We thank both Reviewers for their supportive reviews and the constructive comments. As R2 is happy with our previous revision, here we will only address R1's comments.
##

Re-Review of Muschitiello & Aquino-Lopez CPD

I'd like to thank the authors for their replies to my comments, which cleared up some of my questions but partly also reinforce some of my concerns. The additional analyses lend some support to the robustness of the results, yet, in the current version of the manuscript it is not clear how the synchronization is done exactly and whether it may be prone to biases.

Thank you for your detailed feedback. We appreciate the reviewers time and effort in improving the current manuscript. We appreciate the opportunity to clarify our methodology and the rationale behind the standardization of the data.

A part of my confusion is stemming from the response of the authors to my previous comment on the underlying assumption of a linear relationship between speleothem and ice core data. I remarked, that equation 3 requires a linear relationship between the proxies. The authors argue in their response that this is overcome by standardizing the data to [-1,1] and assigning large uncertainties to the signal. However, the applied changes in the manuscript (L217-221) only outline that a non-linear adjustment of the timescales is facilitated by the method. This is of course obvious (and could be removed from the manuscript) but does not address my original question.

Apologies. There was some confusion around the reviewer's original question. We now understand the concern regarding the implication of assuming a linear relationship between the proxies as interpreted by Equation 3. In response, we would like to emphasize that the standardization of proxy data to a range of [-1, 1] and the assignment of large uncertainties to the signal are methodological choices aimed at mitigating the impact of assuming a strict linear relationship. The standardization process is not merely a procedural step but a crucial approach to minimize parameter estimation errors. On the other hand, the synchronization is performed in the time window of the target core ($t \in (t_0', t_m')$), allowing the method to put both records on a similar window, in both axis, in order to identify the best alignment that maximizes their similarity, as stated by the equation $u(t_i) \approx g(\tau(t_i)) \forall t_i \in \vec{t}$ (new lines 245). We would like to reinforce the point that this function seeks to identify similarities and is not to be taken as an equality of the point estimates, as stated by the use of the approximation symbol (≈) and not the equality symbol (=).

It is important to note that this approach fundamentally differs from subjective point estimation of climate-wiggle matching. Our methodology relies on the uncertainty quantification of the whole record, and the equation u(t') ≈ g(τ(t')), together with the use of the t-distribution and conservative standard deviation, provides a conservative estimate for quantifying the uncertainty around the alignment between the records, resulting in a "best" guess for the alignment between both records, rather than subjective matching.

From the author's reply, I understand (but I am not sure because the manuscript and the reply are incoherent in this respect) that the standardization of the data is done separately for each 180-year segment. If so, this is not clearly stated in the manuscript (compare L253-257). Further, this raises additional questions.

1. A standardization for each segment would allow drastic changes in the relationship (regression slope) between ice core and speleothem data. It could lead to the large peak in the speleothem d18O data around 18-20 kaBP, which has the magnitude of a DO-transition, to be matched to some minor structure in the ice core record, which has no equivalent change there. What would be the reason for such a drastic change in the relationship? How valid is the assumption that this still reflects the same physical driver in both proxies? Allowing for a freely varying relationship between the proxies increases the likelihood of erroneously aligning signals that have no physical connection. I also

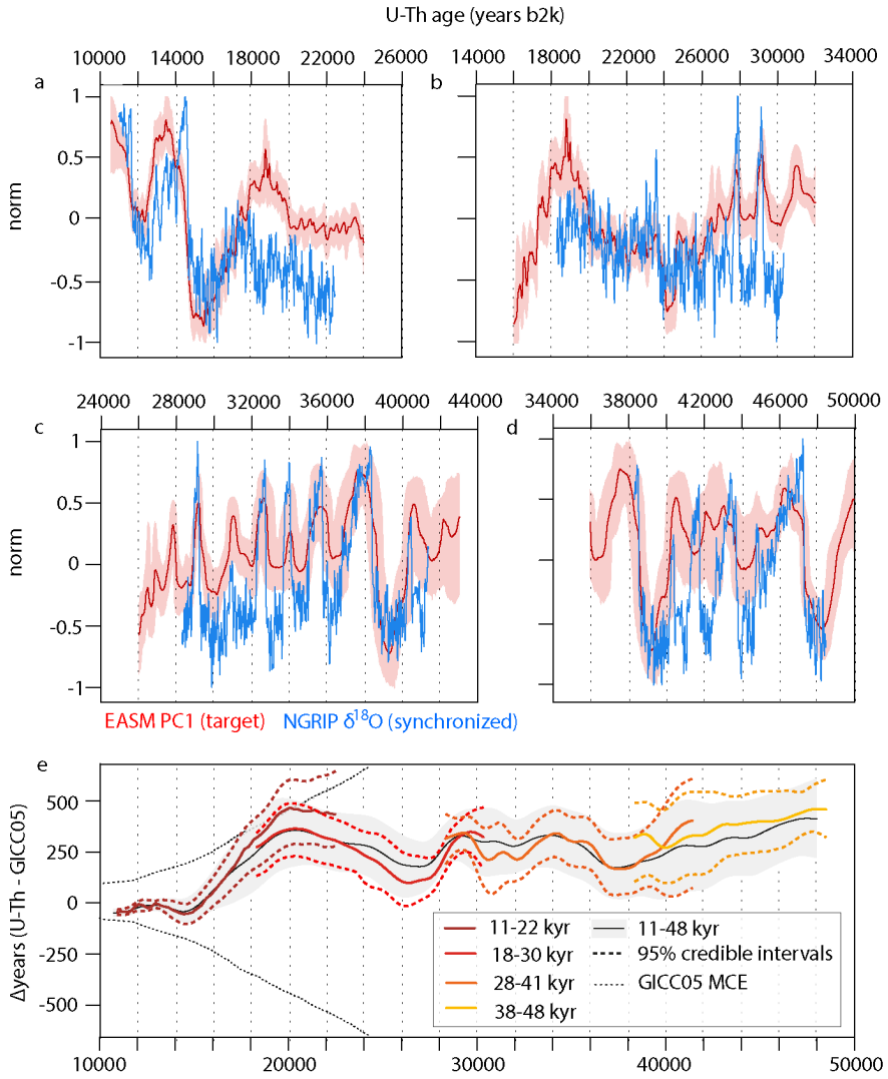
disagree, that the standardization is better at handling non-linearity in the relationship between the proxies, since the full dynamic range of the proxies occurs at DO-onsets within decades which fits inside one standardization-segment and is thus, still treated as a linear relationship. Further, a standardization of each 180-year segment effectively corresponds to a 180-year high-pass filter, while the speleothem data has very little variability in this frequency band, especially during MIS-3.

2. A standardization of the record as a whole (as shown in the figures 4-6) on the other hand, leads to long periods of systematic differences between the records (MIS-2 but also during stadials of MIS-3). Because the method evaluates squared differences between the records (equation 3) this leaves the method prone to minimizing those differences instead of matching structures. To test this, I ran a test on artificial data (figure below), which are composed of a AR(1)-process (the exact same in both datasets) and different linear trends in both datasets, as seen in the real data. Standardizing both datasets (as a whole) to [1,1] and allowing for a linear timescale compression/stretching of one dataset by +/- 5% clearly leads to an erroneous compression of the timescale for the investigated segment. Increasing the uncertainty of the records does not fix this problem in contrast to the statement in the manuscript (L253-257). I understand that this would obviously be different when all segments of the data are jointly evaluated in the MCMC, but it is not clear whether this really avoids the problem as a whole. This effect may for example bias the inference during MIS-3 since the scaling leads to large difference between the records during stadials and small differences during interstadials (see figure 4) and I wonder whether this can explain part of the difference to the results by Corrick et al. (see also comment below).

   Thank you for raising these additional points regarding the potential issues with standardizing the entire record and the impact of systematic differences between the records during certain periods. We now clearly state in the main text that the scaling is performed on the entire timeseries and not by segment (new lines 247 and 256). In the following we will therefore address point 2 from above.

   Regarding the comment about standardizing the entire record potentially leading to long periods of systematic differences between the records, and the method minimizing these differences instead of matching structures, we acknowledge this as a valid concern. However, it is important to note that our methodology quantifies the uncertainty associated with the global alignment itself, evaluating the alignment quality across the entire record. This means that if there is a mismatch between the records in a particular period, it would also affect the sections where the records are properly aligned. While the standardization of the entire record may introduce biases during certain periods, our approach relies on the joint evaluation of all segments in the MCMC process. This global evaluation aims to find the optimal alignment that minimizes the overall differences while accounting for the uncertainties across the entire record. In other words, our methodology provides a conservative estimate of the uncertainty around the alignment by jointly considering all segments and their respective uncertainties. In addition, the use of the t-distribution and conservative standard deviation further contributes to this conservative estimate. We now more clearly discuss all these issues in the main text (new lines 256-261 and lines 345-353).

   We appreciate the reviewer's point regarding the potential issues that may arise in different windows of observation. To test the robustness of our methods, we performed sensitivity tests where we align NGRIP $\delta^{18}O$ against the speleothem stack using short segments of ~10 kyr. We cropped both the input and target data and standardized the timeseries in the same way as per the global alignment. The sensitivity tests show that both alignments agree with each other, and more importantly, the differences between the global and local alignments are statistically indistinguishable (see figure below). The overlapping credible intervals across most of the record demonstrate the robustness of our approach, even when considering localized alignments. These new findings are presented and discussed in the new version of the manuscript (new Supplementary Fig. 1 and new lines 345-353).



Supplementary Figure 1. Comparison between the global and localized synchronizations of NGRIP $\delta^{18}O$ and EASM PC1. The records were cropped in segments of approximately 10 kyr and scaled between -1 and 1 before alignment. Intervals spanning 11-22 kyr b2k (**a**), 18-30 kyr b2k (**b**), 28-41 kyr b2k (**c**), and 38-48 kyr b2k (**d**). The target is always longer than the input by allowing 2 kyr on both sides of the timeseries. **e.** Posterior median and pointwise 95% credible intervals (of the difference ΔT between the GICC05 and U-Th timescales estimated locally (coloured lines) and globally (grey shading and black line).

As to the reviewer's concern about the potential bias during MIS-3, where the scaling may lead to large differences during stadials and small differences during interstadials, we now show that our results are in better agreement with Buizert et al's Δt estimates than previously thought after that the Hulu data is placed on the updated U-Th timescale (new Figs. 4-6, new line 359, and reply below).

This leads to several request for clarification/revision in the manuscript:

1. Please clearly indicate whether the standardization is done for each segment or for the record as a whole.

a. If done for each segment: Please include a supplementary figure where you show the records after standardization of each segment. As it is now, the upper two panels of figures 4-6 are misleading. Furthermore, please discuss (incl. figure in SI) how variable the ratio of the scaling factors is over time (i.e., how variable is the assumed relationship between the proxies) and whether it can still be assumed that this reflects a common climatic process in both proxies.

The standardization is performed for the entire record, not for each individual segment. We have clarified this in the revised text (new lines 245 and 256) and crafted a new Supplementary Figure to

compare the quality of the localized and global synchronizations. We agree that this process may struggle in records with extreme maxima/minima. We have added a discussion of this potential limitation in the main text (new lines 256-261 and 345-354).

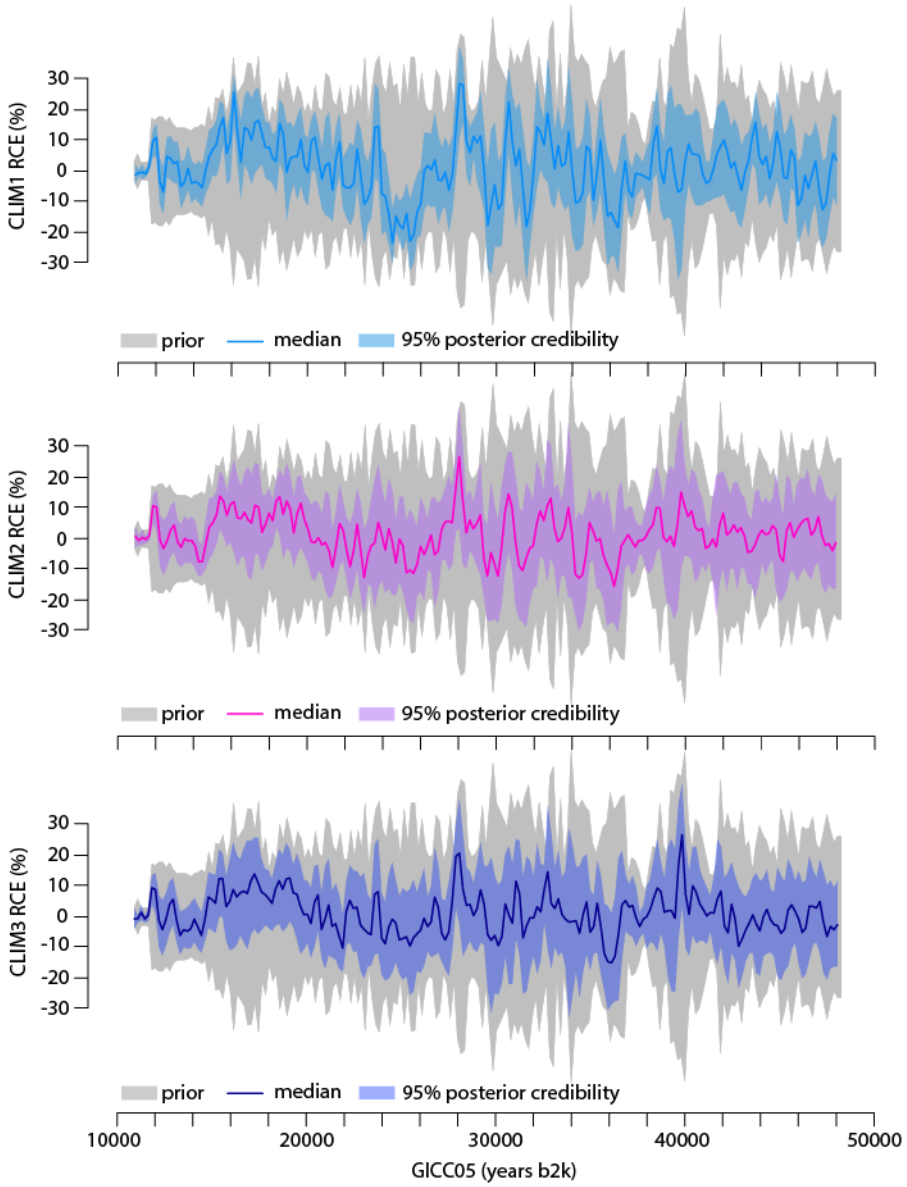
b. If done for the record as a whole, please show that the issue outlined under point 2 above is not affecting the synchronization (for example by analysing subsections of the data and standardizing those).

We have address this in the points above and have also analyzed subsections of the data (see Supplementary Figure 1).

2. In both cases, I repeat my request for additional panels in figures 4-6 that show the correlation per segment before and after synchronization, to allow the reader to evaluate which signals are really driving the synchronization, and by how much the fit between the records is improved by synchronization. It is for example surprising that the synchronizations for the different datasets (NGRIP d18O, GIRP d18O, NGRIP Ca) is so similar in MIS-2 when these records have been shown to diverge during this period (see figure 2 but also Rasmussen et al. 2008 fig. 3, 10.1016/j.quascirev.2007.01.016). I am aware that it is not the correlation of records that is being evaluated, but it is an intuitive measure for the readers, and ultimately, a correlation of the signals is the fundamental reason why climate-wiggle matching is considered a valid tool in paleoclimatology. Further, this would illustrate how "continuous" the synchronization really is and where the transfer function is driven by the priors. If the synchronization is hinging on relatively few tie-points, then the title and main text need to be adjusted accordingly.

We disagree that the three synchronizations are similar during MIS2. Perhaps the trends/trajectories are comparable but the resulting offsets are quantitatively different and involve uncertainties of at least 0.1kyr in either direction. We now show the extent of the credible intervals associated with each synchronization (panels d of new Figures 4-6). These intervals will aid the reader identifying regions where the alignment is less robust. From these plots, it is evident that there are no sudden jumps in the synchronization uncertainty, nor a systematic narrowing of the uncertainty at the onset of stadials and interstadials, thus demonstrating that the alignment is not driven by discrete "tie points". Instead, the synchronizations yield relatively smooth uncertainty bounds throughout, which supports our argument of a continuous synchronization. This approach –which aligns with the Bayesian methodology– should satisfy the reviewers interest and provides the same intuitive measure for the reader.

Regarding the role of the priors, the figure below shows the prior versus posterior distributions for the RCE parameters. The interval widths demonstrate that the posteriors differ substantially from the priors, thus confirming that the data is driving the synchronization estimates, not just the priors, as the reviewer suggested.

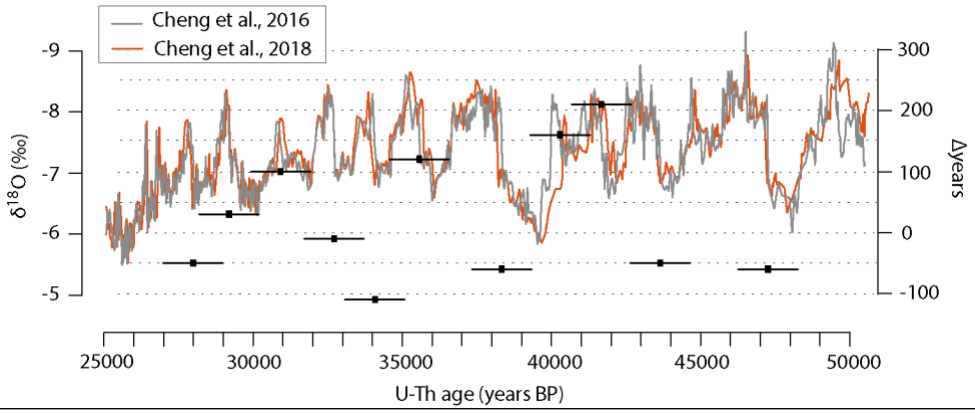


Prior and posterior RCE parameters for each CLIM synchronization.

Another aspect that may need to be revised is the inclusion of the results by Corrick et al. Comparing the way how the authors present the results by Corrick et al. in figure 7 to the equivalent plot in the original publication (Corrick et al. figure S6) it appears that the authors exaggerate the uncertainties by Corrick et al. It is my impression, that they included the GICC05 uncertainty into this figure, which is irrelevant for this comparison. If this is done correctly, it becomes apparent that the results presented here, significantly disagree with those by Corrick et al. between 30-38 kaBP. This should be discussed as this is also the period where there seems to be a systematic disagreement to the match by Buizert et al. 2015, which the authors attribute to Buizert et al's use of Hulu-cave only. However, this argument would not hold for the results by Corrick et al. Further, as the authors state in L365ff, their synchronization of this period is largely driven by DO-onsets, which is similar to the estimates by Corrick et al. This difference may stem from a bias mentioned above (small differences between the records during interstadials, large differences during stadials). Alternatively, this disagreement may arise from the method trying to find a compromise between aligning GS-GI and GI-GS transitions, while not violating the counting error constraints? Please discuss. Again, additional panels with running correlations of similar would help evaluating this.

Thank you for bringing this up. We took this opportunity to revise Buizert et al's match points to

account for the new U-Th chronology presented in Cheng et al., 2018 and to assess whether said match points still hold. To update Buizert et al's Δt estimates we adopted a simple sliding correlation approach using windows of 2000 years around the published tie points that are moved in steps of 10 years with a maximum lead/lag of 500 years (see figure below and Δt corrections on panels c of new Figures 4-6). The corrected Δt estimates are within the 95% credible intervals of our transfer functions and reveal a possible scaling of GICC05 that is slightly larger than previously estimated (0.99%).



Comparison between the speleothem $\delta^{18}O$ stack from Cheng et al. (2016) and the Hulu Cave $\delta^{18}O$ data from Cheng et al. (2018) on the timescale published therein. Black squares and horizontal bars indicate the age shift due to the new chronology at the interstadial transitions identified by Buizert et al. (2015). Positive values imply that the new timescale is older than the previous chronology. Age shifts were estimated using cross-correlation, whereby we correlated windows of 2000 years length with leads/lags of 500 years at steps of 10 years centered around said tie points.

As to comparing our results with the Δt estimates presented in Corrick et al. (2020), we adjusted the uncertainty bars as requested by the reviewer (see new Fig. 7). In general, we are happy to include additional text and discuss further. However, we deem such estimates to be –to some extent– subjective, and therefore inconsistent with our methodology. We strongly suggest that their age offsets should not be over-interpreted and in fact taken with a grain of salt. Admittedly, the authors state that their results differ from Buizert et al's Δt estimates as well as results from cosmogenic radionuclide wiggle matching, and that such difference is potentially associated with their "*methodological approach, including the choice of detrital-thorium correction and depth-age modeling*". More critically, the identification of interstadials is rather qualitative, and reliant on the visual identification of "*the first data point of the steep part that clearly deviates from the baseline level preceding the transition*" using at times low-resolution records that have "*at least three data points per thousand years*". In addition, the authors state that "*it was occasionally necessary to shift the point to a position structurally similar to that of the event's assigned position in NGRIP*". As such, they state that "*statistical methods to identify the onset of interstadial transitions were found to be difficult to implement consistently to all speleothem records*". The qualitative and subjective nature of Corrick et al's approach is at odds with the more objective method presented in our study and therefore differences in the resulting Δt estimates are not surprising. As a final remark, it should be noted that the authors analyzed the Hulu Cave data using the "pre-2018" U-Th chronology (i.e. Wang et al., 2001; Wu et al., 2009; Southon et al., 2012), which may have additionally contributed to the observed mismatch.

We now discuss more openly these issues in the main text (new lines 396-404).

Specific Comments:

L71: "14C concentrations" – change to D14C which is not a concentration L72:
"ocean carbon inventories" – change to "radiocarbon inventories"

Thank you. This is has been corrected (new lines 71-72).

L86: Please include Adolphi et al. 2018 into the reference list, since we the main point of our work was to test the synchroneity of DO-events in speleothems and ice cores.

This has been added (new line 85).

L110-114: These issues have nothing to do with the autocorrelation of cosmogenic radionuclides (d18O and Ca are autocorrelated as well) but are an artefact of analysing overlapping windows. Rephrase or delete.

This has been deleted and rephrased (new line 111).

L121: "when timescales reach their largest offset" – please add "according to cosmogenic radionuclides (Adolphi et al. 2018, Sinnl et al. 2023)"

This has been added (new line 121).

L123: "first continuous" – previous transfer functions where also continuous, albeit based on selected tie-points and various ways to interpolate in between. Given that this method is likely also only driven by a limited number of tie-points it is not that different. Please adjust the formulation and possibly the title.

While previous transfer functions were indeed continuous, they relied on (at times subjectively) selected tie points and various interpolation techniques between those points. In contrast, our approach is not driven by pre-selected tie points but rather a quantifiable assumption of continuity across the entire record. Furthermore we provide an objective uncertainty quantification of the overall alignment between the two records, rather than relying on interpolation between sparse tie-points. This objective uncertainty quantification is a key distinguishing feature of our method, as it allows evaluating the robustness of the alignment across the entire input. With this in mind we believe that no change is needed.

L129: "improve precision and accuracy" – how do you determine that your transfer is more accurate than previous versions? Please elaborate or delete.

This has been deleted (new line 130).

L135: "three independent synchronization" – synchronization should be plural. However, change to "three synchronizations based on independent Greenland ice core climate proxy records" or similar. The synchronizations are not independent (always the same speleothem data).

Thank you. This has been changed accordingly (new lines 136-137) and the term "independent" has been removed throughout the text.

L253-257: See major comments. Is the standardization done per segment? If yes, please clearly indicate.

Thank you. This has now been clarified in the main text. Please see our detailed reply above.

L265 (eq3): There is an error in this equation. The power of two only applies to the numerator of the last term of the equation (squared differences). old

The reviewer is correct we had made a small typo in the equation. This has now been corrected in the main text (new Equation 3):

$$l \propto \sum_{i=0}^{n} \left[-\log(\sigma_{u_i}) - \frac{a}{2} \log\left(b + \frac{(G(t_i) - u_i)^2}{2\sigma_{u_i}^2} \right) \right],$$

L269: “Any underestimation” – of what?

This has been rephrased (new lines 277-278).

L277: See my previous comments. Shouldn’t tau0 be constrained by the MCE instead of the RCE?

Thank you. This section has been edited accordingly (new lines 286-292).

L296: replace “uncertainties” with “credible intervals”

This has been changed (new line 306).

L346: Muscheler et al. 2008 inferred an offset of 65 not 55 years. Please change.

Thank you. This has been changed (new line 364).