(R3 comments in normal typeface; **responses in bold**)

1) Greenland and Hulu Cave data covariation.

The study is based on correlation of NGRIP Deuterium excess data with Hulu d18O. The reason is explained in line 157-160 and may be true for the response to the H events (as it is for D-O events), but it is a critical and very, very daring assumption that Hulu d18O and NGRIP D excess trace the same smaller-scale climatic changes. I do not think that this has been demonstrated previously, and the manuscript does not provide compelling evidence that the smaller-scale features correlate significantly. As I understand, Figure 1 does not separate the hosed D-O-scale/H-scale variability from smaller-scale variability.

Thank you for bringing this up. Fig. 1b & d do already show that climate/hydroclimate changes at NGRIP and Hulu Cave are linearly correlated and synchronous at timescales shorter than DO variability. I will present this finding more clearly in the main text.

I think the manuscript will either have to demonstrate with statistical back up that the Hulu d18O signal correlates significantly with at least one of the Greenland records (e.g. d18O, Calcium, or D excess) also over smaller-scale (and preferably non-forced) changes, or refrain from presenting a match of the records across these rather long periods which do not have D-O- or H-scale variability. That would challenge the concept of a "continuous" transfer function.

Notwithstanding the climate model results presented in Fig. 1b & d (see my previous comment), I agree that this issue should be addressed more rigorously. As also discussed in my replies to R2, I will address this concern in the revision. To demonstrate a physical relationship between hydroclimatic changes in the Asian Summer Monsoon region and Greenland Summit at timescale shorter than DO variability, I will present a novel analysis of climate model output from transient simulations of the last glacial period (Armstrong et al., 2019) and deglaciation (Liu et al., 2009), and from PMIP4-CMIP6 LGM experiments. All simulations reveal the presence of a persistent hydroclimatic covariability at multidecadal and centennial timescales between SE Asia and Greenland, which justifies the alignment approach used here –in particular for the interval 18-24ka.

2) Speleo dating uncertainty

The uncertainties of individual U/Th age determinations are small, but as demonstrated e.g. by the speleothem age-modelling work of Corrick et al., 2020, different but realistic assumptions about growth rates, interpolation methods, purity of samples etc. can lead to differences in ages at a certain speleothem depth that are larger than the raw U/Th age uncertainties (sometimes several times those). Especially at climatic transitions which are not located close to a U/Th-dated sample, this can lead to systematic dating offsets of D-O event onsets. If taken at face value, this forces the duration of the stadials and interstadials to change very significantly and way beyond what it compatible with the constraints from ice-core annual-layer counting (which is exactly what is seen here, described as an ice-core annual-layer-counting bias, line 139-141). This is why Buizert et al., 2014, stretched GICC05 by 1.0063 to fit the Hulu constraints ON AVERAGE and not on a transition-to-transition basis. This can likely be done better with Cheng 2016 data, but both the true age uncertainties from all sources (and not only the raw U/Th age uncertainties) and uncertainty due to that the D-O onset are not always similar between records must be included (and it is not clear if/how this is presently done). I believe that the current manuscript overemphasizes how tightly this one particular record (the Cheng/Edwards data) with its implicit assumptions about growth model, sample purity etc. can properly constrain individual D-O onset ages with realistic uncertainties, and that this introduces unrealistic stretching/compression of the ice-core time scale. Another way to address this problem would be to use data from other speleothems (e.g. the data from Corrick et al., 2020), and investigate if the results are reproducible under other assumptions.

Thank you for your suggestion –using multiple speleothem records is a good idea. I agree that dating uncertainties should be accounted for more explicitly in the manuscript. I have taken on board the Reviewer's suggestion and leveraged information from the speleothem compilation presented by Corrick and colleagues. As also discussed in my response to R2, I have devised a Monte Carlo Empirical Orthogonal Function (MCEOF) (e.g. Anchukaitis and Tierney, 2013) to extract the common large-scale d18O signature across a dataset of 17 speleothems from the Asian Summer Monsoon region that accounts for U-Th age uncertainties at each site.

The MCEOF method uses iterative age modelling of the available U-Th ages and eigen-decomposition of the d18O records to isolate the common d18O patter and estimate uncertainties. By creating ensembles of possible age models for each speleothem record, I generated a 1,000 member ensemble of the first leading mode of the MCEOF analysis (EOF1), whereby each EOF1 realization is dominated by the characteristic regional monsoonal pattern of the last glacial period. The CLIM synchronization was then performed by employing for each MCMC iteration a randomly resampled (with replacement) EOF1 realization from the 1,000 member ensemble, i.e. used as target for the alignment procedure (Fig. A). Note that the speleothem compilation additionally includes an updated version of the Hulu Cave data based on the newly published U-Th ages presented in Cheng et al. (2021), which significantly improve the resolution and dating precision of the cave record across Heinrich Stadial 4.

The MCEOF approach will improve the robustness of the CLIM synchronization and provide a more representative estimate of the alignment uncertainties. Furthermore, the new CLIM results bring about new implications (Fig. B):

1. **Unlike the results based on Hulu data only, there is an evident systematic ice-layer counting bias of GICC05 that is precisely centred around Heinrich Stadials (i.e. undercounting during H1, H3, H4 and H5, and overcounting during H2);**
2. **The inferred GICC05 relative counting error over these climatic events is more conservative than that presented in the previous version of the manuscript (i.e. within 10-15%);**
1. **The overcounting bias during GS-3 does not exceed 10%.**

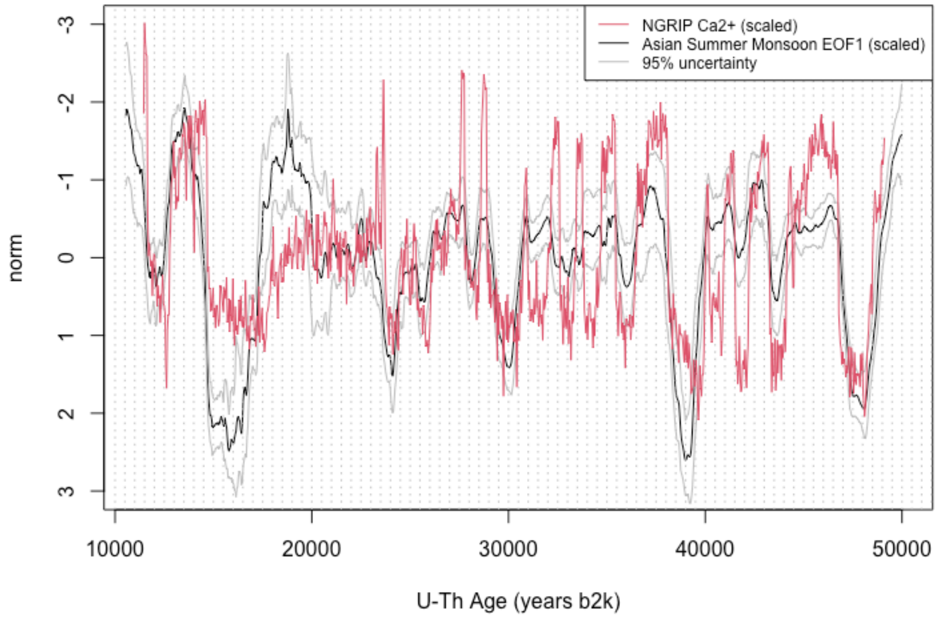

Fig. A – Preliminary CLIM synchronization of GICC05 to the U-Th timescale using the first leading mode from a Monte Carlo EOF procedure based on 17 speleothem d18O timeseries from the Asian Summer Monsoon region (Corrick et al., 2020). The NGRIP Ca2+ data are presented on the U-Th timescale using the posterior median estimate of the MCMC synchronization.

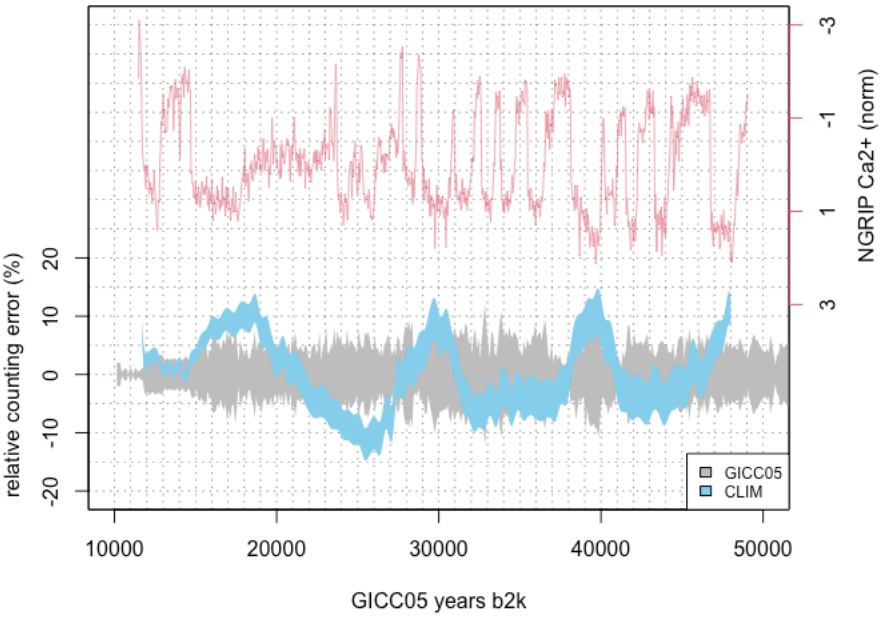

Fig. B – Inferred estimates of the relative annual-layer counting error for the GICC05 timescale based on the preliminary synchronization shown in Fig. A. Shading reflect pointwise 95% credible intervals. NGRIP Ca2+ data are also presented (red line) for reference. Note the systematic ice-layer undercounting during H1, H3, H4 and H5. The counting error exceeds the confidence levels predicted by the GICC05 chronology by up to ~5%.

3) Continuity

The method rests on an assumption that the records can be matched continuously, i.e. that there is robustly correlatable information everywhere in the record. There is always a best match between records being correlated, but the method seems not to address whether “best” is “good enough”. It thus becomes impossible for the reader to figure out in which sections the correlation is statistically significant and where there is nothing but noise (or local climate variability etc.) resulting in a transfer function that essentially just bridges between sections with statistically significant correlation.

I think there is a fundamental misunderstanding here. In a Bayesian context, the term “best fit” is at odds with the classical perspective from inferential statistics rooted in significance testing, hypothesis testing and confidence intervals (in fact, note that I generally refer to “optimal” rather than “best” synchronization). In Bayesian statistics it refers to the best possible fit given some prior estimates of the model parameters (in our case, the parameters defining the synchronization model), and the posterior mean (or median) should be taken as a measure of “adequacy” rather than as a deterministic quantity. There is a fundamental difference in philosophy between frequentist and Bayesian statistics. In the frequentist framework, the model parameters are unknown but deterministic quantities –meaning that the goal is to determine the range of values for the parameters that are supported by the data (i.e. the confidence interval). Bayesian statistics on the other hand is concerned with predicting the data given an estimate of the model parameters –regardless of whether these estimates are meaningful. When the parameters are viewed as deterministic quantities –as in frequentist statistics– it is nonsensical to talk about their probability distributions. Analogously, when parameters are viewed as probability distributions –as in Bayesian statistics– it is nonsensical to talk about their deterministic quantities and statistical significance. For these reasons, Bayesian results do not require validation via deterministic metrics (see also my reply to R2), and estimating the significance of the correlation between the speleothem and ice core data is not informative and overall incoherent with the probabilistic framework adopted here.

Specific comments

I will address all the specific points raised by the reviewer following their suggestions.

Line 287-293 and section 2.3.4: It seems like the method allows that sections of GICC05 are stretched/compressed to fit the assumed perfect Hulu time scale. If needed, sections close to each other can be modified in opposite ways (even though the stretching is smoothed as described in 2.3.3). This seems physically implausible given the nature of the GICC05 counting process: There are likely biases in GICC05, but these are not likely to change abruptly on short time scales (except between interstadials and stadials) because neither the data basis nor the counting method changed quickly. This is mentioned in line 391-395, but I think the results are not at all "approximating the layer-counting structure of the GICC05 timescale" but in stark contrast to the layer counting procedure.

As discussed in my reply to R1 and openly stated in the main text (see lines 356-357, and 391-395), the synchronization model relies on a simplistic formulation that is not meant to reproduce the complexity of the ice-layer counting and its uncertainty structure. That being said, I concede that interpreting GICC05 counting errors relative to only the

Hulu Cave d18O record can be misleading and that some of the features presented in Fig. 7c may be affected by noise and dating uncertainties in the Hulu data set. I think this issue will be mitigated by deploying the MCEOF methodology described above, i.e. following the reviewer's suggestion on combining multiple speleothem d18O records.

This is also what is seen on Figure 7c: The results indicate that at 26-28 ka, the GICC05 counting bias changes from 20-30% in one direction to 20-30% in the opposite direction. A similar slightly smaller feature is seen around 38 ka. These biases do not correlate with climate: e.g., the phases of largest overcounting happen during a stadial dust peak (26 ka) and during a low-dust interstadial (38 ka). If these features are real and unexplained, there is really no reason to trust GICC05 anywhere in the glacial. Extraordinary claims require extraordinary evidence, and I do not think that the manuscript provides sufficient evidence that these features are real, i.e. cannot at least to a large extent be attributed to weaknesses in the assumptions of Greenland - Hulu Cave climate correlation and underestimation of the total uncertainty of the Hulu Cave record, or alternatively, that the synchronization method does not produce statistically significant results in these sections.

As the reviewer suggested, integrating multiple speleothem records within the alignment procedure should improve the robustness of the CLIM results and their implications on the GICC05 counting bias. A preliminary synchronization using the MCEOF method (Fig. A & B) indicates that:

1. **the relative counting biases are much smaller than initially suggested, i.e. generally they are within 5% larger than the nominal GICC05 uncertainty;**
2. **the swings from undercounting to overcounting (and vice versa) are more in phase with abrupt climate shifts, especially during H1, H3, H4 and H5.**

References

Anchukaitis, Kevin J., and Jessica E. Tierney. "Identifying coherent spatiotemporal modes in time-uncertain proxy paleoclimate records." *Climate dynamics* 41, no. 5-6 (2013): 1291-1306.

Armstrong, Edward, Peter O. Hopcroft, and Paul J. Valdes. "A simulated Northern Hemisphere terrestrial climate dataset for the past 60,000 years." *Scientific data* 6, no. 1 (2019): 1-16.

Cheng, Hai, Yao Xu, Xiyu Dong, Jingyao Zhao, Hanying Li, Jonathan Baker, Ashish Sinha et al. "Onset and termination of Heinrich Stadial 4 and the underlying climate dynamics." *Communications Earth & Environment* 2, no. 1 (2021): 1-11.