Response to #1:

We thank the referee for the thoughtful and constructive comments. The perspective on the definition of the NAO index relative to the reconstruction product was especially helpful, as were the points about ice core age uncertainty. We carefully considered all suggestions, carrying out additional analyses and revising several figures to ensure we could confidently incorporate the new discussion points into a revised manuscript. We describe the changes in more detail below, with the referee's main comments in bold and our responses in plain text. Line number references are given for the "tracked changes" version of the revised manuscript.

[T]he water isotopes are clearly telling us something useful about NAO variability and its impacts over the NH, but the reconstruction from them ought not be considered an unbiased expression of the NAO alone. Address the issue I have brought up regarding what they are actually reconstructing: the NAO or NAO impacts and how that might affect the frequency domain properties of their result.

The reviewer raises an important point. In fact all of the NAO reconstructions that we compare ours with are actually reconstructions of NAO-related impacts. We added discussion to the manuscript clarifying this point (L194–198). We also address this in response to a comment made by referee #2 in which they suggest extracting all the proxy sites from the CESM isotope output and comparing to the spectrum of the proxy reconstruction. This method makes a more explicit comparison between our reconstruction and the atmospheric impacts of the NAO in the model, rather than the modeled NAO index. We have updated Figure 7 in the revised manuscript to reflect these changes.

At least mention the issue of accumulated dating errors in ice cores and how that might affect the power spectrum and wavelet spectrum interpretations. I realize this cannot be quantified here, but it really should be noted as something to consider.

We agree with this suggestion and added text discussing the accumulating age errors and their impact on the spectral features of the reconstruction (L497–500). Referee #2 also raised several concerns regarding ice core age models and uncertainty which we have taken into consideration. Specifically, we have performed additional analysis focused on the impacts of ice cores with greater age uncertainty and describe the results in the response to referee #2 (L373–385; Fig. S6).

Response to #2:

We thank the referee for the useful points raised in this review. We appreciated the suggestions and took them into careful consideration, including performing some new analyses and modifying existing figures to incorporate these new discussion points into a modified manuscript. We believe the changes, which we describe below, greatly improve the manuscript. We have reproduced the referee's main comments in bold, with our responses in plain text. Line number references are given for the "tracked changes" version of the revised manuscript.

Age uncertainties:

- **I consider the time scale of speleothems, ice cores from Svalbard and Alpine sites too uncertain to be directly incorporated in this type of reconstruction without taking into account age-scale uncertainties. For example, the Svalbard age-depth scale is based on a flow model with a few volcanic horizons for reference (Isaksson et al., 2001). This should inflate the uncertainty envelope back in time considerably.**
- **For Greenland ice core time-scale pre ~1200 CE an error in time scale has been uncovered after publication and therefore in the inclusion in iso2K database. See e.g., Adolphi & Muscheler (2015). The options are to correct the raw ice data with new dating or shorten reconstruction (I too[k] the latter approach in my work).**

Thank you for these comments. We have added text to the methods (L165–178) and discussion regarding the potential impacts of age model uncertainty in the proxy data (L497–500). Our methodological approach relies on the inclusion of a diverse set of proxy types, which is one of the key strengths of large-scale data synthesis efforts. In particular, inclusion of diverse proxy systems permits the use of sites from a range of geographical areas (rather than being biased towards, for example, locations with glaciers and ice sheets). To the referee's point, we recognize that a downside of this approach is that it makes the reconstruction susceptible to diverse uncertainties associated with multiple proxy types. By design, we restricted our analysis to well-dated records with annual resolution in order to avoid the larger pitfalls of age uncertainty inherent to radiocarbon chronologies and other approaches. The inclusion of the speleothems and flow-modeled ice core records was therefore an explicit choice, and that choice is consistent with many other published studies which also include such records (e.g. Mann et al., 2007, PAGES 2013, PAGES 2019, Falster et al., 2023, etc.). So while other reconstructions may take a different approach of applying more extreme data exclusion protocols and work with a very small number of proxy records, the purpose of this paper is to use the wider-net approach, which is a new contribution to the field.

The diverse set of proxy data in the Iso2k database has the added benefit of limiting the weighting of any single record, so the uncertainty envelope of the resulting reconstruction is not particularly sensitive to the uncertainties of a given age relationship. The spectral biases inherent to a given proxy type or age model are less impactful because their influence is mediated by the information provided by other data, and the common signal can be more effectively identified (Ortega et al., 2015). That, in combination with the relatively well constrained age-depth relationships even in proxies that are not explicitly layer-counted, gives us confidence when incorporating these proxies into the reconstruction. For these reasons we feel the inclusion of speleothems and flow modeled ice core records is well justified.

The referee does raise an important point here about the sensitivity of the reconstruction to the age model uncertainty inherent to the records in question. In response to this comment, we tested that sensitivity by removing the non-layer-counted records, both one at a time and all at once, then re-computing the reconstructions, and re-performing the spectral analysis. Of the 94 records in the reconstruction, there were 8 non-layer-counted records included (Table S1). Removing any one of these does not notably change the power spectrum of the reconstruction, suggesting the uncertainty inherent in those records is not a substantial component of the low-frequency signal in the reconstruction. Removing all 8 of the records does slightly reduce the power of the multidecadal peak in the MTM such that the median of the ensemble falls below the 95% confidence level based on an AR(1) null. However, the wavelet analysis shows that the structure of the low frequency variability remains consistent with the full reconstruction throughout the millennium. It is possible that the reduced multidecadal peak is partially a function of network size, i.e. removing 8 of the records in the reconstruction.

To that end, in response to this comment as well as Referee #1, we updated Figure 6b to propagate the reconstruction ensemble uncertainty into the MTM analysis. This allows us to evaluate whether the random removal of a portion (15% each time; L190) of the dataset would change the spectral results. Effectively, propagating this uncertainty addresses the issue of whether removing 8+ records could change the results of the reconstruction. Results showed that, for the peak in multidecadal power, some of the ensemble members also fell below the AR(1) 95% confidence level. While this analysis does not change the major results and points of our paper, it does suggest that changes in the strength of multidecadal variability are partially sensitive to the size, coverage, or spatial distribution of the records included in the reconstruction, including the 86 layer-counted records with very little age uncertainty. As such, we added a supplementary figure showing the impact of removing the non-layer-counted records from the reconstruction (Fig. S6). We have also included a supplementary table in the revised manuscript highlighting the non-layer-counted records and describing their age modeling and uncertainty (Table S1).

We found the information about the Greenland ice core timescale very interesting and appreciate the referee bringing it to our attention. We investigated how this issue may impact the reconstruction, and in doing so we determined that the impact of the proposed update to the GICC05 timescale is minimal, amounting to about +/- 5 years of age uncertainty in the individual Greenland ice cores between 1000–1200 CE (Sinnl et al., 2022). We added mention of this uncertainty to the revised manuscript (L165–169). We have not seen a solution to this issue yet in other data synthesis products that use the Greenland ice cores, but we will make suggestions to the Iso2k team when the database is updated. We appreciate the referee sharing their knowledge on this topic.

Seasonality:

- **Proxy data: what is the seasonality of the proxydata used in the reconstruction?**
    - **It appears that all proxydata are assigned to calendar year? (L155). I think it doesn't make sense to use the approach of Putman & Bowen for proxy data.**
- **See publication on seasonality, circulation patterns and climate variability: Vinther et al. (2010) and Sjolte et al. (2020).**
    - **The use of the calendar year for Greenland ice core data causes loss of signal.**

The question of seasonal sensitivities is an important one and the referee is right that information is lost when integrating the few subannually-resolved records (3 out of 94 records) to an annual timescale. In our approach, we made the decision to treat all proxy records similarly, so we chose to remain agnostic to the seasonal sensitivities assigned to individual proxy records (either in their original publications or during integration into the Iso2k database, though we note such information is not always available). This approach has good precedent in the literature (e.g., Falster et al., 2023, Ortega et al., 2015, PAGES 2019). Our method does account for seasonality of the NAO itself in that it tests the correlation of each proxy record with the DJF NAO instrumental index and weights the proxy data according to the resulting coefficient. This process inherently incorporates the seasonal sensitivity of the annual data to the winter NAO, with the reconstruction product representing the aggregate isotopic response in that season. We have incorporated additional text addressing this point into the Methods section of the manuscript (L181–190). We have also adjusted the averaging of subannually-resolved records to use a simple arithmetic mean.

Regarding the seasonality of the Greenland ice cores: We appreciated this referee's expertise on this front. During the manuscript revision process we have consulted with the Iso2k ice core team to ensure that the metadata for the Greenland ice cores cited in Vinther 2010 is correct, in addition to discussing the age model concerns raised above. The Greenland ice cores use a mean annual series with the exception of GISP2, Milcent, Renland, Site A, D, G, and E, for which the DJF series from Vinther 2010 was used.

Impact of changing number of records t[h]rough time:

- **Type of proxies partly investigated (Fig S5) but not the influence of varying the number of proxy records. What is the performance of the reconstruction using only spanning the full timeframe? This would indicate what the skill is in the earlier part of the reconstruction.**
- **How does Figure 3 look like for the period prior to the validation period?**

We did in fact investigate the influence of the variable number of proxy records through time; we found that proxy record availability had minimal impact for a given validation/calibration interval (L128–137). Varying the number of proxies is also implicitly tested using the regional subsets, as well as the noise-padded regional subsets. However this comment is well taken in that additional clarification is needed, so to that end we have generated a new supplementary figure which explicitly quantifies the change in the reconstruction validation metric as the number of available records declines with time (L220–222; Fig. S2b).

We note that changing the calibration and validation windows does, in essence, influence the number of available proxy records (because not all have complete coverage during the 20th century). Based on this referee comment, we added additional text to the manuscript to discuss the selection of the calibration/validation windows (L200–208). As expected, alternative validation intervals impact the performance in accordance with the change in available data in the calibration interval. For example, calibrating on the earlier portion of the instrumental interval (e.g. 1823–1923 CE) and validating on the later portion (1824–1999 CE) causes the calibration procedure to miss large portions of the shorter records that inform the reconstruction. Therefore we do not feel it is really possible to assess pre-validation interval stats the way the referee is asking, but we agree that additional details were needed on the calibration/validation decisions and we added those to the text accordingly.

Multi-decadal to centennial variability:

The authors discuss the multi-decadal to centennial variability of other NAO reconstructions, but there is no comparison of different NAO reconstructions. Since you make use of wavelet analysis, cross-wavelet power and wavelet coherence would be appropriate.

This is a great suggestion. The spectral analysis reported in Cook et al. (2019) is directly comparable to the analysis in our manuscript, as mentioned in the discussion (L506–508). In response to this comment we have also performed MTM and wavelet analysis on the calibration- and model-constrained NAO reconstructions reported in Ortega et al. (2015). These analyses also support the conclusions we reported in the discussion (L508–510, Fig. S8).

If the focus is multi-decadal to centennial variability the time window for the validation is very short, and I supposed done on annual data with no filtering?

Yes, the validation was performed against the DJF instrumental NAO with no filtering.

- **How does Figure 3 look like for the period prior to the validation period using a decadal filter (e.g., a Gauss filter)?**

We were not certain what was being suggested here. We interpreted this question to suggest performing the reconstruction after withholding validation data, applying a decadal low-pass filter to the result and the instrumental index, and then testing the correlation between the low-pass filtered reconstruction and instrumental index in the validation interval. If that is correct, we do not believe filtering the validation data would give an accurate measure of reconstruction performance at low frequencies for two reasons. First, the degrees of freedom for the overlapping data become quite small because, as discussed in the validation period response above, the availability of validation data is limited. Second, the non-stationarity of low-frequency variability in our reconstruction is a key component of our results (Section 4.3), and although previous studies have suggested minimal low-frequency variability in the instrumental NAO index (e.g. Cook et al., 2019) we believe this may not be fully representative of the spectral characteristics of the atmospheric impacts of the NAO over the last millennium.

- **It is well-known that models underestimate multi-decadal to centennial variability, in particularly on regional scales (Laepple et al., 2023). Instead of just repeating this, you could investigate where the model underestimates the isotope variability by extracting all the proxy sites from the model output and comparing to the spectrum of the isotope records.**

This was a very interesting suggestion and we incorporated it into the revised version of the manuscript. We updated our spectral analysis of the NAO in CESM by subsampling the CESM precipitation isotope field according to Iso2k proxy locations and then performing a CPS reconstruction calibrated against the CESM NAO timeseries (L228–238). Results from this analysis show that the spectral characteristics of the resulting pseudo-reconstruction are very similar to the CESM NAO, but show slightly stronger decadal variability (Fig. 7).

Reference to previous work on the topic and claim that isotope records are underutilized: L12-13 and L102-104:

- **I think this claim is not correct. Even if we disregard my own work on the topic this ignores the use of ice core isotope records in previous work by Cook et al. (2002), Ortega et al. (2015), Michel et al. (2020) etc. some also referred to by the Flaim et al.**

- **I think part of the perceived underutilization could come from some records not being used, as people (myself included) refrain from using these records due to large dating uncertainties (see point on age uncertainties above).**

- **My work on this topic can be found in (Sjolte et al., 2011, 2018, 2020, 2023, 2025; Tao et al., 2023). I hope you will find this relevant for your study. My reconstructions are available for download. Follow links in publications.**

We appreciate the referee's work on this topic and recognize that some ice cores have been used in previous NAO reconstructions, including the referee's. We have added text and citations acknowledging these contributions (L104–107). However we disagree that this means that isotope proxy records have been extensively utilized overall. Even well-dated, layer-counted archives such as tree ring cellulose $\delta^{18}O$ have been largely left out of NAO reconstructions. The wood cellulose isotope records included in our reconstruction are very well dated and underutilized in previously published NAO reconstructions. Additionally, many of the records included in Ortega et al. (2015) and Michel et al. (2020) are speleothems and lake sediments, suggesting the age uncertainty that the referee is highlighting here is not a primary constraint on the use of isotope records in the literature. Finally, only 8 of the 94 isotope records included in our reconstruction are not annually layer counted (i.e. would fall into the category which the referee describes as having large dating uncertainties).

Of the reconstructions in Figure 3, Cook et al. (2002) uses 365 tree ring records and a small but undefined number of ice core records from only two sites in Greenland. Of the 48 records in Ortega et al. (2015), 5 were Greenland ice core $\delta^{18}O$ and one was speleothem $\delta^{18}O$. Michel et al. (2020) use annually resolved records from the Pages2k temperature database and Ortega et al. (2015), but they do not clearly document their final filtered dataset, making it hard to determine how many isotope records were actually included. Although Pages2k contains some isotope-bearing sites like the Greenland ice cores, using the Pages2k database is not equivalent to reconstructing the NAO from isotope data, as the Pages2k database by definition only includes records deemed temperature-sensitive; therefore overlap with our isotope-based reconstruction is limited.