reject the respective null-hypotheses.

We would like to emphasize that we acknowledge the arguments brought up against our decision criteria and that after a careful review we have refined the explanation of the criteria based on the referees input. Here, for sake of clarity, we will first summarize the status of the discussion. Subsequently, we repeat the argument presented by referee 1 in favour of the criterion (1). We will then examine this argument in detail and explain why we are still convinced that the combination of the criteria (2) and (3) yields a more meaningful assessment than criterion (1). Nonetheless, due to the comments of the referee, with respect to the decision criteria we changed the manuscript as follows:

- We added a paragraph that motivates the propagation of uncertainties with an example, where averaging out uncertainties yields undesired results. (l.282ff)
- We added two sentences on the interpretation of the criteria (2) and(3): 'Given the measurement uncertainty the quantity $\pi(\hat{P} < \alpha)$ indicates the informed estimate of the observer that the true value of the measured sample is in fact statistically significant with respect to the null hypothesis. Thus, the first criterion assesses how 'strongly' the uncertain sample contradicts the null hypothesis and the second criterion evaluates the likelihood of the uncertain sample to contradict the null hypothesis. Depending 315 on η, in many cases both criteria will yield the same decision. If not, the specific situation determines which of the criteria is more convenient.' (l.311)
- In the previous version of the manuscript, we referred to the criteria (2) and (3) as 'criteria for significance' in two instances. This was changed, as the they consitute decision criteria rather, while the significance of the sample cannot be ultimately assessed.

  l.515: 'neither of the two criteria for significance' was replaced by 'neither of the two criteria for rejecting the null hypothesis'.

  l.525: 'The first criterion for significance is hence not met by any of the pairs. Also, the probabilityfor significance is below 60% for all pairs and all tests as shown by the pie charts, so also the second criterion is missed.' was replaced by

  'Also, the probability for significance is below 60% for all pairs and all tests as shown by the pie charts. Thus, for all proxy pairs and for all tests the formulated decision criteria do not allow to reject the null hypothesis of pairwise unbiased populations.'

- In the same sense, in line 102 we changed:

  'This changes the results from significant to non-significant when compared to averaging out these uncertainties at individual transition lags'

  to

  'If detection uncertainties are averaged out at the level of individual transition lags, temporal delays in the δ^{18}O and Na^{+} transitions with respect to their counterparts in Ca^{2+} and the annual layer thickness are indeed pairwise statistically significant. In contrast, under rigorous propagation of uncertainty several tests consistently fail to reject the null hypothesis across all considered pairs of proxies.'

- We emphasize more clearly than before that not-rejecting the null-hypothesis does not provide evidence against the alternative hypothesis, in particular given the large uncertainties in the observations. l.561: ' We emphasize that our results must not be misunderstood as evidence against the alternative hypothesis of a systematic lag. In the presence of a systematic lag ($\mu < 0$) the ability of hypothesis tests to reject the null hypothesis of no systematiclag $((H_0 : \mu = 0))$ depends on sample size n, the ratio between the mean lag $|\mu|$ and the variance of the population, and on the precision of the measurement. Neither of these quantities is favourable in our case and thus, it is certainly possible that we failed to reject the null hypothesis despite the alternative being true.

1.2 Statistical Setting

We are given a sample of observed time lags $\mathbf{\Delta t}^{p,q} = \Delta t_1^{p,q}, ..., \Delta t_n^{p,q}$ between the proxy variables p and q, where each observation stems from a different DO event. We assume that the process which generated these time lags is qualitatively the same, such that we can think of the variable ΔT as a random variable that will assume a specific value Δt when a DO event occurs, that is, when the random experiment is performed. The random experiment is characterized by its population $\mathcal{P}_{\Delta T}$. The sample $\mathbf{\Delta t} = (\Delta t_1, ..., \Delta t_n)$ of observations enables us to test hypothesis regarding $\mathcal{P}_{\Delta T}$.

In classical hypothesis testing, the realizations of the random variable Δt_i are assumed to be measured (or observed) with infinite precission. That is, each Δt_i is assigned a scalar value in the observation process. However, in our case each Δt_i can be estimated only with limited precission. Instead of scalar values we characterize the Δt_i by means of probability density distribution $\rho_{\Delta t_i}(\Delta t_i^*)$ induced by the linear ramp model.

Notation

Due to the uncertainty in the measurement process the true value Δt_i must be regarded as a random variable whose probability is characterized by $\rho_{\Delta t_i}(\Delta t_i^*)$. We denote with Δt_i^* possible values for the true value. The randomness encoded in Δt_i has nothing to do with the initial random experiment $\mathcal{P}_{\Delta T}$ any more, but is only due to measurement uncertainty. We therefore introduced the term 'uncertain variable' in the manuscript and called $\rho_{\Delta t_i}$ an uncertainty distribution.

$\rho_{\Delta t_i}(\Delta t_i^*)$ quantifies the uncertainty about the individual Δt_i in a Bayesian sense, indicating how plausible or probable a certain value Δt_i^* for Δt_i is in view of the data. Importantly, the measurement of the Δt_i cannot be repeated. Therefore, the $\rho_{\Delta t_i}(\Delta t_i^*)$ must really be understood as a measure of plausibility in view of the data and not as the probability to obtain a value Δt_i^* in a repeated process. To distinguish between PDF's that characterize a true random experiment in a frequentist understanding and those that quantify uncertainty (in the Bayesian sense), in the manuscript we introduced the term uncertainty distribution for the latter ones.

This being said, one needs to find a way how to incorporate this uncertainty in the assessment of an hypothesis on the population $\mathcal{P}_{\Delta T}$, and in particular in corresponding statistical hypothesis tests. Three different approaches are discussed in our manuscript.

1. Averaging out the uncertainty of the measurement at the level of the individual Δt_i's yields a sample comprised of scalars.

$$\mathrm{E}(\mathbf{\Delta T}) = (\int \Delta t_1^* \rho_{\Delta t_1}(\Delta t_1^*)\, d\Delta t_1^*, ..., \int \Delta t_{16}^* \rho_{\Delta t_{16}}(\Delta t_{16}^*)\, d\Delta t_{16}^*). \quad (1)$$

   Classical hypothesis tests can be applied to such an expected sample without any further ado. However, this approach is associated with a loss of information. Furthermore, it is insensitive to the degree of uncertainty involved in the problem and would thus likely lead to overconfidence and to too easy rejection of null hypotheses. We argue that a setup to test statistical significance must account for uncertainties of the kind treated in our manuscript. In particular it must at least in principle be possible that the uncertainties are so large, that the null-hypothesis cannot be rejected anymore.

The other possibilities rely on the propagation of the uncertainty to the p-value based on the general notion that any function of a random variable constitutes a random variable itself:

$$Y = f(X) \rightarrow \rho_Y(y) = \int \delta(f(x) - y)\, \rho_X(x)\, dx. \quad (2)$$

This allows to compute an uncertain p-value, whose plausibility is indicated by

$$P_{val}(\mathbf{\Delta T}) \sim \rho_{P_{val}}(p_{val}^*) = \int \delta(p_{val}(\mathbf{\Delta t}) - p_{val}^*) \rho_{\mathbf{\Delta T}}(\mathbf{\Delta t})\, d\mathbf{\Delta t}, \quad (3)$$

where we use the notation $P_{val}(\mathbf{\Delta T}) \sim \rho_{P_{val}}(p^*_{val})$ to indicate that the random variable $P_{val}(\mathbf{\Delta T})$ follows the distribution induced by the density $\rho_{P_{val}}(p^*_{val})$. Note that in Eq. 3 we omitted the intermediate step of propagating the uncertainty to the test statistic. In order to decide between acceptance and rejection of the null-hypothesis based on the uncertain p-value given by the random variable P_{val}, we formulated two criteria:

2. If the probability for the uncertain P_{val} to be less than the chosen significance level α exceeds a certain threshold η,

   $$P(P_{val} < \alpha) \overset{!}{>} \eta, \tag{4}$$

   the null-hypothesis shall be rejected ($\alpha = 5\%$ and $\eta = 90\%$ were chosen in the manuscript). As explained above, we emphasize here that the PDF $\rho_{P_{val}}(p_{val})$ quantifies the uncertainty that we have about the p-value due to the *inability* to measure the individual realizations Δt_i precisely. One could therefore also say that, in order to reject the hypothesis, we want to at least be certain to a level of 90% that the true value of the uncertain sample in fact significantly contradicts the null-hypothesis and thus, that the true p-value is less than α. Otherwise, we prefer not to reject the hypothesis. We would also like to emphasize that a statement on the level of certainty about the significance of the sample with respect to the null hypothesis does not violate the notion of a p-value and its typical interpretation.

3. As a second option we proposed to compare the expected p-value

   $$\mathrm{E}(P_{val}) = \int \rho_{P_{val}}(p^*_{val})\; p^*_{val}\; dp^*_{val}, \tag{5}$$

   to the a priori chosen significance level. In general, the p-value associated with a sample is a measure for the extremeness of the sample with respect to the null-hypothesis. Thus, the expected p-value reflects the overall extremeness of the uncertain sample with respect to the null hypothesis. $P(P_{val} < \alpha)$ only gives information about how likely the sample is to contradict the null-hypothesis at the chosen significance-level. $\mathrm{E}(P_{val})$ takes into account how strongly potential values for $\mathbf{\Delta T}$ that are assigned probability larger than zero contradict or support the null hypothesis. Therefore, these two quantities should be considered in combination. We proposed to reject the null-hypothesis if the expected p-value is less than α

   $$\mathrm{E}(P_{val}) \overset{!}{<} \alpha. \tag{6}$$

1.3 Argument by the referee in favour of criterion (1)

The referee criticizes that the criteria (2) and (3) do not comply with the general meaning and common understanding of the significance level: 'In the textbook meaning of the significance level α, α denotes the probability that the significance test falsely rejects the null-hypothesis given that it is in fact true. Which is the meaning that almost all readers of the study will be familiar with.'

First, we would like to clarify that the criteria we introduced are 'decision' criteria, they are not significance criteria. The significance of any possible value for $\mathbf{\Delta t}$ is decided based on the comparison of the corresponding p-value $p_{val}(\mathbf{\Delta t})$ with the significance level α. Propagating the uncertainty associated with $\mathbf{\Delta t}$ we obtain probabilities for both, the sample being significant and the sample being non-significance w.r.t. to the null-hypothesis. A decision must hence be taken between rejecting or not rejecting the hypothesis, even though the significance of the sample cannot be assessed unambiguously. It seems appealing to place the same requirement on the decision criteria as on the significance criterion itself. Namely, that the decision criteria should reject the hypothesis with probability equal to the significance level, as formulated by the referee.

This being said, we would like to clarify that the interpretation of the p-value given by the referee is in our view not precise. The statement that α indicates the rate of wrongly rejecting the null-hypothesis may hold in many cases, but it does not hold if a hypothesis on a parameter of a population is formulated in terms of an inequality. E.g. in our manuscript the null hypothesis assumes a population mean $\mu \geq 0$. If μ happens to be as large as, say, $\mu = 3$, the null-hypothesis is obviously fulfilled. Nevertheless, the chances for rejection are smaller than the significance level. Only in the 'least favourable' case that still complies with the null hypothesis, namely $\mu = 0$, the probability for rejection is equal to the significance-level. The significance level α thus constitutes an *upper limit* for the probability of wrongly rejecting the null-hypothesis.

Given an uncertain sample, can one formulate a decision criterion that decides between rejection and acceptance of the null hypothesis and that rejects wrongly at most with a probability of the significance level? The referee argues that the first decision criterion would posses these characteristics. We will show in the following why this is not the case and why instead the assessment of the null-hypothesis should be based on the criteria (2) and (3) in cases where the sample is uncertain (in the sense defined above).

1.4 Arguments against criterion (1)

Whenever one averages out quantification of uncertainty, information is obviously lost. Consider a goalkeeper that knows that an opponent player always shoots the ball either in the left or right corner, but never in the middle. If the keeper averaged over this uncertainty distribution, he would always stay in the middle of the goal but never save the ball. Especially when dealing with bimodal or multimodal distributions, averaging can lead to very misleading results.

1.4.1 Illustrative example

To give an example more closely related to the situation in our manuscript, consider the following:

- Let $\mathbf{x} = (x_1, ..., x_n)$ denote the true value of a sample generated from a population $\mathcal{P}_X$ that is known to be Gaussian.
- Assume that in the measurement process you can only observe the absolute values $y_i = |x_i|$. Thus, there is a 50% chance that $x_i = y_i$ and another 50% chance for $x_i = -y_i$ for each individual x_i.
- Assume the low standard deviation of the sample $\mathbf{y} = (y_1, ..., y_n)$ allows to almost certainly exclude that the true values x_i have pairwise different signs. Hence the true value of the sample is either $\mathbf{x} = (y_1, ..., y_n)$ or $\mathbf{x} = (-y_1, ..., -y_n)$.
- Assume you want to test the null hypothesis that the population mean μ is equal to zero.
- If you apply criterion (1) you would average over the uncertainty distribution to obtain the expected sample $\mathrm{E}(\mathbf{X}) = (0, ..., 0)$. Application of the t-test to the averaged sample would in this case always lead to acceptance of the null-hypothesis. Thus, it would neither fulfill the requirement stated by the referee nor deliver any other meaningful insight.
- Application of criterion (2) in turn will yield the correct decision because $p_{val}(y_1, ..., y_n) = p_{val}(-y_1, ..., -y_n)$ and hence, one either finds $P(p_{val} < 0.05) = 1$ or $P(p_{val} < 0.05) = 0$.
- The expected p-value that is considered in criterion (3) coincides with the p-value of the true value of the sample and would likewise yield the correct decision.

This example shows that averaging out the uncertainty on the level of the observations can lead to meaningless decisions, in particular in the case of not unimodal uncertainty distributions. In fact, the posterior distributions of the transition onsets of individual proxies are multimodal for many of the DO transitions under study (for example see Fig.3 manuscript where the posterior distribution for the transition onset of calcium at the GI-12c onset is shown - the figure is included below). If we strictly followed the paradigm of averaging out uncertainties on the observational level, then this should consequently be done prior to the computation of the transition onset lag $\Delta t^{p,q}$ between the proxies p and q, which is a function of the two uncertain transition onset times of the proxies p and q. This would again alter the results of the analysis.

1.4.2 Setup designed by referee 1

Above we have given an example that motivates the propagation of the measurement uncertainty to the level of the p-value and shows that criterion (1) does not necessarily fulfill the requirement suggested by the referee. Here, we discuss the setup introduced by the referee to argue in favour of criterion (1). We argue that the setup must be interpreted slightly differently with the consequence that it cannot serve as an argument for criterion (1).

The referee considers a Gaussian population $P_X = \mathcal{N}(\mu_X = 0, \sigma_X = 1)$ and realizes a sample $\mathbf{x} = (x_1, ..., x_{16})$ from this population. We will denote the true value of the sample by $\mathbf{x}$ in the following. $\mathbf{x}$ corresponds to the ΔT_n in the referees' comment. Measuring the values x_i necessarily involves uncertainty, which referee 1 models as a second-level normal distribution around the true value. It is important to note here that this second level distribution does not constitute an analogue to the uncertainty distributions $\rho_{\Delta t_i}(\Delta t_i^*)$ in our manuscript, which quantify the plausibility that the true value Δt_i equals Δt_i^*. This is explained in detail below. The most obvious difference is that the distributions introduced by the referee are designed such that the expected value necessarily coincides with the true value, but this cannot be guaranteed in the real application case.

Again, if x_i can be observed directly and without the need to introduce anything like the ramp fit model, the measured value y_i will deviated from the true value x_i due to measurement uncertainty. Typically (as done by the referee) one assumes a Gaussian distribution for the probability to measure y_i if the true value is x_i:

$$\mathcal{P}_{Y_i|x_i}(Y_i|x_i) \sim \mathcal{N}(x_i, \sigma_{\text{obs}}), \tag{7}$$

where σ_{obs} quantifies the measurement uncertainty in the setup. Correspondingly, the error $\Delta_i = x_i - Y_i$ that you make in the measurement follows a normal distribution as well

$$\mathcal{P}_{\Delta_i|x_i} = \mathcal{N}(0, \sigma_{\text{obs}}). \tag{8}$$

Generally, $\mathcal{P}_{\Delta_i|x_i}$ is in fact independent of x_i such that $\mathcal{P}_\Delta = \mathcal{N}(0, \sigma_{\text{obs}})$ conveniently describes the errors $(\Delta_1, ..., \Delta_{16})$ as i.i.d. random variables.

If σ_{obs} is known, from a measured value y_i you can in turn deduce a probability distribution that quantifies the probability that the true value x_i is given by some x_i^*:

$$P_{X_i^*}(x_i = x_i^*|y_i) \sim \mathcal{N}(y_i, \sigma_{\text{obs}}). \tag{9}$$

We see here that the expectation of this distribution, $\mathrm{E}(X_i^*)$, does not coincide with the true value x_i but instead with the measured value y_i. These considerations are illustrated in Fig. 1.

We assume that the referee had in mind something like repeated measurements of the true value. Repeated measurement would indeed correspond to sampling from the distribution of measured values given a true value $\mathcal{P}_{Y_i|x_i}$. However, the situation in our study is such that the uncertainty distribution for the true value must be quantified after a single measurement. Assuming that multiple measurements of individual true values x_i were possible, then the measured values $y_{i,j}$ would in the referees setup be distributed normally around the true value (see Eq. 7) where j indicates the j-th repetition of the measurement of the i-th true value. From the set of m measurements of x_i: $\mathbf{y_i} = (y_{i,1}, ..., y_{i,m})$ the mean of the distribution $\mathcal{P}_{Y_i|x_i}$ - and hence the true value x_i - can be estimated. Either a point estimate can be derived as done by the referee by taking

$$x_i^* = \bar{Y}_i^m = \frac{1}{m}\sum_{j=1}^{m} y_{i,j} \tag{10}$$

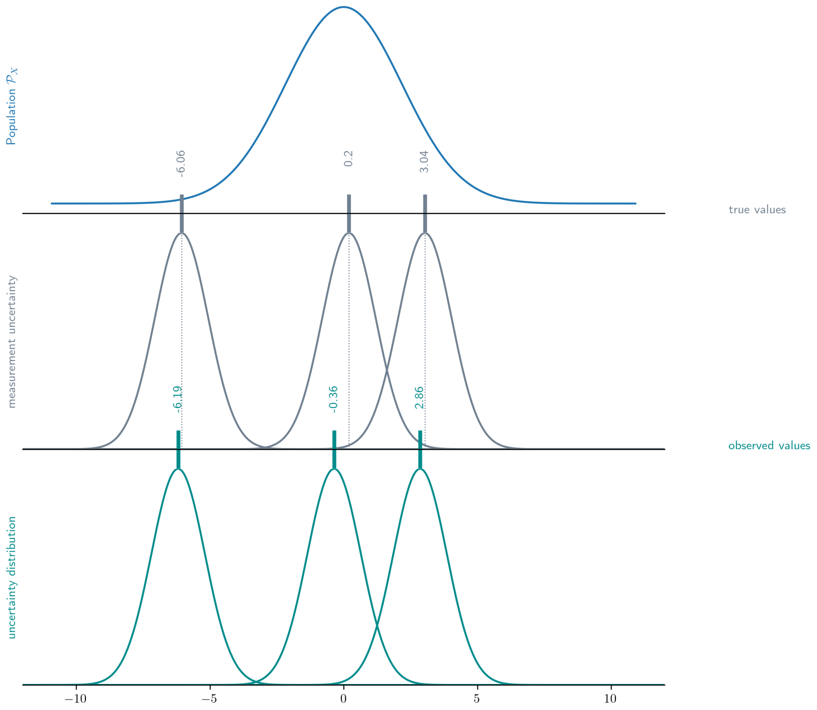

Figure 1: Illustration of measurement uncertainty. On the top level, there is the population that characterizes the random experiment. From this population, three true values are realized, measuring each is associated with uncertainty. The probability to measure y_i if x_i is the true value is shown in light gray in the second level. When the measurement is executed, one value is realized from each of these distributions - the measured value which is indicated in dark green. Given that the uncertainty of the measurement process is known, one can then specify an uncertainty distribution for the true value, based on the measured value (green Gaussian distributions in the bottom panel).

or a distribution $\mathcal{P}_{X_i^*|y_{i,1},...,y_{i,m}}(x_i^* = x_i)$ for the estimator x_i^* can be derived by means of the t-distribution (see original version of the manuscript). This uncertainty distribution, which estimates the true value x_i based on m measurements $y_{i,1},...y_{i,m}$, is much narrower than the distribution $\mathcal{P}_{Y_i|x_i}$. So in the example provided by the referee, the criterion (1) should be compared with the application of the criteria (2) and (3) to the uncertainty distributions that can be derived for the true value from repeated measurement via the t-distribution (assuming that the distribution for Δ were Gaussian). The referee might have had in mind that in the case of repeated measurement a test statistic $T(\mathbf{x})$ should be computed for each measured sample $T_j = T(y_{1,j},...,y_{16,j})$ and that the $T_1,...,T_m$ would correspond to a set of p-values $p_{val_1},...,p_{val_m}$. In that case the $\{p_{val_j}\}_j$ could be interpreted as a representation of the uncertainty on the p-value, but this is not the case in our situation. One aims to compute the p-value that corresponds to the true sample $x_1,...,x_{16}$. As mentioned before, multiple measurement of each x_i substantially reduced the uncertainty on each x_i. Thus, before computing the test statistic, the uncertainty distribution $P_{X_i^*|y_{i,1},...,y_{i,m}}$ can be computed, which will be a t-distribution centered around $\bar{Y}_i^m$. The width of this distribution will be significantly smaller than σ_{obs}. The remaining uncertainty should then be propagated according to Eq. 19.

1.4.3 χ^2-test

We have shown above that what the referee interprets as the uncertainty distribution is in fact not the uncertainty distribution, but the distribution of the measured value Y_i around a true value x_i. Furthermore, we explained why in case of repeated measurements one should not compute p-values for each measured sample $y_{1,j},...y_{16,j}$, but should instead use the multiple measurements to infer uncertainty distributions $\mathcal{P}_{X_i^*|y_{i,1},...,y_{i,m}}$ for the estimator of true values x_i^*. This uncertainty is small compared to σ_{obs} and should then be propagated in the sense of Eq. 19. We would thus politely like to argue that the referee compares mathematical objects that cannot be compared. If the 6000 samples from $\mathcal{P}_{Y_i|x_i}$ really were to be interpreted as multiple measurements of the true value, then proper propagation of the uncertainty to $\mathcal{P}_{X_i^*|y_{i,1},...,y_{i,m}}$ would indeed yield rejection rates of 5% for the criteria (2) and (3) in the setup designed by the referee.

For the case of a single measurement we will show how under the correct interpretation of the measurements y_i criterion (1) in addition does not reject a true null-hypothesis at a 5% rate, as stated by the referee. We will discuss a chi^2-test, because in the particular case of the t-test the 5% rejection holds true due to the symmetry of the problem.

Consider a normally distributed random variable X with $\mathcal{P}_X = \mathcal{N}(\mu_X = 0, \sigma_X)$ and a sample $\mathbf{x} = (x_1,...,x_n)$ of n realizations. Furthermore, assume that the measurement errors are distribution normally as above

$$\mathcal{P}_\Delta = \mathcal{N}(0, \sigma_{\text{obs}}), \tag{11}$$

such that the measured value $Y_i = x_i + \Delta_i$ corresponding to the true value x_i is distributed according to $\mathcal{P}_{Y_i|x_i} = \mathcal{N}(x_i, \sigma_{\text{obs}})$. We distinguish the following two situations:

1. single measurement of the true sample $\mathbf{x} = (x_1, ..., x_n)$ yields a single realization of the measured sample $\mathbf{y} = (y_1, ..., y_n)$
2. repeated measurement of the true sample yields for each true value x_i a set of measured values $\mathbf{y_i} = (y_{i,1}, ..., y_{i,m})$.

We will first discuss the case of a single measurement, because this is the one in analogy to the manuscript. Without any knowledge about the true values x_i the unconditional distribution for the random variables Y_i is given by the convolution of $\mathcal{P}_X$ and $\mathcal{P}_\Delta$

$$\mathcal{P}_Y = \mathcal{N}(0, \sigma_X) \circledast \mathcal{N}(0, \sigma_{\text{obs}}) = \mathcal{N}(0, \sqrt{\sigma_X^2 + \sigma_{\text{obs}}^2}). \tag{12}$$

Hence, measuring a single sample of true values is effectively the same as sampling from a normal distribution with standard deviation $\sigma_Y = \sqrt{\sigma_X^2 + \sigma_{\text{obs}}^2}$ and mean $\mu_Y = 0$. Suppose you were given one measured sample $\mathbf{y} = (y_1, ..., y_n)$ As explained above, knowing σ_{obs} allows you to quantify an uncertainty distribution for the true value from this measurement:

$$\mathcal{P}_{\mathbf{X}^*|\mathbf{y}} = \mathcal{N}(\mathbf{y}, \sigma_{\text{obs}}) \quad \text{or componentwise} \quad \mathcal{P}_{X_i^*|y_i} = \mathcal{N}(y_i, \sigma_{\text{obs}}) \tag{13}$$

This means that, given $\mathbf{y}$, the probability that the true value equals $\mathbf{x}^*$ is normally distributed around the measured $\mathbf{y}$, with standard deviation σ_{obs}.

Now suppose you wanted to test the null-hypothesis $H_0 : \sigma_X < \sigma_0$, knowing that X is normally distributed. Under this condition, the test statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{14}$$

follows a χ^2 distribution of $n-1$ degrees of freedom. Here, $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - u)^2$ denotes the samples' variance with sample mean $u = \frac{1}{n}\sum_{i=1}^n x_i$. σ_0 is the variance that defines the null-hypothesis. In a one-sided significance test $\alpha = 0.05$ the most extreme 5% at the high end of the possible values for χ^2 comprise the rejection region Ω_K. Referee 1 now proposed to first collapse the uncertainty distribution $\mathcal{P}_{\mathbf{X}^*|\mathbf{y}}$ to its expected value

$$\text{E}(\mathbf{X}^*) = \mathbf{y} \tag{15}$$

and subsequently apply the hypothesis test. It becomes clear that in this setup one effectively tests the standard deviation $\sigma_Y = \sqrt{\sigma_X^2 + \sigma_{\text{obs}}^2}$ instead of the desired σ_X^2. For $\sigma_{\text{obs}}^2 \sim \sigma_X^2$ one finds $\sigma_Y \sim \sqrt{2}\sigma_X$. If then $\sigma_X \lesssim \sigma_0$ the probability to reject the hypothesis easily rises above the chosen significance level under the use of criterion (1), since the effective standard deviation exceeds the hypothesized standard deviation by a factor of $\sqrt{2}$ even though the null-hypothesis is in fact true. Importantly, we see here that large observational

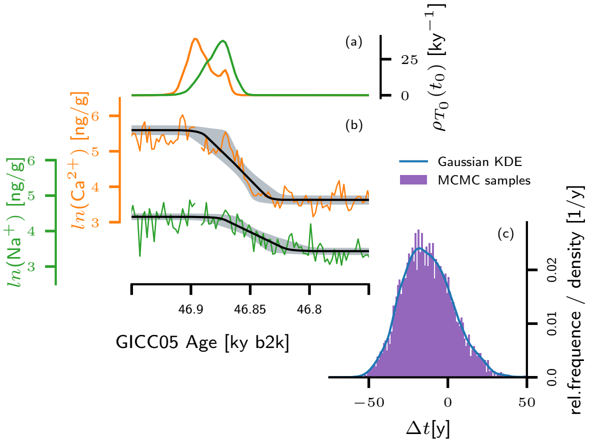

Figure 2: Figure 3 from the manuscript

uncertainty limits the ability to test hypothesis. In the setup discussed by referee 1, the results obtained by the application of criterion (1) were robust against the measurement uncertainty which contradicts our physical intuition. Please note that all criteria converge to the classical hypothesis testing case when Δ or more generally speaking, the uncertainty tends to zero.

In the case where repeated measurements of the true value x_i are possible one can estimate the true value x_i with higher precision by averaging over the observed values

$$\bar{Y}^m = (Y_{i,1}, ..., Y_{i,m}). \tag{16}$$

For a given x_i the variable $\bar{Y}^m$ will be distributed around x_i with variance $\frac{\sigma^2_{\text{obs}}}{m}$. Hence, the generic random variable $\bar{Y}^m$ is normally distributed around zero with an effective standard deviation of $\sigma_{\bar{Y}^m} = \sqrt{\sigma_X^2 + \frac{\sigma^2_{\text{obs}}}{m}}$. The mechanism is the same as above and again, the chances to wrongly reject the null-hypothesis may be higher than 5% if one would follow criterion (1), because the effective variance exceeds the true variance. The impact of the measurement uncertainty is reduced by the repeated measurement.

1.5 Point-by-Point answer to the referee

> Additionally, prompted by the authors statement about the novelty of their "uncertainty propa- gation to p-values" in their response: A cursory literature search (keywords fuzzy p-values, bayesian p-values) brought up a range of papers that seem to be dealing with p-values in settings similar to the setting the authors deal with here. It would be good if the authors set their approach into the context of aforementioned literature or highlight

its conceptual differences should the they decide to further employ it in the next iteration of the study.

During our work on the manuscript, we encountered the concept of fuzzy p-values as well. However, we found that it did not match our case precisely. Filtzmoser (2004) writes with regards to fuzzy data:

> Real observations of continuous quantities are not precise numbers but more or less non-precise. The best description of such data is by so-called non-precise numbers. Such observations are also called fuzzy. The fuzziness is different from measurement errors and stochastic uncertainty. It is a feature of single observations from continuous quantities. Errors are described by statistical models and should not be confused with fuzziness. In general fuzziness and errors are superimposed.

Instead of PDF's, in the fuzzy p-value theory the uncertainty about data is expressed in terms of characteristic functions. Also an adoption of this concept to our case fails due to the properties that characteristic functions are required to fulfill (see for example Filtzmoser, 2004 and Parchami, 2008).

However, we agree with the referee that fuzzy p-values are worth mentioning in the context of our work and added the sentence 'The theory of fuzzy p-values is in fact concerned with uncertainties either in the data or in the hypothesis, however, it is not applicable to measurement uncertainties that are quantifiable in terms of probability density functions' in line 569.

Regarding the Bayesian p-value, Gelman et al. (2004) write:

> *Posterior predictive p-values.* To evaluate the fit of the posterior distribution of a Bayesian model, we can compare the observed data to the posterior predictive distribution. In the Bayesian approach, test quantities can be functions of the unknown parameters as well as data because the test quantity is evaluated over draws from the posterior distribution of the unknown parameters. The Bayesian p-value is defined as the probability that the replicated data could be more extreme than the observed data, as measured by the test quantity:
>
> $$p_B = Pr(T(y^{rep}, \theta) \geq T(y, \theta)|y), \tag{17}$$
>
> where the probability is taken over the posterior distribution of θ and the posterior predictive distribution of y^{rep} (that is, the joint distribution, $p(\theta, y^{rep}|y)$:
>
> $$p_B = \int\int I_{T(y^{rep},\theta)\geq T(y,\theta)} p(y^{rep}|\theta)p(\theta|y)dyrepd\theta, \tag{18}$$

where I is the indicator function. In this formula, we have used the property of the predictive distribution that $p(y^{rep}|\theta, y) = p(y^{rep}|\theta)$.

Hence, the Bayesian p-value assesses the fit of a model to data. In our application, that would be the fit of the linear-ramp model to the transition data. But this concept does not seem applicable for assessing the significance of the uncertain sample of transition onset lags with respect to the null-hypothesis.

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

In the caption of their Fig.3 Erhardt et al write 'No timing results are given for transitions where there are data gaps in one of the necessary datasets.' However, it is true that DO events with minor gaps in the data around the transition are used for the analysis nonetheless.

We corrected the statement, which now reads: 'For their analysis, Erhardt et al. (2019) only considered time series around DO events that do not suffer from **substantial** data gaps.'

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

The referee is certainly correct that mathematically, there is no difference between the distributions that we term 'uncertainty distributions' and standard probability distributions represented by probability density functions (PDFs). However, their interpretation is somewhat different: a PDF that characterizes a random experiment can be thought of as the probability to observe a certain value of the random variable in a repeated next execution of the experiment. Contrarily, an uncertainty distribution is a measure of plausibility but the uncertain variable cannot be observed repeatedly. We believe that the term 'uncertainty distribution' is a useful way to highlight this difference and in fact helps the reader not to confuse the different origins of randomness involved in the analysis. We would therefore like to keep using the term 'uncertainty distribution' and have not changed this in the revised manuscript.

For sake of clarity, and to avoid ambiguous notation, we have marked all uncertaint quantities with a hat - that is, all random variables that inherit their randomness from the Bayesian transition onset detection. The same holds true for potential values they might assume. We added explanation on how true values and uncertain values are related and how the two-level randomness must be understood in line 267.

'The left panel in Fig. 4 illustrates this situation: from an underlying population $\mathcal{P}_X$ a sample $\mathbf{x} = (x_1, ..., x_6)$ is realized, with the x_i denoting the true values of the individual realizations. However, the exact value of x_i can not be measured due to measurement uncertainties. Instead an estimator $\hat{Y}_i$ is introduced together with the uncertainty distribution $\rho_{\hat{Y}_i}(\hat{y}_i)$ that expresses the observers belief about how likely a specific value $\hat{y}_i$ for the estimator $\hat{Y}_i$ is to agree with the true value x_i. The $\hat{Y}_i$ correspond to the $\Delta\hat{T}_i^{p,q}$. For the x_i there is no direct correspondence in the problem at hand, because this quantity in practice cannot be excessed and hence must not be denoted explicitly in the practical case. We call the vector of estimators $\hat{\mathbf{Y}} = (\hat{Y}_1, ..., \hat{Y}_n)$ an uncertain sample in the following.'

Also, in the cause of this, the notation especially in Section 3 changed in many places.

However, we agree that 'high uncertainty probability' is not a well comprehensible term and have thus replaced 'that is, we observe high uncertainty probabilities for negative ΔT_i across the sample according to Eq. 7' with 'that is, the corresponding uncertainty distribution indicate high probabilities for negative ΔT_i across the sample according to Eq. 7.' (l.326)

L 208ff Prescribing a "fixed pattern of causes and effects" is an overly strong interpretation. It is quite easy to imagine a range of mechanisms that have an indistinguishable imprint in the proxy record but trigger a transition from stadial to interstadial conditions. For example the ocean processes alone, that the authors list in the introduction are probably very difficult to distinguish using the proxies presented here as they partly invoke very similar feedback mechanisms. I would suggest to reformulate "similar pattern of cause and effects" to convey the possible ambiguity as a discussion of the imprint in the proxy records by the different causes clearly goes beyond the focus of this manuscript. The same holds true for later implications of the one trigger of DO-Events that the authors make throughout the manuscript.

The referee is of course right that different mechanisms could easily have left the same or at least an indistinguishable imprint in the proxy data. However, we express very clearly that the assumed one to one mapping

- δ^{18}O → temperature

- Ca^{2+} → state of the atmosphere
- Na^{+} → sea ice
- λ → local precipitation

is potentially oversimplified and that we leave the discussion about the correct proxy interpretation to the experts. In order to think of the Δt_i from different DO events as an i.i.d. random variable, one *has to* assume that all DO events were triggered by the same physical process, which in turn necessarily 'prescribes a fixed pattern of causes and effects for all DO events' - at least on the scale of interaction between the climatic subsystems represented by the proxies. In turn, if the pattern of causes and effects was different between DO events, then the physical mechanism was not the same and we could not treat Δt as an i.i.d. random variable.

For clarity, we added: 'at least on the scale of interaction between climatic subsystems represented by the proxies under study' to the corresponding sentence in line 213.

L 210 The imperative "should" in regards to the setup of the frequentist analysis that follows should be replaced with "could" or "can".

Thank you, we fully agree and replaced 'should' with 'can'.

L 220 Either "be" or "bear".

Corrected: 'bear'.

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

The footnote is very technical and addresses the mathematically interested reader. Placing the statement in a footnote clearly signals that it this content is not required to follow the reasoning in the main text. If the editor thinks that using footnotes in this way is not adequate, we will be glad to incorporate this footnote into the main text.

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

The formulation 'the sample no longer only carries the randomness of the population' suggests that now the sample carries the randomness of the population and the randomness due measurement uncertainty simultaneously. However, the sample of DO transition onset lags has already been realized and therefore does not carry the randomness of the population any more. It only carries the randomness due to the measurement.

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

With regards to this, Romano and Lehmann (2006) write:

> We now begin the study of the statistical problem that forms the principal subject of this book, the problem of hypothesis testing. As the term suggests, one wishes to decide whether or not some hypothesis that has been formulated is correct. The choice here lies between only two decisions: accepting or rejecting the hypothesis. A decision procedure for such a problem is called a test of the hypothesis in question.

The term 'null-hypothesis' is used to expressed that this hypothesis entails the 'no effect', 'no difference' or 'no causal relation' that one aims to reject. If the test fails to reject the null-hypothesis it must certainly be acceptance, for the time being. However, in the revised manuscript we highlight more strongly than before that in our case we cannot reject the null-hypothesis due to the large uncertainties and that this result should not be interpreted as evidence against the alternative (see l.562ff).

At the beginning of the discussion section some text parts have been moved around and some parts have been shortened. Now the section reads more consise while the content has not changed.

In the previous version we used the term 'biased' to characterize a population with mean different from zero. Since a 'bias' in statistics usually means a systematic distortion of measurements, we replaced the terms 'biased' and 'unbiased' with the terms 'non-neutral' and 'neutral', respectively.

References

Peter Filzmoser and R. Viertl: Testing hypotheses with fuzzy data: The fuzzy p-value. Metrika: International Journal for Theoretical and Applied Statistics, 2004, vol. 59, issue 1.

Abbas Parchami, Mahmoud Taheri and Mashaallah Mashinchi: Fuzzyp-value in testing fuzzy hypotheseswith crisp data. Stat Papers (2010) 51:209–226.

Lehmann, E. L. & Romano, J. P. Testing Statistical Hypothesis. Design vol. 102 (Springer US, 2006).

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

2 Answer to Referee 2

2.1 General Remarks

First of all we would like to thank the referee for the careful second review. Before we address the comments point by point, we will discuss the main critizism, which is why we do not use a mixture distribution to statistically assess the significance of the sample of uncertain DO time lags. The simple answer is: we do, but we failed to make this clear earlier. We therefore apologize for our reservation towards this comment, which was brought up by the referee already in the first review and which we had misunderstood. Accordingly the paragraph from line 572 onwards (previous manuscript) was removed in the revised version.

The referee proposes to use a mixture distribution to test whether the population mean is greater than or equal to zero. In fact, the bootstrap test that we carry out does exactly this, though it is not immediately obvious. Given a sample $\mathbf{x} = (x_1, ..., x_n)$ generated from a population $\mathcal{P}_X$, the idea behind bootstrapping is that the empirical distribution (or mixture distribution as termed by the referee) $\bar{\rho}_X(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - x_i)$ approximates the population to a certain degree. Let $H_0 : \mu_X \geq 0$ denote the null-hypothesis that the mean μ_X of the population P_X is greater than or equal to zero. In the absence of any further information on the populations shape, a null-distribution for testing the hypothesis can be constructed by modifying the mixture distribution such that it complies with the null-hypothesis.

In order to test the mean of a population, Lehmann and Romano (2006) propose the use of the test statistic

$$T^n = n^{1/2}\, u, \tag{19}$$

with $u = \frac{1}{n}\sum_{i=1}^{n} x_i$ denoting the sample mean. To construct a null-distribution for the test statistic, they take the mixture distribution shifted by the sample mean

$$\rho^n_{X'}(x') = \frac{1}{n}\sum_{i=1}^{n}\delta(x' - x_i + u), \tag{20}$$

such that $\rho^n_{X'}(x')$ has mean 0. Hence, it fulfills the null-hypothesis while simultaneously capturing the characteristics of the mixture distribution. Furthermore, this choice guarantees that 1) $\rho^n_{X'} \overset{n\to\infty}{\to} \rho_{X'}$ such that $\rho_{X'}$ has mean 0 as well (criteria for convergence are given in Lehmann and Romano (2006)), and 2) that in this limit the probability for rejection is not higher than α (the chosen significance level) for any original population P_X that fulfills H_0. For testing the inequality considered in H_0, the case where the null-distribution has zero mean is the decisive one. From the shifted mixture distribution $\rho^n_{X'}$ a data-driven null-distribution can be computed by resampling m samples of size n $\mathbf{X}^*_j = (x^*_1, ..., x^*_n)_j$ from $\rho^n_{X'}$ and computing the test statistic $T^n_j = T^n(X^*_j)$ for each of these 'synthetic samples'. For a given significance level α the $1-\alpha$-th percentile of the set $\{T^n_1, ..., T^n_m\}$ establishes a rejection region for the original $T^n(\mathbf{x})$.

In the previous version of our manuscript we used the sample mean itself as a test statistic. This has been corrected and we now use the statistic T^n as proposed by Lehmann and Romano (2006). Furthermore, in our study the original sample $\mathbf{x}$ as well as its sample mean $u(\mathbf{x})$ and its corresponding test statistic are uncertain. This uncertainty is propagated rigorously through all the steps described above.

When the referee proposed to use the mixture distribution to 'test whether the population mean is greater or equal zero', he or she might have had in mind to simply resample samples of size n $\mathbf{X}_j^* = (x_1^*, ..., x_n^*)_j$ from a non-shifted mixture distribution. Then one could compute their means $u_j^* = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}^*$ and compare the α-th percentile to the mean of the null hypothesis $\mu_0 \geq 0$.

Lehmann and Romano (2006) and Hall and Wilson (1991) provide arguments why for rigorous hypothesis testing the shifting of the mixture distribution is required. While taking this into account, effectively we indeed use the mixture distribution to assess the significance of the given sample with respect to the null-hypothesis, just as proposed by the referee.

2.2 Point by Point answer to the referee

L380 In response the referee's remark on using the mixture distribution, we rewrote section 3.3.3 and changed the test statistic. Accroding to these changes, Figures 6. and 7. as well as Table 1 and Table B1 were updated.

L86ff "In order to review the statistical evidence for a potential systematic lags, we formalize the notion of a 'systematic lag': We call a lag systematic if it is enshrined in the random experiment in form of a population mean different from zero. Samples generated from such a biased population would systematically (and not by chance) exhibit sample means different from zero. Accordingly, we formulate the null hypothesis of a pairwise unbiased transition sequence, that is, a population mean equal to zero." Could this maybe be reformulated, in order to also acknowledge the fact that whether or not a truly biased sample can be expected to systematically exhibit sample means (significantly) different from zero depends on the sample size? To me, this seems to be an important motivation for the procedure proposed here.

We agree that the sample size is one of the factors that limits the ability to detect a systematic lag between the proxy variables. The probability to reject the null-hypothesis assuming a truly biased population is a non-trivial quantity and depends on the sample size, the strength of the bias, and the variance of the population simultaneously. Since these factors interact in a non-trivial way, we do not believe this point should be raised already at the stage were we still aim to properly define the statistical problem. Instead we have included a paragraph in the discussion that discusses the chance to reject the null-hypothesis.

Caption Fig. 2 I was not sure at first whether the posteriors shown are from the authors or from Erhardt et al. Maybe this could be made specifically clear in the caption.

We reproduced all posterior probability density estimates adopting the method and the data provided by Erhardt et al. We agree that this was not made sufficiently clear in the caption. Hence, we changed the sentence

'The probability density estimates for the transition onsets with respect to the timing of the DO event according to Rasmussen et al. (2014) are shown in arbitrary units for all proxies' to 'The posterior probability densities for the transition onsets with respect to the timing of the DO event according to Rasmussen et al. (2014) are shown in arbitrary units for all proxies. They were recalculated using the data and the method provided by Erhardt et al. (2019).'

L187 indipendent → independent

Corrected.

L797-798 I don't understand what is meant here, maybe just a grammatical error?

We are not quite sure we understand what the referee finds problematic here. We changed the sentence 'For a given proxy pair the starting point for the statistical analysis however, is the uncertain sample $\boldsymbol{\Delta}\mathbf{T}^{p,q} = (\Delta T_1^{p,q}, ..., \Delta T_n^{p,q})$ characterized by the n dimensional uncertainty distribution $\rho^{p,q}_{\boldsymbol{\Delta}\mathbf{T}}(\boldsymbol{\Delta}\mathbf{t}^{p,q}) = \prod \rho_{\Delta T_i^{p,q}}(\Delta t_i^{p,q})$.' to 'For a given proxy pair the starting point for the statistical analysis is, however, the uncertain sample $\boldsymbol{\Delta}\mathbf{T}^{p,q} = (\Delta T_1^{p,q}, ..., \Delta T_n^{p,q})$, **which is** characterized by the n-dimensional uncertainty distribution $\rho^{p,q}_{\boldsymbol{\Delta}\mathbf{T}}(\boldsymbol{\Delta}\mathbf{t}^{p,q}) = \prod \rho_{\Delta T_i^{p,q}}(\Delta t_i^{p,q})$.' and hope that this clarifies the statement.

At the beginning of the discussion section some text parts have been moved around and some parts have been shortened. Now the section reads more consise while the content has not changed.

References

Lehmann, E. L. & Romano, J. P. Testing Statistical Hypothesis. Design vol. 102 (Springer US, 2006).

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