Response to reviewers (*Deficient ocean–atmosphere feedbacks constrain seasonal NAO prediction* by Erik W. Kolstad)

First, I apologise for the multiple errors on my part, which introduced challenges for the editors and the reviewers. I would like to thank all for their flexibility and both reviewers for their thoughtful comments. In fact, the suggestions were so useful that they inspired major changes to the manuscript, which I think is vastly improved in its new version. I know that it can be irritating to have to review major changes that were not asked for, but I hope the improvements justify these changes.

I'll respond to each comment in turn below (original comments in italics and my responses in upright font), but I also summarise here the main changes with respect to the first submission.

The analysis now uses November initialisations, following the suggestion from Reviewer #1. This change makes it possible to relate the mediated effect directly to the model's NAO prediction skill, rather than to the SST–NAO correlation used in the previous version. As a result, the "skill sections" (§4.2 and §4.6) have been completely revised. The new study is more relevant, as it is based on the actual anomaly correlation between the SEAS5 NAO and ERA5 NAO. Additionally, the method does not require any subjective choice of reference regions for heat flux and baroclinicity.

I have taken Reviewer #2's reservations into account by expanding the analysis regarding the "reverse" pathway, i.e. the NAO influencing the mediators. In the new version, I have replaced partial correlations with the new Eq. 5:

$$Z = \alpha' X + \gamma Y.$$

This way of regressing out Y (the DJF NAO) is more consistent with the other regression equations. It also allows for a comparison between γ and β. If the former is greater than the latter, then the NAO -> mediator pathway dominates over the SST -> mediator one.

I now use the SST reference region used in Czaja and Frankignoul (2002) and Kolstad and O'Reilly (2024). This choice isolates the influence of extratropical SST anomalies from that of tropical SSTs, which is consistent with the physical mechanisms examined in the paper. The change did not materially alter the results.

Response to Reviewer #1

Major points

A) Use of October hindcasts

I remain unconvinced about the use of October hindcasts. November starts are normally used for DJF forecasts so why not use the forecasts that are relevant to the problem? We should at least be reassured that November forecasts show similar, if perhaps weaker errors.

The motivation for using the October initialisations was to ensure sufficient variation in the November SST states. In the November runs, the SST fields are nearly identical across ensemble members due to oceanic inertia. This is reflected in the high correlation between the November SST index between SEAS5 and ERA5 (r = 0.91). I first thought this dependence would reduce the usefulness of the data for mediation analysis since fewer SST

states would be included in the samples. However, having looked closer at the differences between the results based on October and November runs, I now see that the differences are marginal. I agree that November forecasts are operationally the most relevant for DJF predictions, so I have decided to base the analysis on the November initialisations as you suggest. This change also led me to derive a more relevant metric for the model's skill in predicting the NAO.

B) Use of ensemble means

The analysis is novel, relevant and interesting but there is a serious flaw. The analysis is carried out entirely on ensemble means (L186) and then compared to the observations (L325, L360 and throughout). This comparison is not valid. A simple example can illustrate why: assume for example that the NAO is entirely formed from unpredictable variability or 'noise'. In this case there would be no ensemble mean signal and no regressions between the modelled variables. However, the observational analysis will still show relationships, albeit from unpredictable 'noise'. In reality the difference will be less extreme as the NAO contains predictable and unpredictable components but the presented analysis would only be valid if the NAO is formed from entirely predictable variability. Fortunately, the problem is easily corrected as it simply needs to be redone on ensemble members. I hope this can be done as I still think this has the potential to be a very useful contribution but it is essential before publication.

I fully agree with the reviewer that comparing ensemble-mean relationships to observational relationships is not appropriate, since the ensemble mean largely removes the internal variability that drives much of the NAO. The reviewer's example illustrates this well: if a large part of the NAO arises from unpredictable variability, then regressions based on the ensemble mean will artificially suppress physically meaningful relationships that are present in the observations. This was an oversight on my part.

Following the recommendation, the analyses is now based on the full set of ensemble members. This necessitated a modification of the bootstrapping strategy for SEAS5. Using all ensemble members increases the effective sample size by a factor of 25 relative to ERA5, which causes almost every relationship to appear statistically significant. To allow a fair comparison between the datasets, I now bootstrap SEAS5 by drawing samples of length 43 (matching the number of winters in ERA5) rather than 43×25. This yields a significance assessment that is directly comparable across the two datasets.

Minor points

The article seems to be overly positive about empirical forecast methods. Several of the examples cited have not performed well after publication in real out of sample cases. This is often the case with such methods which have often been inadvertently tuned to non-causal relationships in sections of the past observational record. Please therefore refine the language to better represent this, for example by saying "...achieved potentially useful levels of skill (but note the comments below about real time forecast skill)..." and at L34: "often appear to outperform" as this is not really outperforming if based on noncausal factors.

This is a fair point. I have now updated the paragraph starting with "Owing to", and I also added a new sentence in the next paragraph ("In sum, empirical models"). I hope this better reflects the problematic nature of empirical models.

L45: Suggest "high surface NAO" as some studies claim NAO skill from high level circulation fields that is not reflected in surface NAO predictions

Changed to "surface-defined".

L46: Baker et al 2024 reported similar levels of skill for the NAO from later generations of forecasts and similar ranking of systems so a better phrasing here would be "However, there is a wide range of performance between systems and system upgrades have not significantly improved overall skill". Please also remove comments about reducing skill as the reported changes are not significant.

Corrected.

L110: typo 'aa'

Corrected.

L138: I did not understand why this implies 'many pathways'

I understand why this could be confusing. I didn't mean that there has to be many pathways, only that the one through Z is one of potentially many pathways. The updated paragraph clarifies this, I hope.

Sec3.1: why is this particular system (ECMWF SEAS5) used? Is it because it has lower skill than some of the others (c.f. Sec 4.1) and so useful to detect errors? If so please say this.

No, it was because it has a long hindcast period. This was already stated in the Discussion, but I have now moved this to §3.1 ("The reason only one model…").

L201–205: How are anomalies calculated in SEAS5 and ERA5?

I have now added a sentence at the end of §3.1 to explain this.

P7 line 1: This seems odd as there are only N values to start with so by definition there are many repeates and samples are not independent. This will reduce spread and affect results like those in Fig.5. Is there a simple inflation of spread that can be done to correct and compensate for this?

I assume you mean §3.3 (on the bootstrapping). I may have misunderstood what you meant. As far as I know, the procedure described follows the standard nonparametric bootstrapping approach. Drawing N samples with replacement from an N-point series is the conventional way to approximate the sampling distribution of a statistic. The fact that individual data points are repeated across resamples is inherent to the method and does not bias the resulting confidence intervals. No additional "inflation" of the spread is required; the uncertainty is estimated directly from the variability across the bootstrapped samples. In any case, §3.3 has changed now that the full SEAS5 sample is used.

L268: what is the mean bias in the NAO?

Good question. As I defined the NAO as the leading EOF of winter SLP, there is no unique scalar "mean NAO bias". To provide a quantitative measure, I estimated a NAO-like mean-state bias from the DJF mean SLP difference between representative centres of action

(Stykkishólmur, Iceland, and Ponta Delgada, Azores). The SEAS5–ERA5 bias is +1.0 hPa in Iceland and –0.4 hPa in the Azores, giving a north–south difference of +1.4 hPa, demonstrating that SEAS5 underestimates the climatological NAO-like dipole, with too weak westerlies across the subpolar North Atlantic. This clarification has been added to the new version (in §4.3).

L290: typo 'gyrefor'

Fixed.

L297: please state if this represents a positive feedback

This sentence has been deleted.

L384: robust

I'm not completely sure what you meant here, was it to replace "coherent" with "robust"? I made no changes.

L394, L405: grammar at the start of these sentences, please reword

Both changed.

Thanks again for your insightful comments.

Response to Reviewer #2

This paper uses a statistical casual framework called mediation analysis to diagnose atmosphere-ocean feedbacks associated with NAO predictability and to compare them between ERA5 and the seasonal prediction system SEAS. It finds that surface heat fluxes and Eady growth rate mediate the effect of November SST on DJF NAO, which the paper refers to as an "indirect effect". These indirect effects are found to be substantially weaker in SEAS than ERA5.

Thanks for this summary. Just a quick comment on the indirect effect: this is standard notation in mediation analysis, but I detected a slight disapproval in your comment – I agree that it's not the best term, so I've now consistently changed it to "mediated effect", which is also standard but works better.

I find this to be an interesting and well written paper, albeit with some potentially important interpretation issues in its current form. My chief concerns are that the autocorrelation of the NAO is not considered, that much of the results hinge on the potentially coincidental correlation between the DJF NAO and the particular November SST pattern studied, and that nothing about the analysis demonstrates a causal role of subpolar heat fluxes in the SST-to-NAO feedback. I foresee that these issues could be addressed with some additional analyses and more careful wording, which could be addressed in a round of major revisions.

As described in more detail below, I've expanded the directionality analysis, and I've also tried to use more careful wording (although the new results demonstrate causal roles of both the fluxes and the baroclinicity).

Please note that this review is based on an updated version of the manuscript, provided by the editor, in which the data handling error was corrected and where individual ensemble members were used instead of ensemble means.

Major Comments:

1. When the effects X -> Z -> Y are discussed, X -> Y -> Z is considered as an alternative, which motivates Figs. 3b, 3e, 4b, 4e. This is all fine and well, but the paper does not consider the alternative that Y -> X and Y -> Z. That Y -> X may seem a bit silly when Y is DJF NAO and X is November SST, but not necessarily if the autocorrelation of the NAO is considered. To rule this out, I think it is necessary (1) to consider the correlation between November NAO and DJF NAO and/or (2) to investigate the sub-seasonality of the X -> Y relationship. I can see in Kolstad and O'Reilly that this effect is largest in February, which certainly helps to address this concern, but I still more discussion of this is needed in this paper.

Thank you for this thoughtful comment. The new Sections 2.2 and 4.5 now address the reverse pathway $X \rightarrow Y \rightarrow Z$ in more detail, using a regression framework that is consistent with the other diagnostics in the paper.

In §2.2, the pathway $Y \rightarrow X$ is now addressed by accounting for the November NAO index explicitly through the coefficient γ_0. However, this parameter was negligible for both mediators. In other words, the $Y \rightarrow X$ pathway can be ruled out.

A related concern is that the mediation effect αβ could simply be the coincidental agreement between α and β. This is easy to imagine, because β is strong (everything co-varies with the NAO, albeit usually with NAO as the causal driver). Then any α that are large by coincidence (and of the same sign) will show as a strong indirect effect. My concern is that the correlation between the NAO and the November SST pattern is at least partially coincidental (combined with choices made to maximize this correlation, as discussed in the text). Then the comparison SEAS analysis is at a disadvantage in terms of correlations (compared to ERA5) throughout the rest of the analysis, because the ERA5 SST pattern was chosen, rather than choosing whatever November SST pattern would maximize the correlation in SEAS. To address this, I think at minimum requires repeating the SEAS analysis with the November SST pattern most correlated with the DJF NAO within the model.

Thank you for raising this important point. I agree that using the SST pattern most strongly correlated with the NAO within the model can be informative in studies focused on internal model consistency. However, the primary aim of the present paper is different: it is to evaluate whether SEAS5 reproduces the observed physical pathway linking November SST anomalies to the winter NAO, as documented in ERA5.

Using the ERA5 SST and NAO spatial patterns for both datasets was therefore a deliberate choice. It ensures that the analysis assesses the model's response to real-world SST variability rather than the pattern that happens to maximise NAO co-variability inside the model's own mean state. If a model-specific SST loading pattern were used, the SST–NAO

correlation in SEAS5 would, as the reviewer notes, increase by construction. This would test internal coherence rather than the realism of the feedback mechanisms under study.

Your point that "any α that are large by coincidence (and of the same sign) will show as a strong indirect effect" is addressed in the reply to your next comment.

Even after the above two statistical issues are addressed, I still don't think it's possible to fully conclude that DJF subpolar heat fluxes play a causal role in mediating the relationship between November SSTs and the DJF NAO. The reason is that heat fluxes in this region are strongly correlated with the NAO, where this primarily represents the reverse causality (i.e., NAO -> subpolar heat fluxes). This means that ANY other causal pathway that relates the November SST pattern to the DJF NAO will also show up as a strong mediation effect in the heat fluxes. I think this requires much more careful discussion throughout the manuscript as well as softening of the conclusions relating the role of subpolar heat fluxes. This all of course applies to Eady growth rate as well, but there the underlying physical explanations in the manuscript make more sense.

Thank you for this thoughtful and important comment. I agree that, in the original version of the manuscript, the directionality issue was not treated with sufficient depth. As the reviewer notes, the strong contemporaneous NAO–flux and NAO–baroclinicity relationships mean that a large mediated effect does not by itself prove that the mediator is forced by the SST anomalies rather than reflecting NAO-driven feedbacks or another pathway entirely. In the revised manuscript this concern is addressed in three substantial ways:

As mentioned, the new sections (§2.2 and §4.5) introduce and evaluate the coefficient α', which isolates the NAO-independent influence of November SST anomalies on the mediators. This term allows an assessment of how much of the mediated effect originates from SST forcing alone. To further quantify the balance between forcing and feedback, the revised manuscript now includes the ratio γ/β, which measures whether variability in the mediator affects the NAO more strongly than the NAO affects the mediator. This index is used to identify exactly where the indirect effect could be inflated by NAO-driven covariance. In ERA5, the regions of strong mediated effect do not show dominance of the $Y \rightarrow Z$ pathway. In SEAS5, the lack of such dominance is primarily due to the weak SST–NAO link.

Throughout the Results and Discussion, I have sought to soften or reframe the conclusions regarding mediation. The manuscript no longer claims that fluxes (or baroclinicity) causally force the NAO in a unidirectional way. In fact, the revised manuscript now treats the reviewer's concern as a central interpretative issue. I hope the reviewer is satisfied by these amendments.

References to the subpolar gyre, or in some cases even just "the Gyre" are vague. When discussing the heat flux biases and heat flux mediation plots, the subpolar gyre was referring to a small northern part of the Gyre near Iceland (most egregiously on L. 273). Then the references to the gyre were to a location further south when the Eady growth rate results were discussed. Additionally, there are unlabeled emphasis boxes on several figures, which are not in the same place across figures. References to the subpolar gyre should be made with a map of the subpolar gyre streamfunction in mind, and the paper needs clearer labeling of which regions are being referred to where (especially the averaging regions used in Figure 5).

Thanks for pointing this out. I do not use fixed reference regions anymore and have minimised references to the Subpolar Gyre.

While the term "suppression" makes sense, and it seems that "inconsistent mediation" comes from the literature, I think the term "correct mediation" is misleading, because it's also possible for there to be other effects (i.e., in completely other problems) where the correct effect (i.e., in reality) is one of inconsistent mediation or suppression. Is there another term that could be used for this? This terminology sometimes comes off as overconfident about the true direction of the various effects in reality (e.g., on lines 373-375).

I agree and have now removed all references to inconsistent mediation. Additionally, I have tried to provide clearer interpretations of suppression.

Line Comments:

Figure 1: The figure caption should specify that the SEAS values are from the ensemble mean

This is done in the new version.

72-79: After the preceding discussion about S2N paradox, I thought this paragraph could benefit from distinguishing between what has been found based on observations and what has been found based on models.

That's a very good suggestion. I've added a paragraph starting with "A useful distinction…". I hope this aligns with your intentions.

102-106: Of these 3 extensions, it seems to me like the first has already been done by Kolstad and O'Reilly (2024) and deserves less emphasis

I agree. The reason I wrote this was to highlight that I provide a more thorough introduction to the methodology. The text has been amended to focus on the two other extensions, plus a new extension which involves relating the mediated effect directly to NAO skill.

218: "appropriate latitude-based weighting" is not specific enough, because there are two common choices, one where the data is cos(lat) weighted and one where the data is sqrt(cos(lat)) weighted such that the covariance matrix is area weighted. See discussion at https://climatedataguide.ucar.edu/climate-tools/empirical-orthogonal-function-eof-analysis-and-rotated-eof-analysis and in Baldwin et al. 2009 (https://doi.org/10.1175/2008JCLI2147.1).

You're right. I used sqrt(cos(lat))-weighting and have specified this now.

229: Just checking, this is still the SST field and not the surface temperature field, right? Surface temperature has sea-ice surface temperature, which can be much colder than the freezing point, whereas SST should be no less than the freezing point (approx. -4°C).

Yes, it's the SST field.

302: "Its close resemblance to the indirect-effect pattern in the Subpolar Gyre underscores the feedback nature of this coupling" – It's not a feedback on the NAO, because the NAO is forcing the heat fluxes, not the other way around. So maybe the heat fluxes of the same sign can be said to

reinforce the heat fluxes, but this wording seems to be implying a feedback on the NAO, which cannot be diagnosed from the sign of the heat fluxes alone.

That's right, this couldn't be diagnosed. Now this text has been deleted as a consequence of the new method used.

Figure 2: Caption is for 6 panels instead of 8, 2 missing.

Correct. Fixed in the new version.

268: High west of Gibraltar = Azores High

Thanks. Fixed.

Figure 3, 4: It's important to note somewhere that the boxes added for emphasis are not in the same place across figures

I agree, but now the boxes are no longer shown.

322: "broadly negative" is not correct. Approximately just as much positive as negative.

This sentence has been deleted.

332: "strengthens horizontal temperature gradients and reduces lower-tropospheric stability" – have you checked this, or is this an inference?

Also deleted.

339: "unmistakable" is a bit strong. They look similar in pattern, but different in amplitude. Keep in mind that there hasn't been any statistical test of the amplitude difference, just the sign

I agree that this was a bit too strong. The sentence has been deleted.

368: typo in p-value?

No, I meant that it was much less than 0.05 but have changed this now.

416-420: A larger feedback of the ocean is not necessarily all about ocean resolution. See for example Czaja et al. 2019 (https://doi.org/10.1007/s40641-019-00148-5) and Wills et al. 2024 (https://doi.org/10.1029/2023MS004123) on the role of atmospheric resolution

Yes, thanks for pointing me to these articles. I have cited both now.

I thank you again for taking the time to review the paper.