We thank the reviewer for their time and thoughtful re-evaluation of our manuscript. We carefully considered the recommendations made in the second round of review, carried out additional analysis, added several main text figures, and expanded the text to address the reviewer's comments. We describe the changes in detail below, with the reviewer comments reproduced in plain text and our responses in bold. Line number references are given for the "accepted changes" version of the revised manuscript.

Referee #2: Report

This a review of the revision of the study of Flaim et al., following up on my first review. To me there are still several issues that remain unclear after the revision, and I think my comments from the first round of reviews are not sufficiently taken into account. Now that I read the paper again, I think part of the problem is that the description of the goal of the paper and science questions do not match the main focus of the paper. The paper is presented as a new NAO reconstruction "Here we reconstruct the NAO index for the last millennium (1000–2000 CE) …" (L108), in which case more care should be put on the details that I commented on in my first review (consistent data handling, uncertainty envelope, validation, comparison to existing NAO records). However, the actual focus of the study is on exploring the climate proxy signal in isotope records and a model simulation, and the potential for using isotope records for climate pattern reconstructions such as for the NAO.

Since there are still unresolved issues, I see two ways to bring the paper closer to being publishable.

1) Reframe as a method paper, rather than a new reconstruction. I believed this to be the most straight forward solution to make progress on this paper. My main suggestion would therefore be to reformulate the last paragraph of the introduction (L108-118) along with a clear goal and science questions that matches better what is done in the study.

Suggestion for new framing of the study: "Here we investigate the multi-decadal variability in reconstructed NAO and its dependency on selection of proxy data from a collection of isotope-based climate archives." For this option I also think a different title would be appropriate.

The reviewer's framing represents the science questions being considered in this manuscript well, though we believe the existing text was largely consistent with these goals. That said, we have revised the introduction (L108–109) as requested and made some changes to the abstract (L13–16).

We thank the reviewer for this suggestion and believe it is complementary to the previous revisions that were intended to make clear that the goal of the reconstruction is to examine multidecadal variability (see, for example, L198–201). The existing title is still consistent with this framing, especially given the additional multidecadal analysis added based on reviewer comments discussed below, so no changes were made to the title.

2) If indeed this should be viewed as a new reconstruction, a more extensive part on validation of the decadal to multidecadal variability in reconstructed NAO is needed, including a more complete comparison with existing NAO reconstructions. Presently, there is no direct comparison of the multidecadal variability with existing NAO records, or even with observed NAO. I asked for this in the first review, and although I see you have added a wavelet analysis

of the NAO reconstruction by Ortega et al., I think this is not sufficient. This is not something to be buried in the supplementary.

I know the overlap with observations is relatively short, but you can start by simply showing time series of the decadally filtered data of observations and reconstructions. There are many methods to test for significance when correlating records with strong auto-correlation, such as filtered data, for example, Ebizusaki (1997). Then, for example, do a wavelet coherence and cross wavelet power analysis with observed and reconstructed records.

We thank the reviewer for their detailed recommendations. We have made substantial revisions to address the request for additional comparisons of multidecadal variability in our reconstruction to other reconstructions and to the instrumental NAO index. We added two main text figures, the first of which (now Figure 7) compares the decadally smoothed Iso2k NAO, the decadally filtered Cook et al. (2019) NAO, and the decadally smoothed Ortega et al. (2015) NAO with the decadally smoothed instrumental Vinther et al. (2010) NAO index. We show the time series as requested and perform correlation significance testing using the Ebizusaki (1997) method. We added another main text figure (Figure 8) with wavelet and MTM analysis for the non-smoothed Cook et al. (2019) reconstruction, as well as the non-smoothed Ortega et al. (2015) calibration constrained reconstruction, which we moved to the main text from the supplement. We have added methods text (L244–256), an additional results section (Section 3.5; L425–458), and discussion text (L575–589) to describe these results and their relation to NAO signals in Iso2k.

We performed wavelet coherence and cross wavelet power analysis between the reconstructions over the observational interval as well as the last millennium. Additional methods text was added to describe the analysis (L257–266). We explain the outcomes in the results (L436–440; L454–458) as well as the discussion (L581–589). This analysis supported the individual wavelet results but did not materially change our main discussion points or conclusions. Figures detailing the wavelet coherence and cross wavelet power analysis can be found in the supplement (new Supplementary Figure S7 and Figure S8). We agree that this more thorough treatment of multidecadal variability in the other reconstructions is beneficial to the paper.

In any case, I think there are some inconsistencies in the handling of proxy data which are important to mention.

Dating uncertainties.

The results of Sinnl et al. (2022), does not show an uncertainty of +/- 5 years. We know that the time-scale is wrong by +7 years at CE 775, and that this propagates through the records until about CE 1200. The shift in GICC05 at CE 1000 is +5 years. So, if you are looking at 10-year cycles at CE 1000 the ice core data are shifted by 1/2 of a cycle. I don't think you can necessarily see this shift easily in the power spectrum of the reconstruction, but since we already know the data is shifted, we also know that the reconstruction is not going to be great for that time period.

This uncertainty has been acknowledged in the manuscript and was considered when interpreting the results. Importantly, as the reviewer comment also highlights, there is no clear indication that this chronological uncertainty affects the results or conclusions

presented here. While a shift of several years could alter the precise phasing of decadal-scale variability, we do not observe that such an offset significantly influences the reconstructed variability or the broader interpretations drawn from it. In particular, we are careful to focus our interpretations on spectral characteristics and multi-decadal patterns and we do not make direct comparisons at the year-to-year scale between the reconstructions. We believe that our approach is the most appropriately cautious one to take with a climate phenomenon like the NAO that inherently contains such strong interannual variability. Therefore, although the timescale uncertainty is appreciated, it does not introduce discernible impacts on the analyses or conclusions of the study.

Seasonality of ice core data: possible error in Iso2K entry for Vinter et al. 2010.

From the author reply: “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.” There is no DJF data from Vinther et al. 2010, which use Nov-Apr to represent winter. I double checked the Iso2k meta data and it appears not to be correct, and the data entry appears incomplete. The Vinther et al. data has 20 cores in seasonal and annual resolution. I would have expected all of this data to be part of Iso2K.

We thank the reviewer for the clarification. They are correct that Vinther et al. 2010 estimates NDJFMA as the seasonal range for winter in the cores listed. While the mislabeled metadata fields in Iso2k were flagged during the first round of reviews (as noted in our response), the changes have not yet been applied to the publicly available version. Also as highlighted in our previous response, the approach used in our manuscript is agnostic to the seasonality of individual records. The Iso2k metadata has no impact on the results or conclusions presented in this manuscript.

The DYE-3 d18O data can be layer counted 8000 years back, and is available in seasonal resolution back to CE 551 from the Vinther et al. data reference, not ending CE 1778 (minimum year according to Iso2k). Due to these inconsistences I strongly advise to use the data from the original publication
(http://www.iceandclimate.nbi.ku.dk/data/Vinther_etal_2010_data_02feb2010.xls).

In addition, I think that the GISP2 data has somewhat diverging variability, and is not as coherent with the rest of the Greenland ice core data as you would expect from a central Greenland record. I have excluded GISP2 data from my own work as it seems to degrade the quality of the reconstruction.

We appreciate the reviewer’s attention to detail on this issue. We have recommended changes to the Iso2k database to address the missing DYE-3 data, as well as several other metadata issues relating to the Vinther et al. (2010) data. With respect to the NAO reconstruction presented here, we have added the additional DYE-3 data to the proxy dataset and updated the figures accordingly. The additional data changes the reconstruction slightly, and particularly improves the validation tests using only long proxy records. Ultimately the major results and conclusions of the paper remain unchanged. The GISP2 data has sufficient coverage and resolution to meet our filtering conditions, so we include it. Additionally, the GISP2 data is significantly correlated with the instrumental Vinther et al. (2010) NAO index.

Influence of changing number of records over time.

I think the new results in Figure S2 are very important and should be included in the main paper. The results are hardly mentioned in the main text and I think it deserves more space, as it highlights the importance of data availability to maintain a consistent reconstruction. It is a wide-spread problem with reconstructions that the data is validated for a time frame where we have lots of proxy data, but is the skill disappears when going back more than a couple of hundred years.

It is a bit confusing with the “Only 1800-yr records …” label. As I understand this means you are using all records spanning CE 1800 to present? In which case it shows very clearly the loss of skill when using fewer records. I am surprised to see the skill so low when only using data that span all, or most, of the reconstruction interval, and I would expect an improvement if you use the complete Vinther et al. data like I mention above. A correlation of ~0.5 with observed NAO is similar to what other reconstructions have obtained (e.g., Ortega et al., 2015; Sjolte et al., 2018, 2020), and I think that level of skill should be obtainable with relatively few records.

We have moved the key results from Figure S2 to the main text, adding them to Figure 3. Adding the DYE-3 data as mentioned above has improved the validation scores of the versions of the reconstruction that only include the longest proxy records. The reviewer’s interpretation is correct that the validation experiments were performed by using all records with data spanning 1500 CE to present and again using only records with data spanning 1000 CE to present. We have updated the figure axis labels to clarify the methodology. We have also added additional methods (L212–215) and results text (L311–321) to emphasize the limitations imposed by data availability. We would like to highlight that the distributions represent the range of validation scores produced after randomly removing a portion of the available proxy data, which has a greater impact on the reconstructions performed with fewer records. The maximum validation scores for reconstructions using only records spanning 1000 CE to present reached as high as r=0.4.