Response to reviewers

Reviewer 1

Only the typos mentioned.
1. In line 333 of the track-changes version, replace "Van Meersbeeck" with "Van Meerbeeck"
Thank you for pointing out this typo! We have now replaced it.

2. In Table 1, replace "Burjachs & Ramon (1994) " with "Burjachs & Julià (1994)"
Thank you for pointing out this typo! We have now replaced it.

Reviewer 2

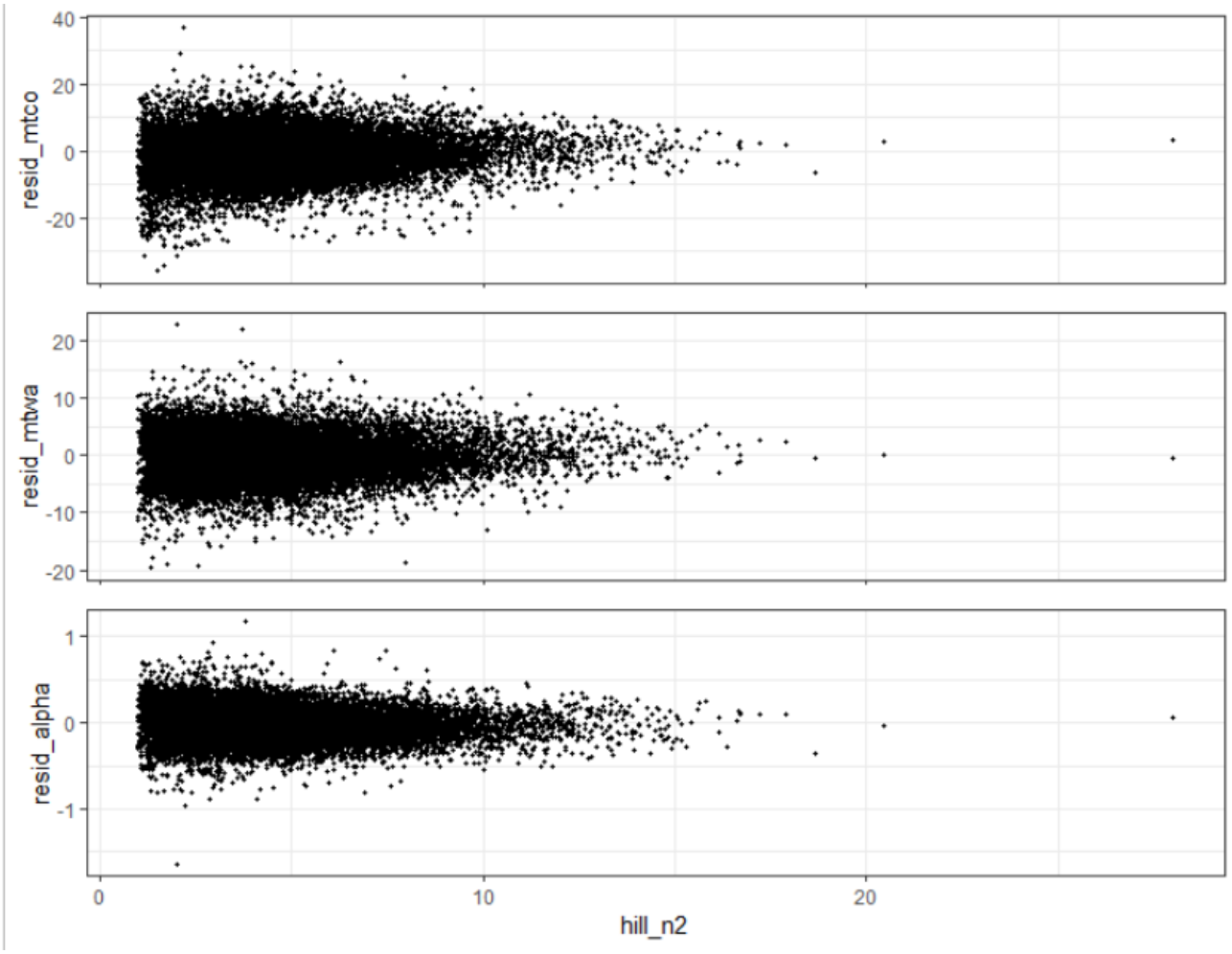
In this revised version of the manuscript, Liu and colleagues have significantly improved their manuscript. Many elements that needed to be clarified before are much better explained. The results are very remarkable and cohesive. Still, I am worried that some results were inverted just by changing the scale of the study (site-based vs gridded). How data are processed or represented impacts conclusions. Hence, I must ask the authors to show me more of their intermediate data a second time. Your results appear coherent, but your data are hyper-processed, and risks exist at every step. I want to see what is happening at each step to be able to accept your conclusions. In particular, I want to see:

1. Where the validation errors are located and identify types of vegetation and/or climates that cannot be reasonably reconstructed with that dataset. I hardly believe you have enough calibration data to analyse African and Latin American records properly.

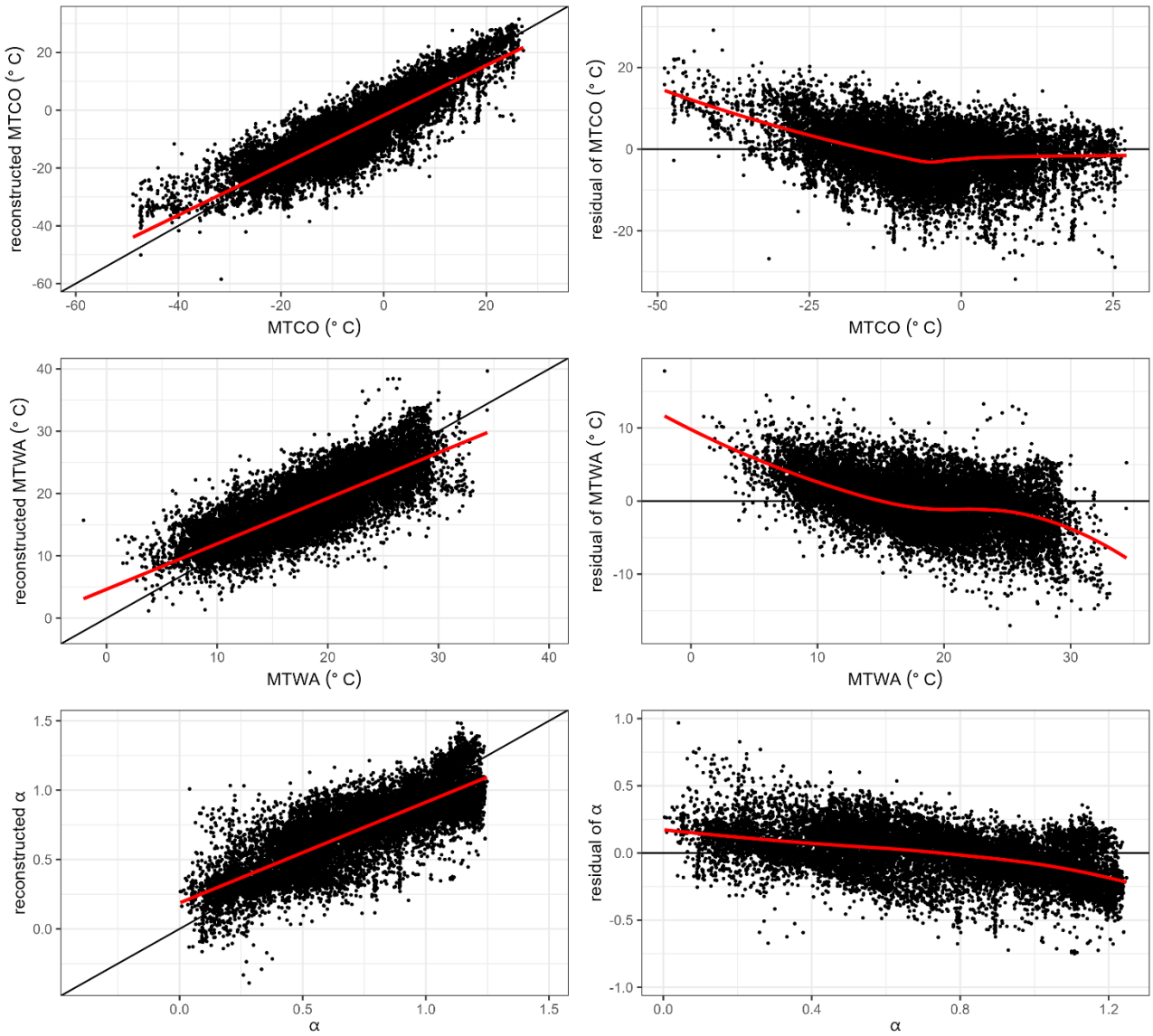
As pointed out in this question, what matters for the reconstruction is the vegetation composition and climate range the sample is in. The location doesn't matter, since the climate at the same location might be different at different times, and the modern climate at one location might correspond to climate at another location in the past.

Here is the relationship between Hill's N2 diversity (the effective number of taxa at each site) and the residual (reconstruction using the last significant number of components minus the actual value

in the modern training dataset). The general trend is that the more diverse the taxa, the less bias there is. However, the bias is also influenced by whether the taxa present have good quality, in other words, whether the taxa present have optima and tolerances with small uncertainty. Therefore, a specific reconstruction error should be obtained individually for each sample, that's why sample-specific error by bootstrapping is invited.

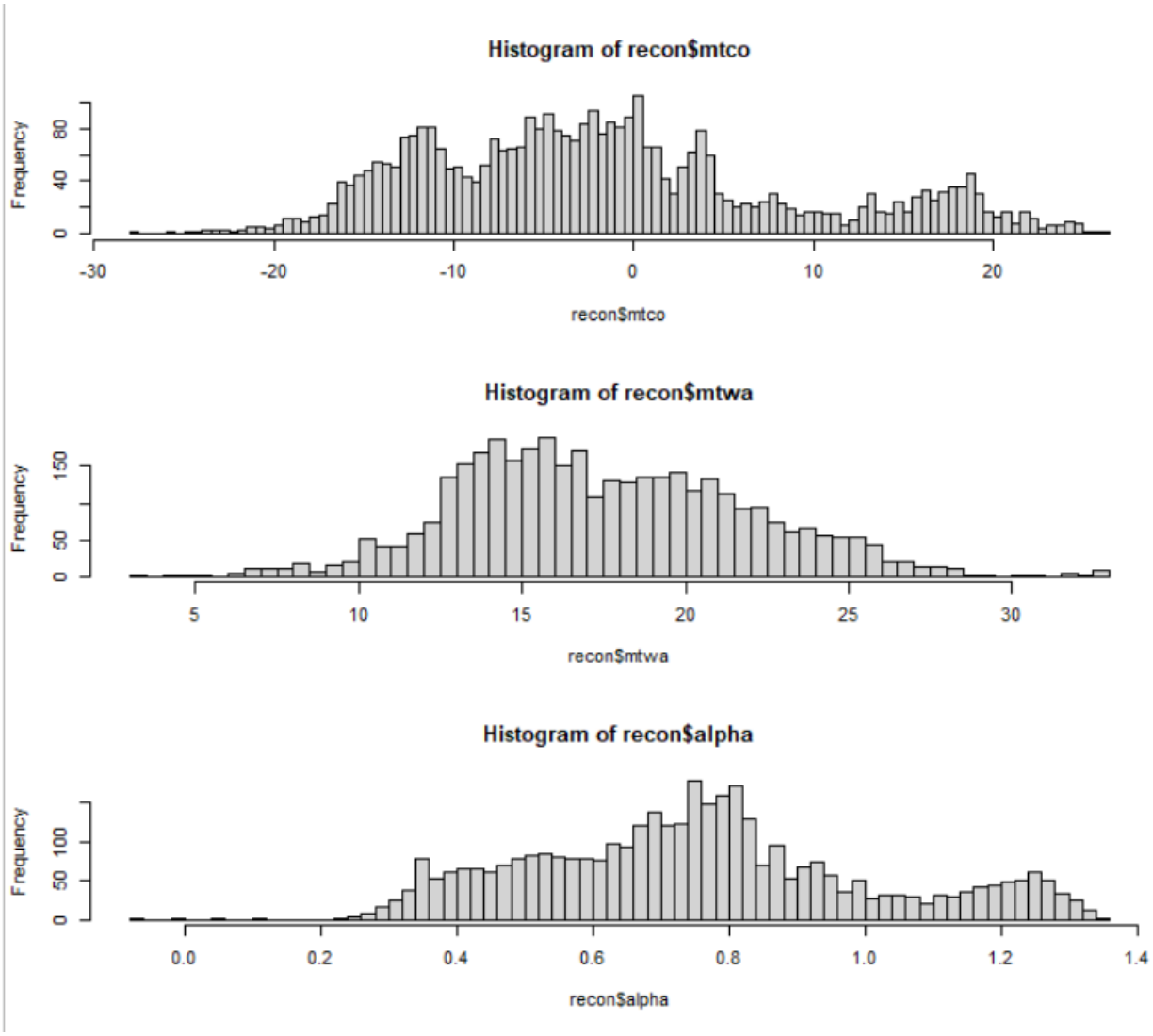


The figure below shows the training results using the last significant number of components for mean temperature of the coldest month (MTCO), mean temperature of the warmest month (MTWA) and plant-available moisture (α). The left panels show the relationship between reconstructed climates and actual climates in the modern training dataset, the black line is 1:1 line, the red line is the linear regression of the points to show the overall compression; the right panels show the relationship between the residual of reconstructed climates and actual climates in the modern training dataset, the black line is zero line, the red line is the loess (locally estimated scatterplot smoothing) regression of the points to show the local compression.



We can see that the climates have good coverage in −25 ~ 25 °C for MTCO, 10 ~ 30 °C for MTWA and the whole range for α, and that the biases in the centre of the range are relatively small. ***Climates outside these ranges might not be able to be reasonably reconstructed.***

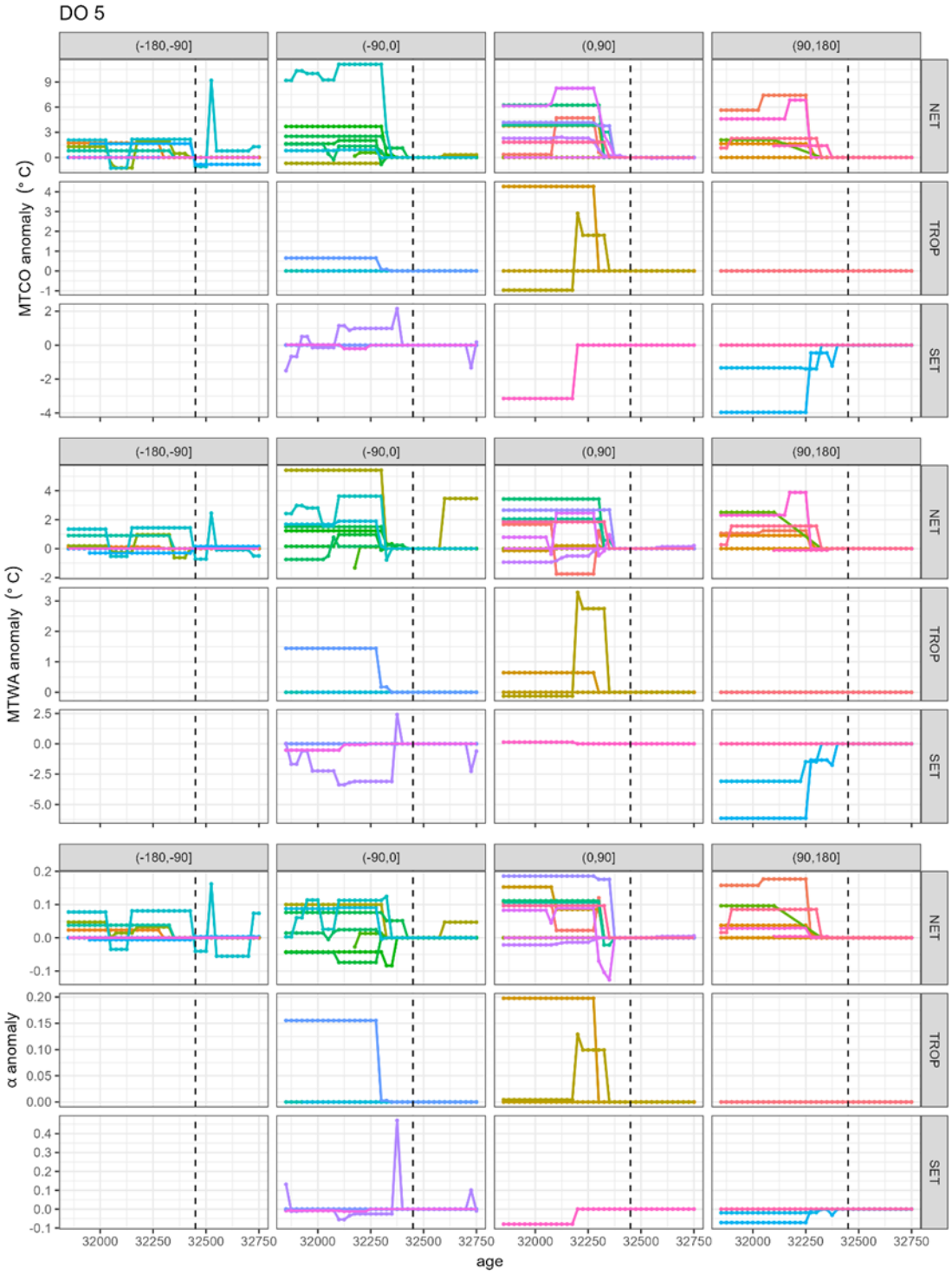
The figure below shows the distribution of the reconstructed climates during 50~30 ka (the period used in this paper). We can see that most of them are in the climate ranges with good coverage and low compression bias.



2. (some of) the reconstructions. Calculating peak-to-peak distance between DO events is ok, but what do the general trends look like?

Thanks! We have now added the time series for individual sites into supplementary.

Here is D-O 5 for example. Different sites are shown in different colours and segmented into different longitude + latitude combinations, using reconstructions with dynamic time warping adjusted ages. The anomalies are anomalies to the values at official start date corresponding to Greenland D-O warming (shown in vertical dashed lines).



Besides, we have made further improvements to our identification methods. In the previous version, the minimum and maximum of the polynomial were got within the period of official start date − 600 yrs ~ official start date + 300 yrs.

```
f_polynomial<-function(x){ a*x^3+b*x^2+c*x+d }
min_t_polynomial <- optimize(f_polynomial, c(t-600, t+300) )$minimum
max_t_polynomial <- optimize(function(x)-f_polynomial(x), c(t-600, t+300) )$minimum
```

However, sometimes the reconstructions don't fully cover this period, resulting in NAs. Therefore, we restrict the minimum and maximum to be found at where there are reconstructions.

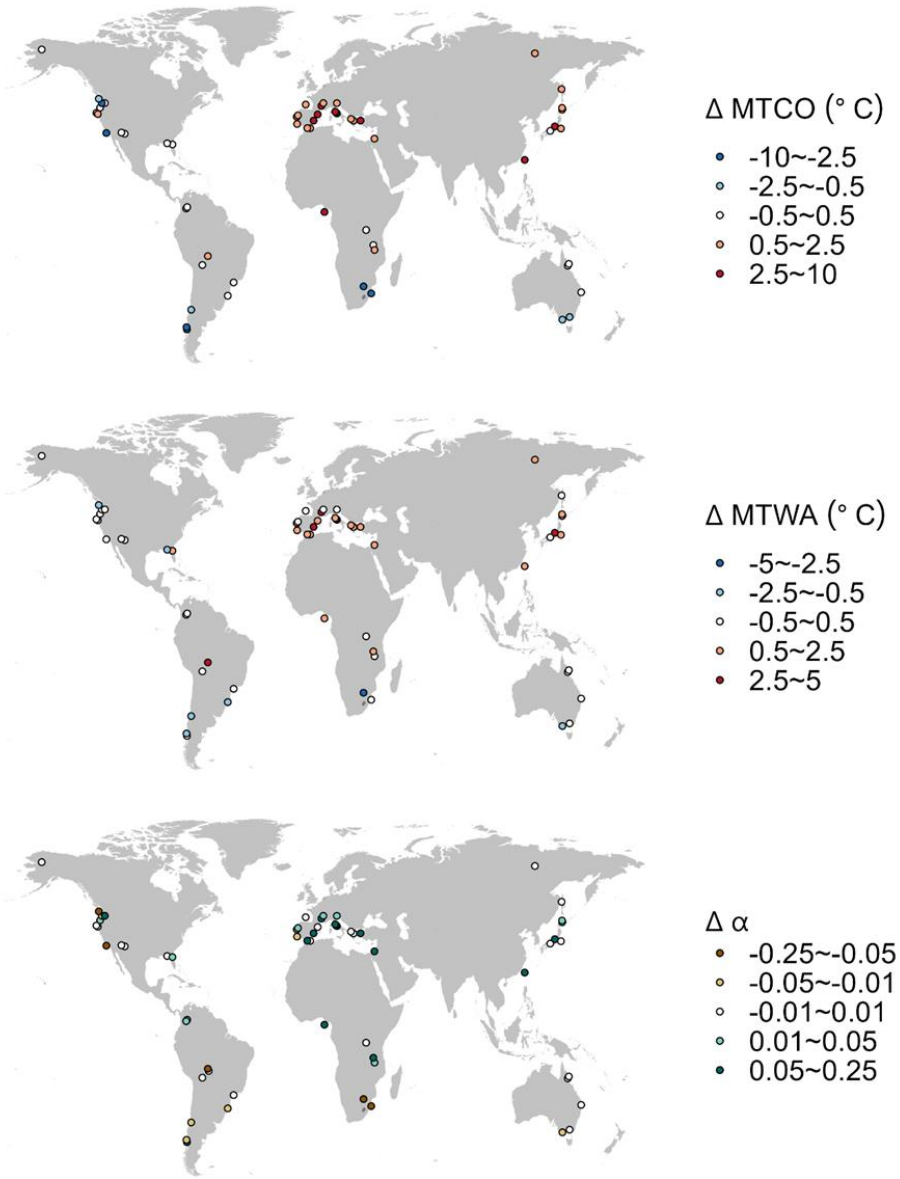
```
f_polynomial<-function(x){ a*x^3+b*x^2+c*x+d }
min_t_polynomial <- optimize(f_polynomial, c(min(sub_map_win$age), max(sub_map_win$age)) )$minimum
max_t_polynomial <- optimize(function(x)-f_polynomial(x), c(min(sub_map_win$age), max(sub_map_win$age)) )$minimum
```

Here is the change to the text:

We used the ages corresponding to the minimum and maximum in the fitted polynomial ($t_{min\ polynomial}$, $t_{max\ polynomial}$) but restrict $t_{min\ polynomial}$ and $t_{max\ polynomial}$ to be found at where there are reconstructions.

By doing so, the number of events identified increased from 278 to 298.

Here is the new figure 5. The general trend is very similar to the previous version.



3. The RMSEP presented in Table 2 are – sorry for the term – horrible, and I am not convinced by your argument that "most" of the reconstructed differences between the warm and cold phases of the DO events are larger than this. You used the 1-sigma error, which is extremely conservative. If you used a more common 2-sigma interval, most of your reconstructions will fit within the "v-shaped" black lines. On average, the reconstructions are 6.5°C off target for MTCO. This is huge, especially for a leave-one-out cross-validation. If we propagate that error to the reconstructions and take the 2-sigma interval, your error bar will be +/- 13°C, a 26°C range. It is surely possible to do better than that. This takes me back to my original request to see the data this analysis is built on. Where are the samples with such high errors, and are they potential candidates for the D-O periods? If so, how do you deal with them in the analysis?

Thanks! First, RMSEP for modern training dataset and the sample specific errors for real fossil reconstructions are different. In the modern training dataset, there are sites with good-represented taxa (which will cause small reconstruction errors) and sites with poor-represented taxa (which will cause large reconstruction errors). All of them account for the final RMSEP. However, what matters for a specific reconstruction is the quality of the fossil taxa in that sample itself, as long as the fossil taxa in that sample is well-represented, we will have small errors. That's why we are calculating sample specific errors.

We agree that the error is an issue when looking at the spatial patterns. Here is the figure 5 only using samples with $|change| \geq 2|sample\ specific\ error|$ for each climate variable. The version without this filter has 342 samples, after applying this filter individually for each variable, there are 90 samples for MTCO, 52 samples for MTWA and 127 samples for α. Although the number of samples are reduced, we can see that the general trend is very similar to the version without this filter, and even slightly more apparent. We have now put this into supplementary.

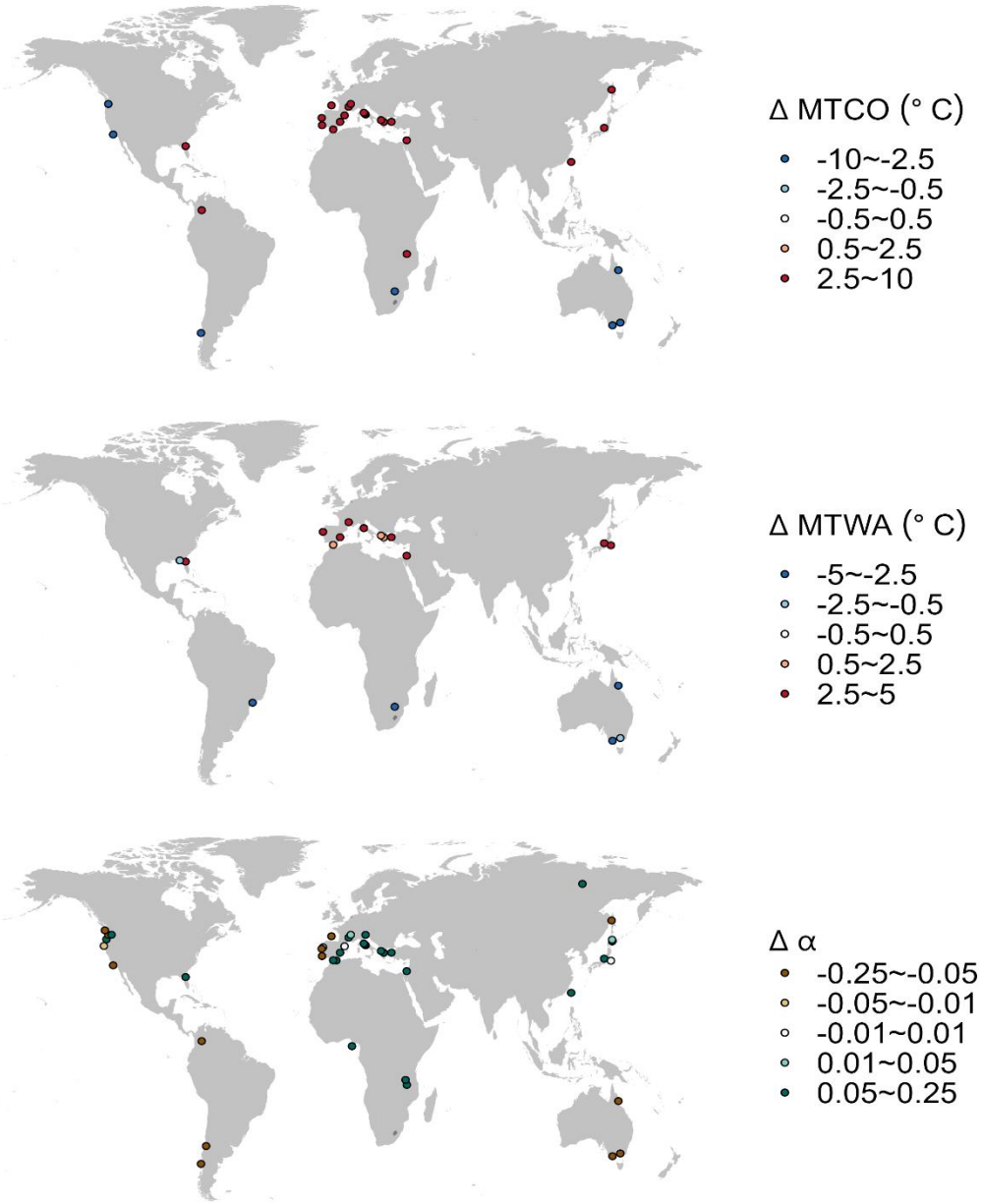


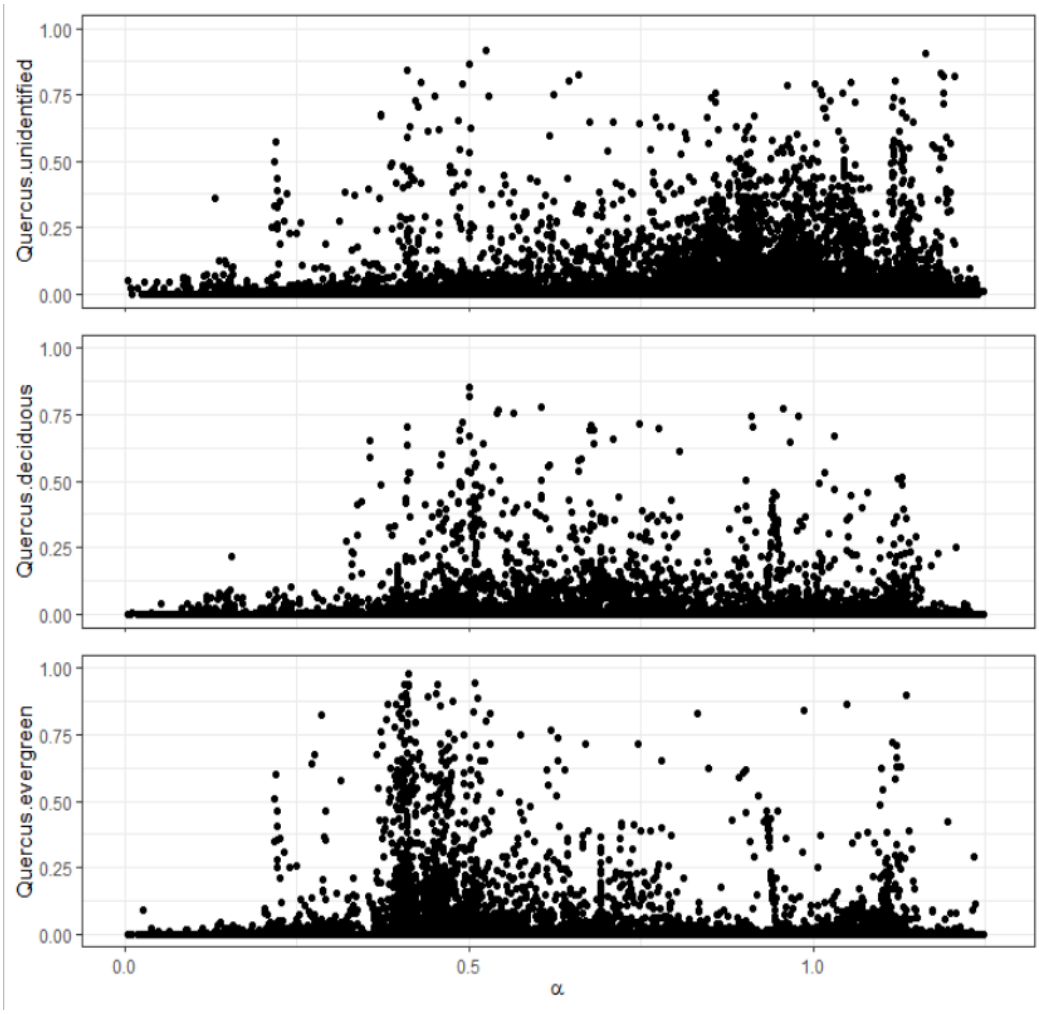
Table 3 and Table 4 are not influenced since maximum likelihood estimation of the ratio already considers the different errors between samples (both the values and the errors of the values are inputs when using this method, and both axes are considered as well).

4. L238-253: This string of arguments is very theoretical, and I don't see many convincing applications of this. You mention Turner et al. (2020). But even if the calibration dataset was extensive in that study, it was still confined to the Palearctic realm, which contains ecosystems that

have co-evolved. Here, you are mixing data at an even larger scale and assuming a transitivity of the property relative to the spatial scale. I'm doubtful that mixing South American, European, and Japanese Quercus make much sense, especially in a regression model where all data are combined. As mentioned in the previous round by the other reviewer and myself, mixing the deciduous and evergreen Western European Quercus is already limiting, as these taxa bring very distinct climate information, even if not much can be done here. Finally, this is an assumption of all your results, not an element of discussion. Bring that part to the methods part.

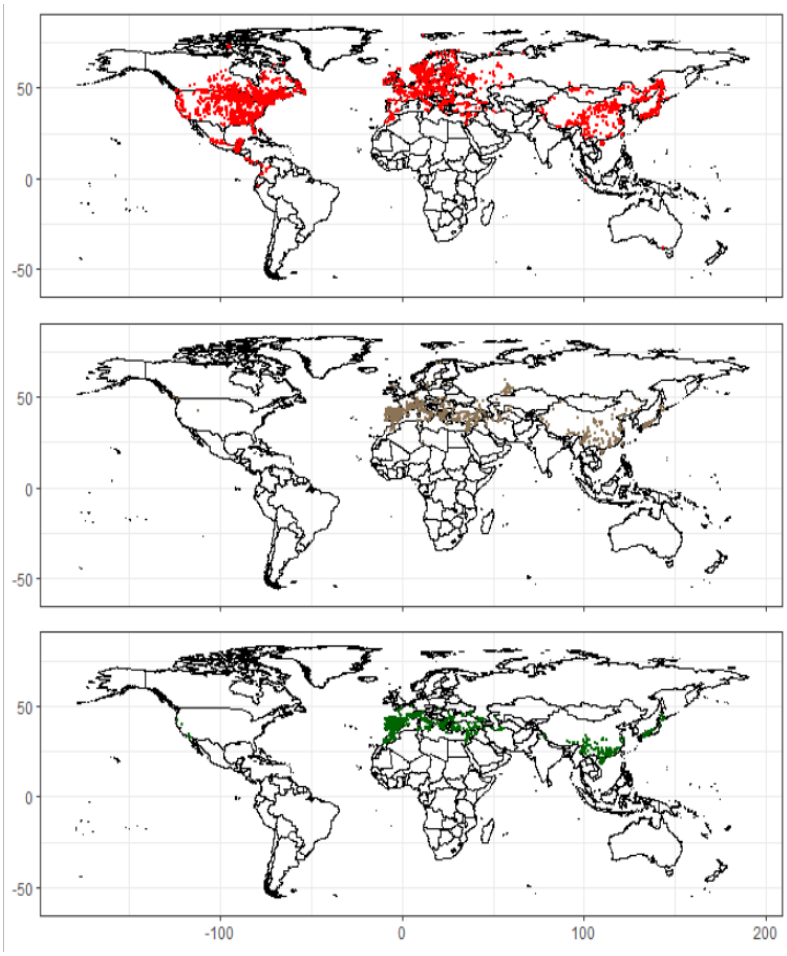
Thanks! We have now moved it to the methods part (section 2.1).

The figure below shows the distributions of Quercus unidentified, Quercus deciduous and Quercus evergreen. There are 17547 samples in total. The problem is that Quercus unidentified is present in 6772 samples, while Quercus deciduous is only present in 2949 samples and Quercus evergreen is only present in 2903 samples. Quercus deciduous and Quercus evergreen seem to have similar patterns with slightly different optima, while Quercus unidentified seems to have an extra cluster at the wet end.

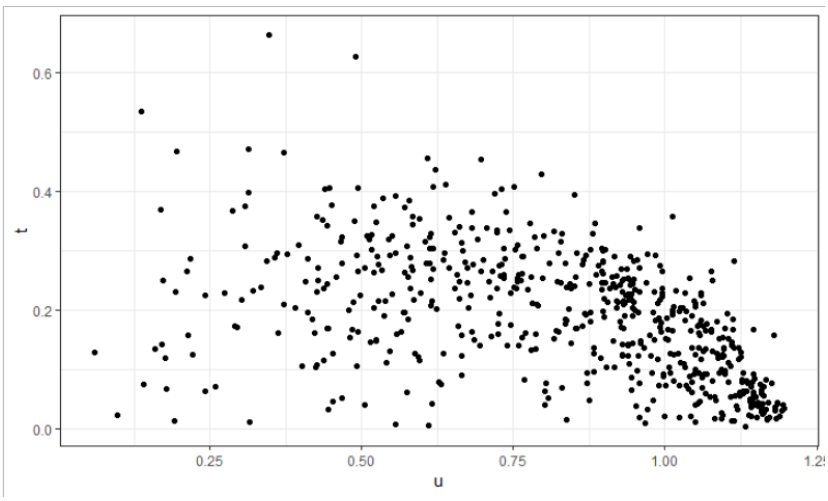


The optimum for Quercus unidentified is 0.81 with a tolerance of 0.28, the optimum for Quercus deciduous is 0.65 with a tolerance of 0.24, the optimum for Quercus evergreen is 0.53 with a tolerance of 0.21, based on calculations using modern data set.

We also checked where these Quercus unidentified are. The figure below shows the spatial distribution of Quercus unidentified (red), Quercus deciduous (brown) and Quercus evergreen (green). The unidentified Quercus is widely present in Eurasia, where both deciduous and evergreen are present, which we can't assign arbitrarily. The Quercus seems not identified in the whole America, which might be the reason why Quercus unidentified has a different optimum.



We agree that not much can be done here, especially when half of the Quercus are unidentified. We can either combine all the Quercus or remove all the Quercus. *The figure below shows taxa information for α, with taxa optima (u) in the x-axis and taxa tolerances (t) in the y-axis. We can see that there are still enough taxa with similar optima and tolerances with Quercus. So even when we combine or remove all the Quercus, the reconstruction of α can still be made.*



Here are the results when we remove all the Quercus in the reconstruction. We can see that the general trends are quite similar to the version when we combine all the Quercus.

Figure 5 (no Quercus): Maps showing the median change of site-based reconstructions for Dansgaard-Oeschger (D-O) events.

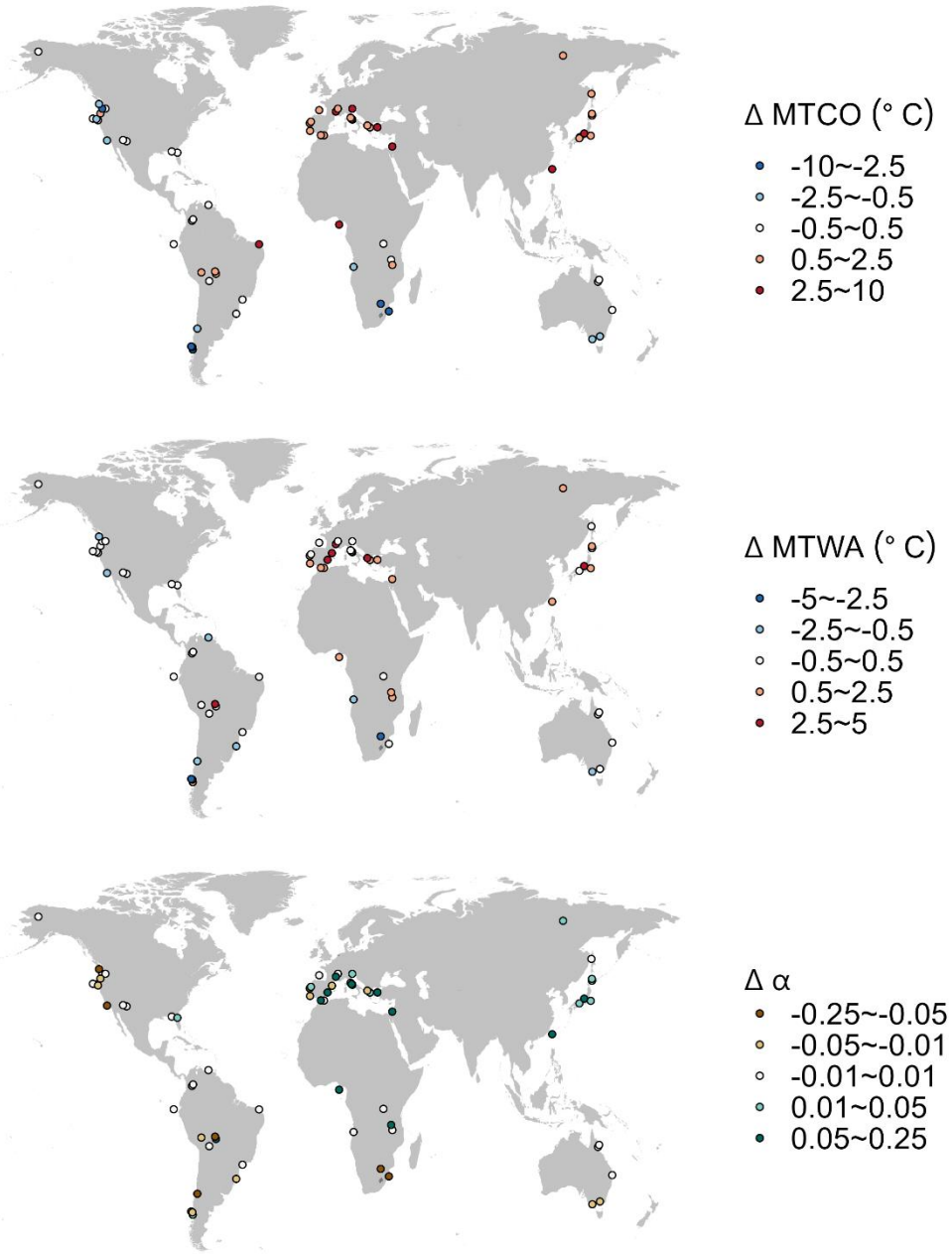


Figure 4 (no Quercus): Scatter plot of the change in plant-available moisture (Δα) versus the change in mean temperature of the warmest month (ΔMTWA) during individual Dansgaard-

Oeschger (D-O) events at individual sites. The points are colour-coded to indicate whether the sites are from the northern extratropics (NET, north of 23.5°N), the tropics (TROP, between 23.5°N and 23.5°S) or southern extratropics (SET, south of 23.5°S).

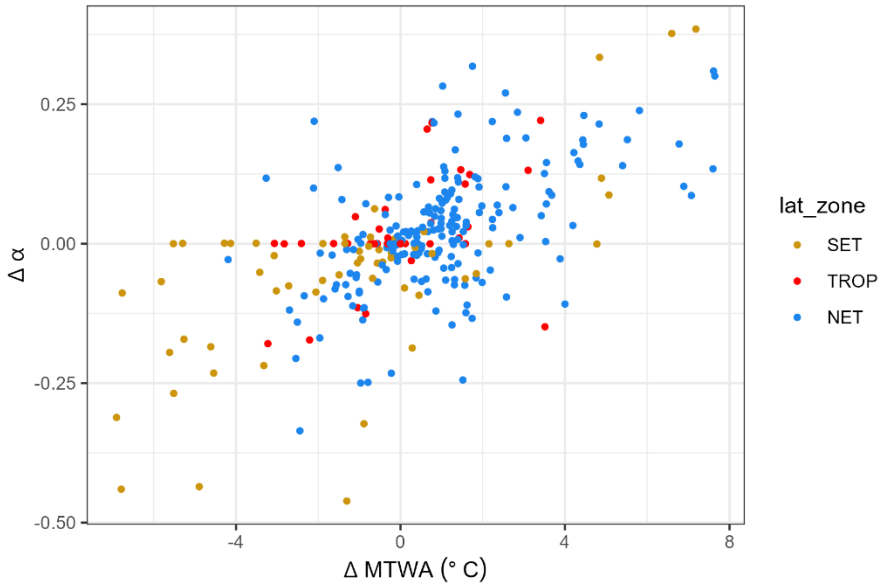
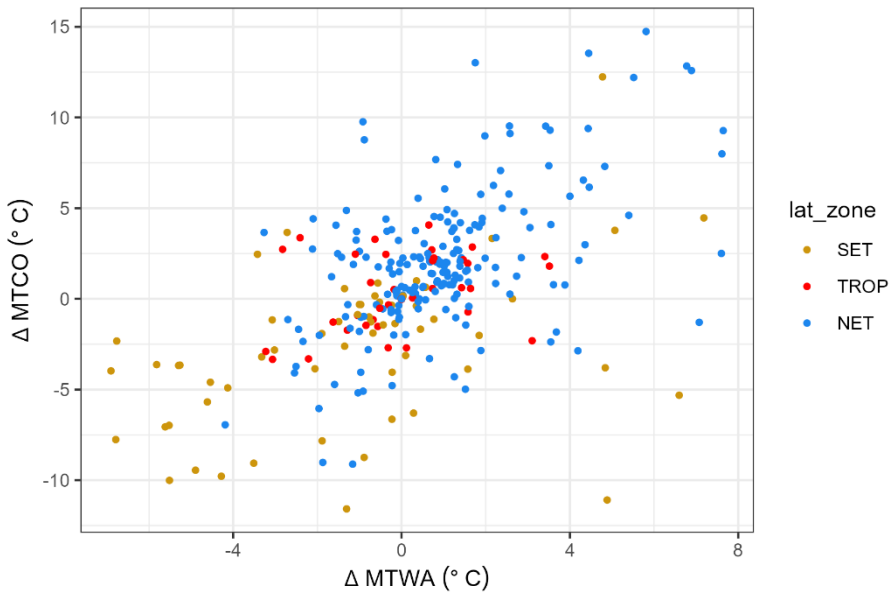


Figure 3 (no Quercus): Scatter plot of the change in mean temperature of the coldest month (ΔMTCO) versus the change in mean temperature of the warmest month (ΔMTWA) during individual Dansgaard-Oeschger (D-O) events at individual sites. The points are colour-coded to indicate whether the sites are from the northern extratropics (NET, north of 23.5°N), the tropics (TROP, between 23.5°N and 23.5°S) or southern extratropics (SET, south of 23.5°S).



We can either 1) keep results using combined Quercus in the main text and put results using no Quercus (with the same number of components) at Supplementary or 2) re-run leave-out cross validation and replace all the current results with results using no Quercus. Both options are OK for us, we can apply in the next round of revision, according to the reviewer's preference.

5. Finally, at L125, I would like to understand the rationale behind excluding climatically close samples. While I appreciate the need to remove spatially close samples to account for spatial autocorrelation, even if regression techniques are not the most sensitive to it, I'm unclear about the decision to exclude geographically distant yet climatically close samples. This exclusion may be contributing to the high RMSEP and I believe it's important to justify this decision.

Yes, excluding climatically close samples will definitely have higher RMSEP than not excluding them! In other words, using leave-out cross validation will definitely have higher RMSEP than the traditional leave-one-out cross validation. The purpose of it is to make sure that we can reconstruct climates that are not in the training dataset at all to check the predictive power. For example, the whole range of MTCO in SMPDSv2 is − 49 °C to 27 °C, 2% of the full range is 2%×76 °C = 1.5 °C, if the sample to be reconstructed is 0 °C, we need to make sure that there are no training samples in the range – 1.5 °C to 1.5 °C within 50 km horizontal distance from the site. By doing so, we are more convinced that we can reconstruct climates even when our training set can't cover the climate range we want to reconstruct.

Reviewer 3

Since it is the second round of review, I am not a specialist in pollen records but the two reviewers are, I am not assessing the derivation of the quantitative estimates.
This compilation of warm and cold month temperature change, with the majority of the proxy on land presents a new aspect of DO variability that will be very useful to the community. I however think that the authors could improve the discussion part by providing some comparisons with existing modelling studies and thus providing some information about the processes at play.

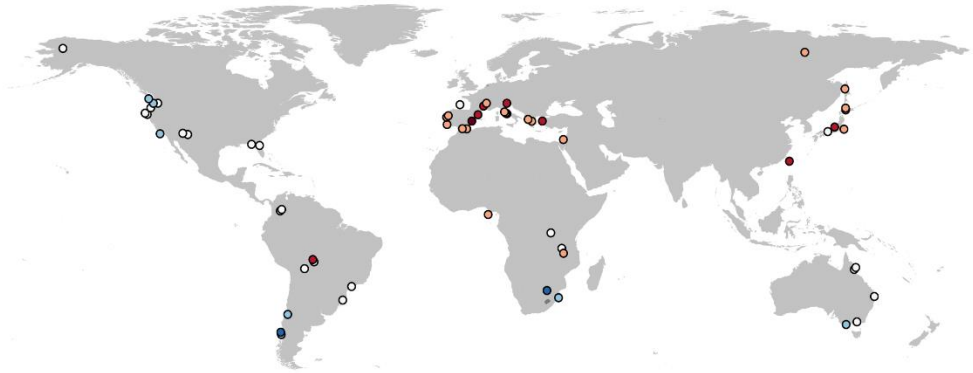
1. The authors repeat 3 times that this dataset will be used to compare with upcoming PMIP meltwater exp. But why don't they already assess existing exp? I understand that there are only a handful of modelling exp are available for MIS3 but there are also some meltwater exp. performed under LGM conditions (eg Kageyama et al. 2013). Please also compare your moisture estimates to changes in precipitation (if that's what it means). For example, the most interesting aspect of your compilation is that the data suggests a cooling over western North America during DO events. That is only consistent with AMOC shutdown exp. in which the NPIW formation strengthens, as in Menviel et al. 2014 for example.

Thanks! α is plant-available moisture, which is influenced by precipitation, temperature and sunshine hours (see the next question for details). Therefore, it can't be directly compared with precipitation. We do have temperature simulations which can be compared. LOVECLIM (Menviel et al. 2014) provides MAT (mean annual temperature) simulations for D-O 5~12, we can use the average of MTCO reconstructions and MTWA reconstructions as an approximation to MAT reconstructions.

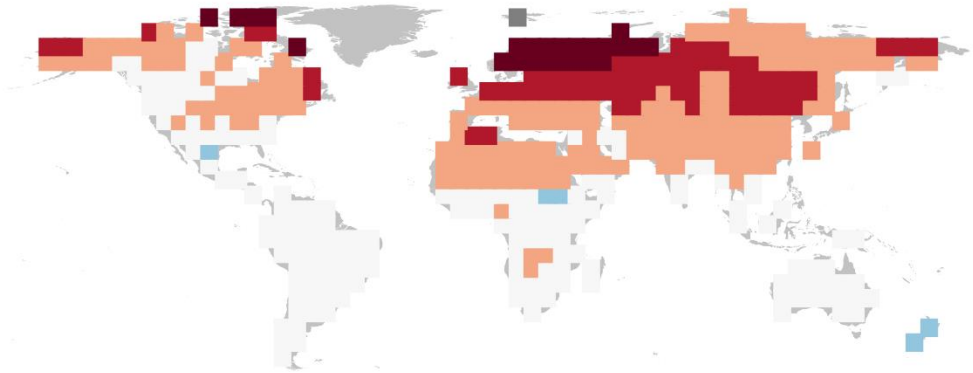
Here is the comparison of median MAT change of D-O 5~12 for reconstruction (upper panel), LOVECLIM simulation (middle panel), and 3D var assimilation (using LOVECLIM simulation as prior and using reconstruction to adjust, more details can be found in Liu et al. (2024)).

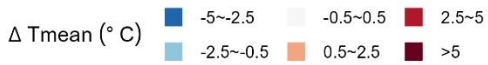
Mengmeng Liu, Iain Prentice, Laurie Menviel et al. Rapid ice-age warming events amplified by strong vegetation-albedo feedback, 26 March 2024, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-4000395/v1]

reconstruction

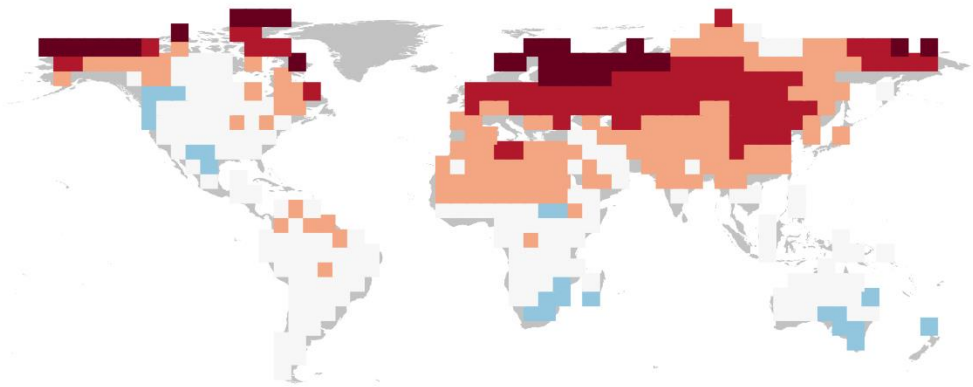


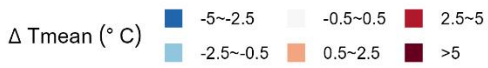
model





3D var





We can see that the general patterns are very similar. However, there are a few mismatches between the reconstruction and the LOVECLIM model, for example, the cooling in the western North America and southern extratropics may not always reach the land and may just stay in the ocean, so there are no changes seen when you consider the median. The 3D var results use reconstruction

to adjust the pattern, so there are no such problems. This indicates the potential of future models to incorporate reconstructions to improve the predictive powers.

The 3D var paper is still under review, which might need further modifications. Hopefully we can get this comparison figure set and put into the supplementary before the next round of revision for this paper, or we can just compare the reconstruction and the LOVECLIM model.

2. It would be good to clarify what alpha represents for modellers: is it precipitation or evaporation-precipitation or is it more related to soil moisture?

Thanks! α is plant-available moisture, calculated as the ratio of actual evapotranspiration to equilibrium evapotranspiration. It is a transformation of the commonly used moisture index MI, to emphasize the differences at the dry end of the climate range, which have a more pronounced effect on vegetation distribution than differences at the wet end. And MI is calculated using SPLASH based on daily values of precipitation, temperature and sunshine hours obtained using a mean-conserving interpolation of the monthly values of each. More details can be found at:

Liu, M., Prentice, I.C., ter Braak, C.J.F, and Harrison, S.P.: An improved statistical approach for reconstructing past climates from biotic assemblages, Proc. Royal Soc., Math. A, 476, 20200346, 20200346, https://doi.org/10.1098/rspa.2020.0346, 2020.

We have now modified the sentence in the Methods:

The SMPDSv2 also provides climatic information at each pollen site, specifically the mean temperature of the coldest month (MTCO), mean temperature of the warmest month (MTWA), and a plant-available moisture (α) calculated as the ratio of actual evapotranspiration to equilibrium evapotranspiration. α is a transformation of the commonly used moisture index MI, to emphasize the differences at the dry end of the climate range, which have a more pronounced effect on vegetation distribution than differences at the wet end (Prentice et al., 2017). Detailed relationships are in Supplementary Materials. These bioclimate variables reflect mechanistically distinct controls on plant growth.

Prentice, I. C., Cleator, S. F., Huang, Y. H., Harrison, S. P. and Roulstone, I.: Reconstructing ice-age palaeoclimates: Quantifying low-CO_2 effects on plants, Glob. Planet. Change, 149, 166–176, doi:https://doi.org/10.1016/j.gloplacha.2016.12.012, 2017.

We have now put the following paragraphs into Supplementary which explains the relationship between precipitation, MI and α.

Transformation of MI into α

The definitions of moisture index (MI), aridity index (Φ) and α are as follows:

$$MI = \frac{P}{E_p} \quad (S1)$$

$$\emptyset = \frac{E_p}{P} = \frac{1}{MI} \quad (S2)$$

$$\alpha = \frac{E_a}{E_q} \quad (S3)$$

where P denotes the annual precipitation; E_p denotes the annual potential evapotranspiration; E_a denotes the annual actual evapotranspiration; E_q denotes the annual equilibrium evapotranspiration.

There is a relationship between E_p and E_q:

$$E_p = 1.26\, E_q \quad (S4)$$

The parametric Fu-Zhang formulation derived from the Budyko relationship is:

$$\frac{E_a}{P} = 1 + \emptyset - (1 + \emptyset^{\omega})^{\frac{1}{\omega}} \quad (S5)$$

where P denotes the annual precipitation; E_a denotes the annual actual evapotranspiration; ϕ denotes the aridity index; ω is a parameter functioning as a curve-shape parameter in the model. Values of ≈ 3 have been used in many applications.

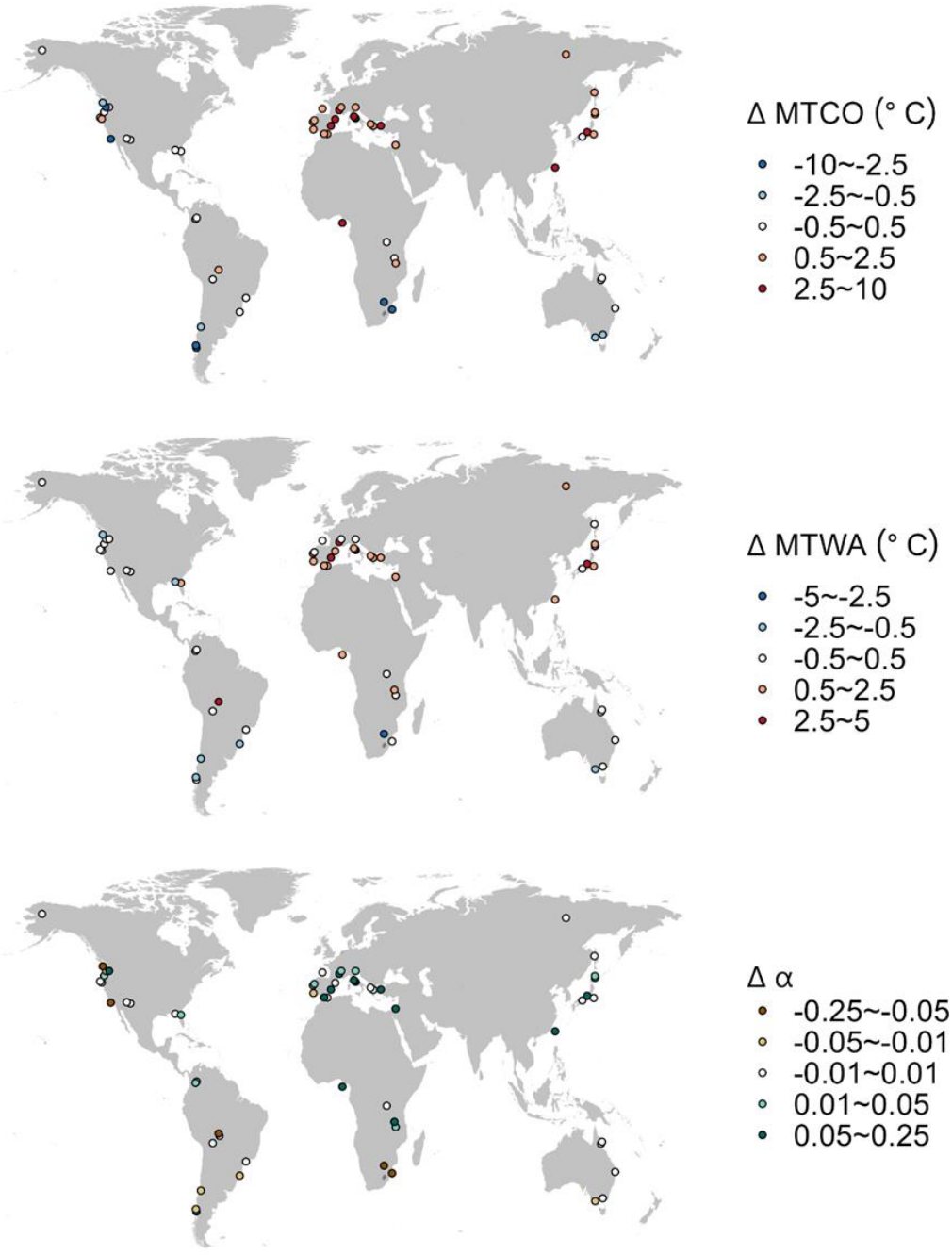
Therefore:

$$\alpha = \frac{E_a}{E_q} = 1.26\, \frac{\frac{E_a}{P}}{\frac{E_p}{P}} = 1.26 \cdot MI \cdot \frac{E_a}{P} = 1.26 \cdot MI \cdot \left(1 + \frac{1}{MI} - \left(1 + \left(\frac{1}{MI}\right)^{\omega}\right)^{\frac{1}{\omega}}\right) \quad (S6)$$

3. Figure 5 is the main figure with important information, but it is a bit difficult to see the colors

and differentiate between the two blues and two reds. Maybe you could make the markers bigger and/or choose reds and blue shades that are more different.

Thanks! We have refigured it as below:



4. L 185-186: all the equal signs are missing.

Thanks! We have now added equal signs.

The performance is best for MTCO (R^2 = 0.75, RMSEP = 6.51, slope = 0.85) but is also good for MTWA (R^2 = 0.59, RMSEP = 3.68, slope = 0.71) and α (R^2 = 0.65, RMSEP = 0.18, slope = 0.71).

5. L 293-294: you already mention that at least in 2 other parts of the manuscript. Please remove at least one occurrence.

Thanks! We have mentioned PMIP in the abstract and in the paragraphs above, so we will remove the one in L 291-295 in the original version.

~~*Nevertheless, this first compilation of quantitative climate reconstructions through multiple D-O events during MIS3 provides an opportunity for evaluation of the transient D-O simulations planned as part of the next phase of the Palaeoclimate Modelling Intercomparison Project (Malmierca-Vallet et al., 2023).*~~