Summary:

A small number of climate-sensitive proxies are used with a simulation from an isotope-enabled coupled Earth system in a field reconstruction of North Atlantic and European climate. The reconstruction spans annual, seasonal, and monthly timescales depending on the climate variable. The reconstruction is validated against instrumental products, including 20CRv3 for the atmospheric variables, and a suite of SST reconstructions for the ocean. Agreement depends on the climate field and the time period, with the best agreement in locations with abundant high-quality proxies (e.g., summer temperature in regions with latewood density records). Skill at monthly timescales, and especially for precipitation and mean-sea-level pressure, is generally poor.

There are two main issues with this study, concerning insufficient description of the method and weak validation, that need to be cleared before I can recommend publication. I elaborate on these points below and then give minor comments.

Thank you for the thorough review and constructive comments. We appreciate this opportunity to improve our work and clarify the writing in our manuscript.

Main Points:

1) *I've read the method section twice, and I would be unable to reproduce the results in this study. The method uses a weighted-analog approach from a single model simulation. The weights are determined by misfits to the proxy records, but I don't see anything about how this comparison is made when the proxy is not a climate-model variable (e.g., MXD and ring width). This suggests that these records have been inverted for temperature, but perhaps they all have (including d18O)? This is very important as it affects not only the weights, but the relevance of the validation process.*

Thank you for pointing out this oversight. The tree-ring (MXD, TRW, BI) data are matched with model JJA T2m. All isotope records are matched with precipitation weighted model d18O. This is the same approach as Sjolte et al. 2020, but this should of course be clear in this paper as well. We will include a table in the revision to summarize this information.

The ad hoc variation inflation method has no justification other than the reconstructed time series has more variance. Lacking any evidence, I assume that this simply adds noise and not signal; i.e., it serves no purpose.

The inflation neither adds noise nor signal if we view it in terms of correlation. We agree that the motivation of the variance correction can be clarified in the manuscript. The variance of the model (ECHAM5-MPIOM) is the variance of the sample prior, which is the best estimate of the variance we can obtain when using this model for the assimilation. In Figure S5 and S6 we show that the variance corrected reconstruction performs well compared to reanalysis data. We will include this explanation in the revision.

2) *The validation of the results is particularly weak. If, as I suspect, that many proxy records have been inverted for temperature, then those inversions have been*

calibrated on the instrumental record. If true, then the validation is in sample, and doesn't independently measure skill. If I have that wrong, then this should serve as motivation to describe this aspect very clearly.

The reconstruction is not calibrated. Our method is only based on matching the model output with proxy data. As shown in Eq. 1 the data used for summer are normalized, and no tuning to observations is involved. For winter we only use d18O-based proxies, and match the model d18O with ice core and tree ring cellulose d18O (not normalized, as explained L132-135). Consequently, the validation against observed datasets (reanalysis, long-term temperature records and SST compilations) serve as a truly independent measure of skill. We will clarify this point in the revision.

Given how computationally cheap the reconstruction method is, another important validation approach is to leave out proxies for validation, and repeat the reconstruction process. This can take the form of a jackknife, or leaving out sets of proxies. The reconstruction can then be used to predict the proxies that were left out as a way to independently validate the results, and to test how sensitive they are to the excluded proxies. I would be particularly interested to see results when the Greenland proxies are left out, since Greenland ice core d18O correlates very weakly with European temperature (e.g., Hörhold et al., 2023).

We investigated the sensitivity to choice of proxy data in Sjolte et al. (2020) with a similar method and proxy data set. This is discussed in L178-182. We do appreciate the reviewer's interest in this question and have carried out the requested test of using Greenland vs European proxies.

We made two new reconstructions based on our annual reconstruction. One based only on ice core data ($RECON_{IC}$) and once based only on tree-ring data ($RECON_{TR}$). We then extracted the data for tree-ring sites from $RECON_{IC}$ and the ice core data from $RECON_{TR}$. The results show that the prediction of tree-ring variability from $RECON_{IC}$ has weak skill, although some correlations are significant (see Figure R1-4). The best correspondence is seen for Scandinavian tree-ring records (TAA, EFmean) and winter d18O correlated to CZEC d18Ocell. Conversely, the prediction of ice core variability from $RECON_{TR}$ also has weak skill, with the best correlation to GRIP d18O (0.33). This underlines the different properties of the proxy data due to differences in seasonality, region and what is recorded by the proxies. In broad terms, the main strength of the ice core data is to record large-scale atmospheric circulation for winter, while the tree-ring data capture temperature changes on a more local scale, but the wider availability of tree-ring data ensures a reasonable spatial coverage to capture a large-scale summer signal. Our main conclusion from this test is that the properties of the different sets of proxy data are complementary and constrain different seasons and regions of the reconstruction.

We will include these new figures (Figure R1-6) in the supplementary and discuss the results in section 3.1 *Ensemble reconstruction and initial evaluation* of the revised manuscript.

Minor points:

1) *You should account for pattern correlation in the significance calculation; I suspect the counts of significant grid cells given in the figures is not far different than random at monthly timescales*

The significance (Student's t-test) is calculated for each grid cell. The number of samples for monthly anomalies is large (8 730) and relatively low correlation correlations can be significant ($p < 0.01$) (L209-210). For individual months the correlations for T2m peak at 0.70 (L211-212) (Figure S7). We don't view the count of the number of grid cells with significant correlation as a strict test, but as a rough indication of how wide-spread the significant correlation is. In any case, grid cells are not exactly area, depending of the grid cells and we are aware the number of grid cells will depend on season, variable and location. We will include details on correlation, significance and how we view the number of grid cells with significance in the methods section of the revision.

2) *line 8: As I indicated above the benefit of this variance inflation needs to be shown. In particular, that the inflated variance is signal, not noise.*

See reply to Main Point 1).

3) *line 54: terms*

We will correct this.

4) *line 83: one simulation only?*

Yes, one simulation. As stated, it is the same simulation as used in Sjolte et al. (2018, 2020).

5) *line 110: reconstructions*

We will correct this.

6) *line 126: What does this mean when the proxy is not a model variable?*

See reply to Main Point 1.

7) *line 130: time series variance? over what time period? why would this be error?*

We normalize by the combined standard deviation (entire timeframe 1241-1970) which is the standard formulation of the Chi-2 measure. It is the same as normalizing alle records individually before matching, but the equation is more compact in this formulation. We will explain this in the revision.

8) *line 163: This sounds problematic. If the "non main" modes of variability were robustly expressed in the proxy data, then larger ensembles should be better.*

The signal of the second and third mode is quite subtle, and this signal is partly lost if we introduce strong spatial smoothing with many ensemble members. This means that the

loading of the modes will go to the first mode of variability. We will include this explanation in the revision.

9) line 191 that that

We will correct this.

10) line 193: coeval?

We will reformulate to “consistent with” as suggested by Reviewer 1.

11) line 220: which is

We will correct this.

12) line 270-271: I do not understand this

For the annual reconstruction all data is evaluated simultaneously using Eq 1. The monthly data can be extracted from the model output based on this matching of model data to proxy data. For the seasonal reconstructions the summer data is evaluated using Eq 1. and the winter data evaluated using Eq 2. Monthly model output can then be extracted for the extended summer (May-Oct) based on the matching with summer proxies and extended winter months (Nov-Apr) can be extracted based on the winter proxies. We then compare the two monthly reconstructions, one based annual matching and one separated in seasons. We will include this extended explanation in the revision.

---

references

Hörhold, M., Münch, T., Weißbach, S. et al. Modern temperatures in central–north Greenland warmest in past millennium. Nature 613, 503–507 (2023). https://doi.org/10.1038/s41586-022-05517-z

New figures for revised supplementary

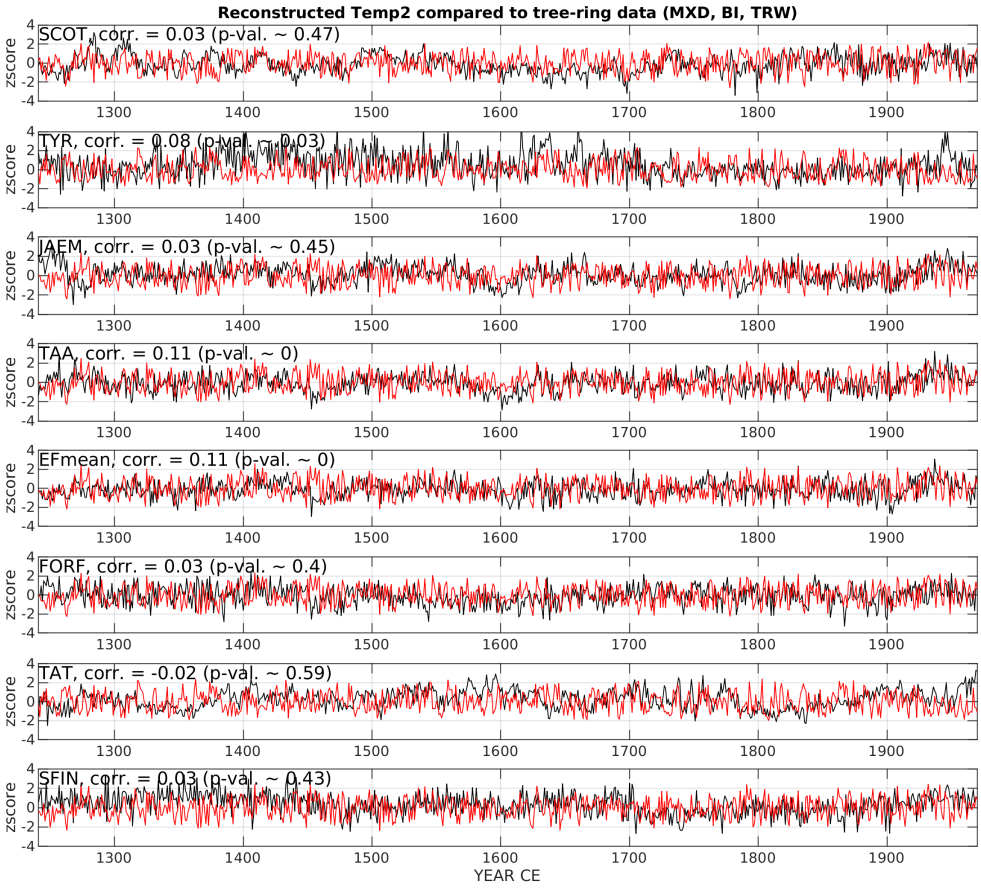

Figure R1 European tree-ring data predicted by RECON$_{IC}$.

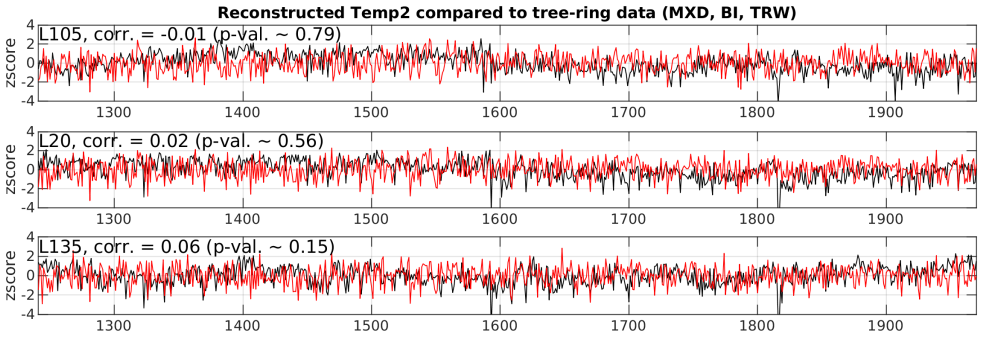

Figure R2 Canadian tree-ring data predicted by RECON$_{IC}$.

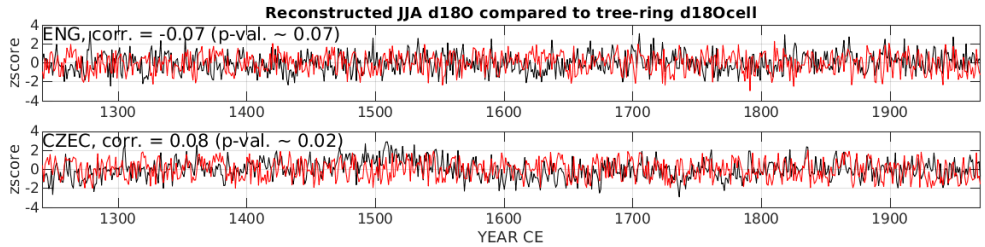

Figure R3 European d18Ocell predicted by RECON$_{IC}$ JJA d18O.

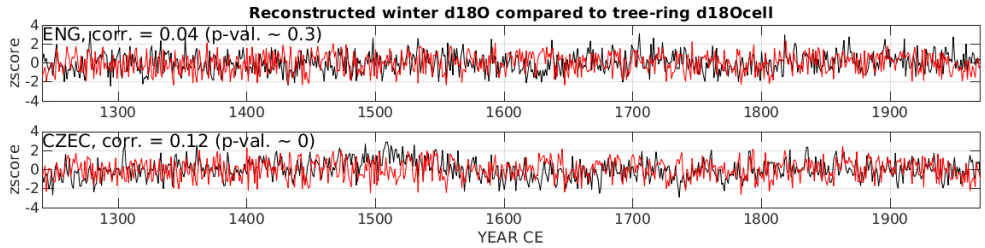

Figure R4 European d18Ocell predicted by RECON$_{IC}$ winter d18O.

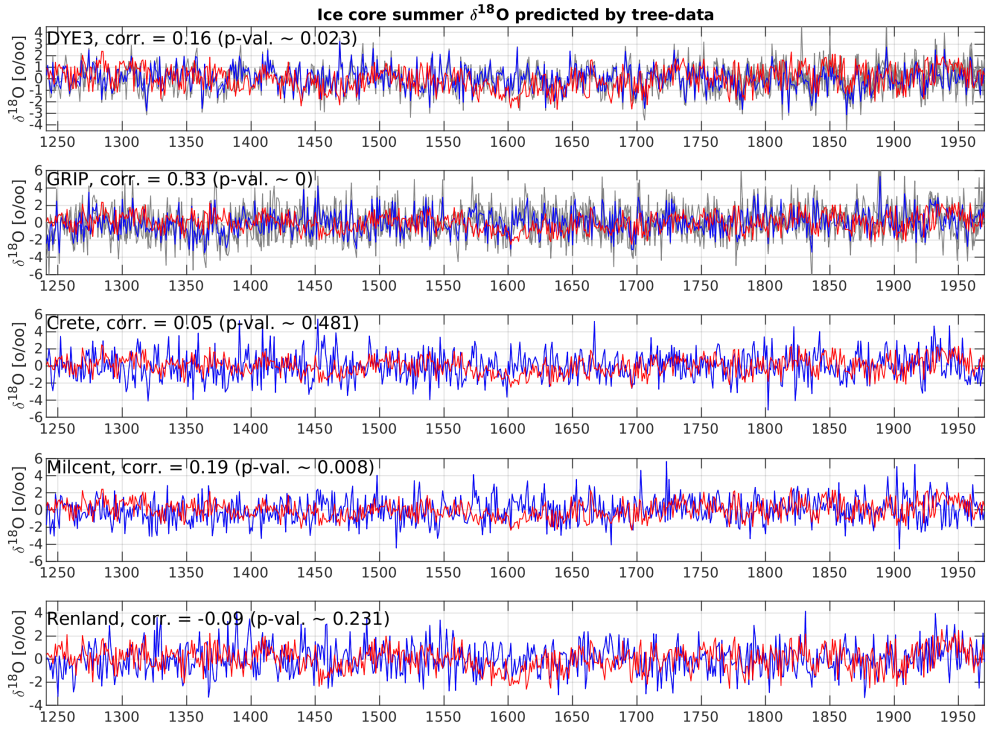

Figure R5 Greenland ice core summer d18O predicted by RECON$_{TR}$.

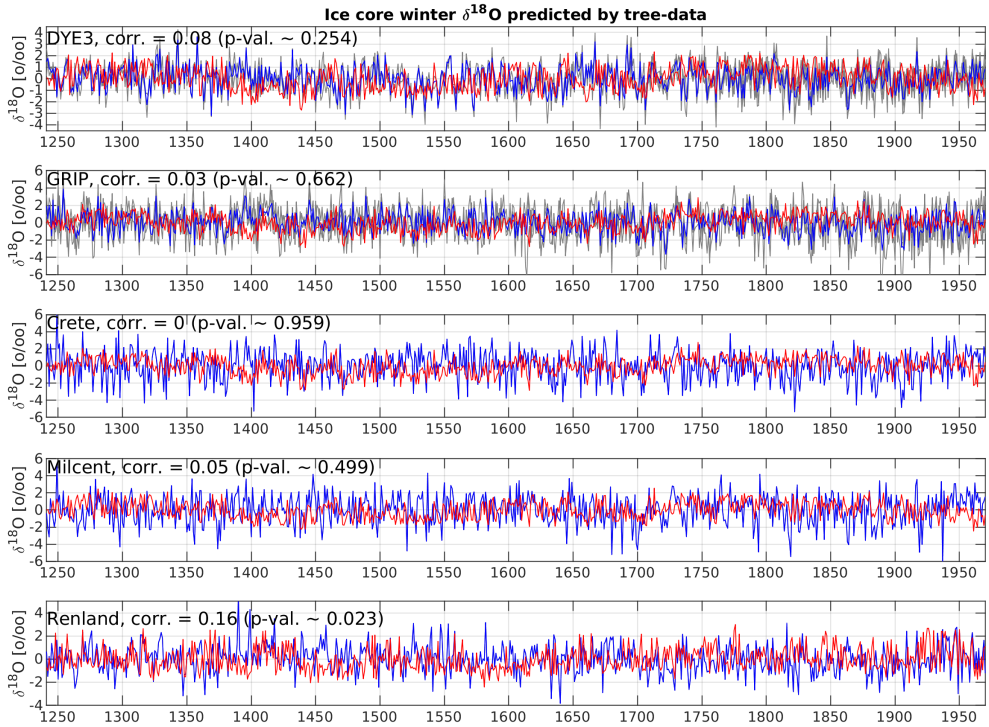

Figure R6 Greenland ice core winter d18O predicted by RECON$_{TR}$.