# GENERAL RESPONSE

We thank all three reviewers for their detailed and extensive feedback, along with pointers to several highly relevant references which we had overlooked. The paper has been substantially revised based on your suggestions.

Before outlining our main responses, we need to point out that last November the authors were able to obtain additional supercomputing units and have since used these to double the ensemble size from 3 to 6. The new ensemble members have proven to be consistent with the original members, which adds considerable confidence to the hypothesis that the stochastic schemes are genuinely improving the teleconnection. All Figures and diagnostics in the revised version of the paper have therefore been expanded to include these members. As a result of this, some minor details of the discussion have changed. We hope this will not cause a nuisance to the reviewers, who might rightfully wish we had waited with submitting in the first place until this larger ensemble was obtained. Unfortunately, it was not clear prior to submission if the required computing units would be obtainable.

Besides the increased ensemble size, the other main change is related to the interpretation of the LIM analysis. All reviewers had asked about aspects of this, and when trying to address this we became aware that we had been thinking about things wrong. In brief, the failure of CTRL to have a teleconnection is *not* accounted for by the variations in the LIM coefficients we computed. These variations affect the magnitude of the signal but not the correlation, which mostly depends on the persistent forcing exerted by the long-lived sea ice anomalies. Rather, it is that the LIM hypothesis simply fails to be valid for CTRL. This suggests the CTRL model is failing to satisfy one or more of the hypotheses of the LIM, pointing in particular to one or more of the following: 1) the crucial role of adjustments to ice and SSTs in regions remote to the Barents/Barents-Kara; 2) disruptions to the initial ice anomaly in CTRL from external atmospheric forcing, such as from ENSO; 3) non-linear impacts not captured by our analysis. No clear evidence is found that the CTRL model does a poor job at generating a realistic initial heatflux response to sea ice anomalies, but some hints are found that adjustments to the ice from heatflux forcing may be bad in CTRL (and better in OCE), which would relate to point 1). We also speculate concretely that adjustments around Labrador and Greenland may be important. All of this is now discussed extensively in the revised Section 5.

Additionally, in response to Reviewer 1, we have included sensitivity tests of our results to the choice of sea ice region. Qualitatively similar results are found if using Barents-Kara for all data sets, but the OCE correlations become slightly smaller in magnitude.

We now address each reviewer in turn.

# REVIEWER #1

**Major comments:**

**RC1:** *"The use of different sea ice regions for the model and observations is problematic. The authors have correlated the NAO with sea ice concentration at all gridpoints and cherry-picked the regions with the largest correlations (which is different in the model and observations). Given the weak correlations combined with large internal variability, there is a good chance the internal variability is contributing to the regions with the highest correlations. This means all the subsequent analysis and discussion about statistical significance is not reliable because the region was not selected a priori. The authors should use the Barents-Kara (BK) Sea for both observations and model correlations. I don't even think this will have that large of an effect on the analysis and conclusions because there are clearly differences in correlations over just the BK Sea (Figure 4).*

*The justification for this is not at all convincing. The authors claim that because models have different biases, the regions with the most sea ice variability is different across different models and the real world. However, The sea ice in the BK Sea in the OCE does not look that different than in ERA5, so I don't see why they cannot use the same region. The leading EOF in ERA5 looks very similar around the BK region (Figure 2). I can see maybe shifting the regions slightly to account for biases (e.g. if the model ice edge is 1° too far south in the model, shift the region definition 1° to the south), but to use a very different region is not justifiable and introduces additional issues.*"

**Response:** Concerning the choice of sea ice region, we agree that a differing choice for the model and observations leaves us open to accusations of cherry-picking, and at the very least some discussion of sensitivity of results to the choice should have been included. We have now done more extensive testing of the use of different regions and can report the following. If one uses Barents-Kara for all data sets, then the conclusions are qualitatively similar, in that there is a consistent improvement of the ice-NAO correlations when adding stochasticity, and these improvements can be explained using the LIM model. However, quantitatively speaking the results are somewhat weaker, with the correlations in OCE being generally smaller (and not as comparable in magnitude to ERA5) when using Barents-Kara as opposed to Barents-Greenland. We also found that using just the Barents sea for the model gave quantitatively almost identical results to using Barents-Greenland, and the increased ensemble size now singles out the Barents sea anyway (revised Figure 4). On the other hand, the Barents November sea ice in ERA5 has zero correlation with the NAO: it is definitely necessary to extend the region out to the Kara sea for ERA5.

After careful consideration, we believe it is still justifiable to somewhat adjust the sea ice region in the model compared to observations. The results discussed above have led us to use Barents-Kara for ERA5 and Barents for EC-Earth. The difference between the two regions is therefore even smaller now, with EC-Earth simply omitting the Kara sea, where EC-Earth3 has clear biases. An equivalent table to Table 1 which uses Barents-Kara for all data sets will be included in Supporting Information of the revised paper, and we will clearly highlight and discuss the fact that qualitatively (but not quantitatively) similar results are obtained with this uniform choice (new Section 4.3). We hope this will go a long way towards

addressing the reviewer's objections.

We now expand on our justification. There are two key points. The first is that both the mean state and the seasonal evolution of the sea ice edge is clearly different in CTRL compared to ERA5.It's true that the bias of CTRL and OCE in the mean sea ice in the Kara sea (Figure 1a,b) is on the order of 10% less ice than in ERA5, and this not huge on the face of it. But the biases in the standard deviation (Figure 1c) clearly point to a big change in how far equatorward the ice edge tends to extend to every year: the sign of the pattern (negative near pole, red equatorwards) says that in CTRL, the ice edge tends to extend further outwards. This is important because the heatflux anomalies are dominated by the variations in the location of the ice edge: if the ice edge has moved, so will the largest heatflux anomalies. The 10% difference in the mean state is therefore in all likelihood misleadingly small, smoothing out more important interannual variations in the ice edge in the Kara sea. This change in the seasonal ice edge evolution in EC-Earth3 is further corroborated by the visibly different EOFs (Figure 2). It is true as the reviewer states that the *local* magnitude of the patterns in the Barents-Kara region are similar between ERA5 and OCE, but clear visible differences still remain. In ERA5, the typical November pattern is evidently an increase (decrease) of ice in Barents-Kara and a decrease (increase) in the Barents sea closer to Russia as well as in the Laptev sea. In OCE, the typical behaviour is an increase/decrease along the entire ice edge from Greenland up to Bering. In particular, sea ice anomalies in Barents-Kara may, in the model world, be expected to often come hand-in-hand with sea ice anomalies elsewhere that don't look anything like that of observation. Since it has been noted in previous papers ([1,2] and others that the reviewer themselves provide) that sea ice anomalies in regions other than Barents-Kara may have different, even opposing, impacts on the atmospheric circulation, we do not consider it obvious that the effect of this can be considered negligible.

The second key point is, as discussed in our paper, that there is evidence in the literature that the teleconnection depends on the atmospheric mean state, in particular the position of the storm track. Since the storm track is almost always biased to some degree in climate models, it does not seem unreasonable to suggest that the sea ice region in models best placed to interact with the storm track is slightly different than that in observations.

The fundamental issue here is that external forcing, including that from teleconnections, very often projects onto the dominant modes of variability (e.g. [3,4]). Not only do these differ between models and observations (Figure 2), but in the case considered here, there is non-linearity embedded at both ends: with sea ice as discussed in [1] and with the North Atlantic Oscillation in the visible multimodal behaviour of the jet [5]. We therefore take the view that model biases, in both the mean and the variability, cannot be easily ignored, and indeed many studies have examined the influence of such biases on teleconnections (e.g. [6] for just one recent example). There are also several precedents in the literature for using sea ice EOFs to compute Arctic-NAO teleconnections (e.g. Wang, Ting and Kushner 2017, or the Strong et al 2009 paper you pointed us to in your comments), and such approaches would inevitably highlight different regions when applied to models vs observations. It is certainly true that allowing for regions or patterns to shift in models opens up the possibility of cherry picking, and so sensitivity to such shifts should be clearly discussed, which we failed to do. But the flip side is that allowing for no model-dependent diagnostics may overly penalise models and give the impression that model skill (or inter-model consensus) is

weaker than it is.

It is the authors' impression that there has perhaps been too little consideration in the literature on potential (small) shifts in the key sea ice region, and we think this is an important point that we wish to highlight as part of our work. The revised version will expand on all the above points to better justify the choice made. Of course, we accept that the reviewer may disagree on some or indeed all of the above points, or be of the opinion that a proper justification of the above points would require more work which would likely be inappropriate to include in this paper. We hope that if this is the case, that our emphasis of the qualitatively similar results obtained with Barents-Kara, and the change from using Barents-Greenland to Barents for the model, will nevertheless allow you to consider your objection adequately addressed.

References:


1. Koenigk, T., Caian, M., Nikulin, G. et al. Regional Arctic sea ice variations as predictor for winter climate conditions. Clim Dyn 46, 317–337 (2016). https://doi.org/10.1007/s00382-015-2586-1
2. Sun, L., Deser, C., & Tomas, R. A. (2015). Mechanisms of Stratospheric and Tropospheric Circulation Response to Projected Arctic Sea Ice Loss, Journal of Climate, 28(19), 7824-7845.
3. Shepherd, T. Atmospheric circulation as a source of uncertainty in climate change projections. Nature Geosci 7, 703–708 (2014). https://doi.org/10.1038/ngeo2253
4. Corti, S., Molteni, F. & Palmer, T. Signature of recent climate change in frequencies of natural atmospheric circulation regimes. Nature 398, 799–802 (1999). https://doi.org/10.1038/19745
5. Woollings, T., Hannachi, A. and Hoskins, B. (2010), Variability of the North Atlantic eddy-driven jet stream. Q.J.R. Meteorol. Soc., 136: 856-868. https://doi.org/10.1002/qj.625
6. Karpechko, AY, Tyrrell, NL, Rast, S. Sensitivity of QBO teleconnection to model circulation biases. Q J R Meteorol Soc. 2021; 147: 2147– 2159. https://doi.org/10.1002/qj.4014


**RC1:** *The model the authors use may be an outlier and the results may not be that relevant to other models. This is very briefly mentioned in the discussion, but I think there are reasons to think this may not work as well in other models. Most models tend have a weak connection between reduced sea ice and a negative NAO. In addition, as mentioned in the introduction, model experiments forced with reduced sea ice also tend to show a weak negative NAO response. However the control model used here shows the opposite sign correlation compared to most models, and a previous study (Ringgaard et al. 2020, doi:10.1007/s00382-020-05174-w) shows that a version of this model shows no NAO response to reduced sea ice in the BK Sea. In addition, the improved correlation in the OCE version are still weak. Could it not be the case that the OCE is just improving the flaws in this particular model, which brings it more in line with other models? This would then mean that applying the same methods in other models may not have as large of an effect.*

**Response:** We would challenge the assertion that “most models tend to have a weak connection”. The range of correlations between Barents-Kara and the NAO found across the coupled CMIP6 models is very well approximated by a normal distribution with mean 0, standard deviation 0.17 and a 95% confidence interval of 0.28. While the exact mean of 0.018 is positive, almost half the CMIP6 models have negative correlations. The EC-Earth3 CTRL ensemble, with its average correlation of -0.06, is in no way an outlier in this distribution and is in fact dead average: this was extremely briefly noted in the submitted paper (line 337), and we have now made this more clear by revising Figure 5 to include the CMIP6 distribution. The inclusion of additional ensemble members has also now produced CTRL members with slightly positive correlations in the period 1980-2015, so there seems to be even less cause to find EC-Earth3 particularly objectionable. Its biases in the mean ice state are also in no way notably worse than many other models.

Note that the slightly positive mean of the CMIP6 distribution is consistent with findings in earlier literature reviews which report that `most' models show a positive association, but it is clear that this consensus is weak. Another point here is that many of the experiments carried out in the literature are not directly comparable with each other: e.g. many model experiments analysing the role of sea ice use fixed anthropogenic forcings, while the models we consider here are using historical forcings. This may account for any remaining discrepancies.

That being said, the point that the stochastic schemes may have differing impacts in other models should have been emphasised more. There are examples from earlier work which show consensus across models in some cases and lack of consensus in others. This will be expanded on in the revised manuscript.

Finally, it is perhaps worth remarking that if these stochastic schemes have little effect on the teleconnection when applied to a model with already realistic sea ice variability and a realistic teleconnection, then that would be a good thing. The purpose of the schemes is to fix variability where it is bad, not apply a uniform impact across all models. This is actually seen to happen as well: stochasticity added to a model with a poor ENSO led to a dramatic improvement of ENSO, but the same scheme applied to a model with an already good ENSO led to no real change in ENSO: https://doi.org/10.1007/s00382-019-04660-0 and https://doi.org/10.1175/JCLI-D-16-0122.1

**RC1:** *3. The authors claim that mean state changes cannot explain the differences, but I don’t find their arguments that convincing. They argue that AMIP ensemble with prescribed SSTs and sea ice show weak correlations. First of all, taking the correlations of the AMIP ensemble at face value would suggest that close to half of the difference can be explained by the mean state. Second, there are many other difference related to the coupling of sea ice and SSTs that could cancel out the improvements made by correcting the mean state biases in the AMIP experiments. It is likely that the improved mean state explains at least some of the differences and it can’t be ruled out that it is entire explanation.*

**Response:** The potential importance of the mean state is a point raised by all of the reviewers, and upon further consideration we agree. We have revised the discussion in several places to make clear that the mean state changes may be playing a role.

**RC1:** *4. The authors conclude that the link between sea ice and the NAO is stronger because of improved ice-ocean-atmosphere coupling. This is a bit vague and could be investigating a little further. What about the coupling is actually being improved? Because the authors argue that coupling on short timescales can explain the difference, there could be a lot value in doing similar analysis to what was done in Figure 7, but with other variables. For example, does the OCE ensemble have a stronger upward heat flux and temperature response following reduced sea ice?*

**Response:** We carried out a variety of such analysis during the revision period, but struggled to find anything extremely conclusive when looking at the local coupling between heatfluxes and sea ice. The most suggestive plot, Figure B7 in the revised, shows lag correlations between daily heatfluxes and daily sea ice, both averaged over the Barents sea region. There is no clear difference when sea ice leads the heatfluxes, but there is some suggestion that CTRL is doing a worse job when the heatfluxes lead the ice. This generally seems consistent with the original analysis, where we didn't see evidence that CTRL was doing worse with the initial anomaly, it was simply failing to evolve the anomaly correctly as the season progresses.

We have also come to understand that our interpretation of the LIM results were not entirely correct. The conclusion that ice-ocean-atmosphere coupling is important and likely improved in OCE still remains, but the more accurate interpretation now points directly to some effects which CTRL may be doing wrong. In particular, we now highlight the role of more remote adjustments to the ice and ocean from the initial sea ice anomaly, which may be done worse in CTRL (as hinted at by Figure B7). We have also highlighted the potentially disruptive role of the unrealistic ENSO teleconnection in CTRL. All this discussion, and more, is now included in the revised LIM section and also the Discussion at the end.

Unfortunately, we still cannot point to a clear mechanism. Looking into the role of the aforementioned possibilities (remote adjustments, ENSO, etc) is simply beyond the scope of what we are able to achieve in this paper! We hope the added discussion and pointers will satisfy the reviewer anyway.

**RC1:** *5. The title and abstract need to be more specific. Many different links between the Arctic and the midlatitudes have been hypothesized via a number of different mechanisms. It is misleading to refer to Arctic-midlatitude links very generally, when the authors have only investigated one specific link between November Barents-Kara sea ice and the winter NAO in interannual variability. Even with this correlation, the authors have only looked at one mechanism (they have not investigated the stratospheric mechanism).*

**Response:** We have edited the title and the abstract to more specifically refer to teleconnections with the North Atlantic Oscillation. We also included a line in the Discussion

and Conclusions pointing out that we made no attempt to separate the tropospheric and stratospheric pathways.

## Other comments

**RC1:** *L35: What is meant by 'More seriously'? Are the model experiments with imposed sea ice anomalies not serious?*

**Response:** This was poorly phrased of us. What was meant was "Perhaps more alarmingly, [...]", the entirely unstated point being that while variations between models could plausibly arise due to model biases even in the absence of significant decadal variability, variability within a single model seems to point more unambiguously towards the role of chaotic internal variability. We have simply reworded to "In addition, [...]", since this will be discussed explicitly later anyway.

**RC1:** *L35-38: Another recent study that could be cited/discussed here is Siew et al. 2021 (doi:10.1126/sciadv.abg4893).*

**Response:** We have included a citation to this.

**RC1:** L*30-42: Somewhere in this discussion it should be mentioned that observed correlation seems to be highly intermittent when looking at the much longer record (Kolstad and Screen 2019, doi:10.1029/2019GL083059). In the middle of the 20th century, the sign of the connection appears to be opposite compared to the recent period.*

**Response:** We think the analysis in Kolstad and Screen is flawed, because there appears to be a clear degradation in the sea ice data quality prior to the satellite era. For example, HadISST is what is used for ERA20C and CERA20C (data sets used by Kolstad and Screen), and the documentation states outright that the data is "mostly climatologies before the 1950s":

https://climatedataguide.ucar.edu/climate-data/walsh-and-chapman-northern-hemisphere-sea-ice

In fact, it is easy to see the degradation in sea ice data by plotting the daily Barents-Kara time series and visually observing that there are long periods between 1900 and 1979 where there is little/no variability or just repeating climatological values over several years. We attach a snapshot of the timeseries below in Figure 1 for the reviewers benefit. Kolstad and Screen do not adequately address this, and in fact seem to barely comment on it at all, despite the fact that this would be expected to have a big impact on the correlations.

The authors hope to address these issues more directly in future work, but since it is mostly tangential to the aim of the present work, we have simply included a very brief line and a

footnote in the introduction where we cite the paper but indicate how it is confounded by poor sea ice data prior to 1979.

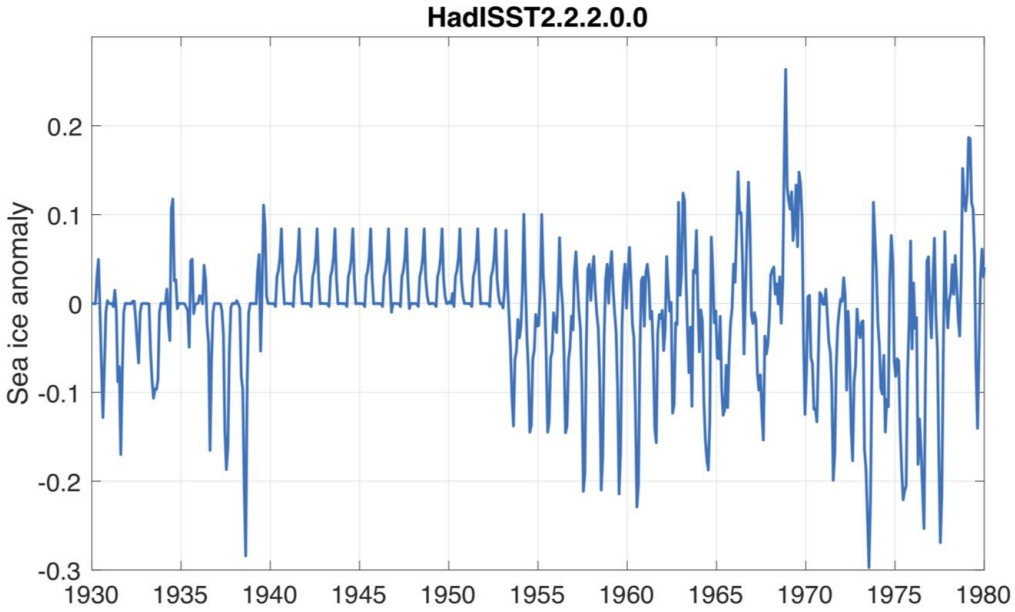


**Figure 1:** *Timeseries of detrended monthly Barents-Kara sea ice, using HadISST data. Covering 1930-1980.*

**RC1:** *L38-42: This is not an accurate description of Blackport et al. 2019. This study has nothing to do with the connection between November BK sea ice and the winter NAO and is not that relevant for this study. A much more relevant study that argues that the correlation between November BK sea and winter NAO may not be causal is Peings, 2019 (doi:10.1029/2019GL082097).*

**Response:** Thanks for the pointer to Peings, which we will include (see also below point). About Blackport et al 2019, we have clearly been too cavalier with our statements about what it does and doesn't say. Firstly, ibid uses DJF quantities for both sea ice and the circulation, so is not clearly relevant for November-leading-DJF teleconnections, and secondly the focus is on surface temperature anomalies, with the NAO barely mentioned. However, we disagree with your assertion that it is not relevant for this study. The circulation response aspect of Blackport et al's Figure 4 (clearly an NAO signal despite not being explicitly identified as such) is almost entirely reproduced in our paper using a partitioning of years based on November ice/heatfluxes and DJF pressure anomalies. This is what our Figure 9 shows, though we of course mischaracterized this as a reproduction of Blackport et al there too. The arguments of Blackport et al could therefore be transposed to the teleconnection we discuss with little/no edits required. More generally, we consider Blackport et al to be an influential and important paper on the role of internal variability in Arctic-midlatitude links, so wish to cite it and use its methods. In fact, we have been explicitly asked to consider using its methods on 3 separate presentations of our work, which further motivated this.

Of course, the existence of Figure 9 is not yet known, so we have minimised the reference to Blackport et al 2019 here, and rather discuss it more in the relevant section.

**RC1:** *L41: Warner et al. 2020 do not suggest tropical forcing as a common driver of sea ice and the NAO. They did suggest this may be the case for other aspects of the mid-latitude circulation, but not the NAO.*

**Response:** Thanks for the pointer. You're right, upon rereading the paper it's clear they seem to carefully avoid explicitly making the hypothesis for the NAO itself, referring only to internal variability as an explicit hypothesis in that case. We will now rather refer to Peings et al 2019 and Ural blocking as a suggested common driver.

**RC1:** *L198-207/Figure 1: The main takeaway from this is that OCE reduces the sea ice everywhere. The changes in variability are also entirely consistent with just a reduction in sea ice extent everywhere .*

**Response:** We partially agree and have simplified the discussion. The changes in the sea ice standard deviation seem to us more consistent with a change in the sea ice edge, rather than a blanket reduction in sea ice: the difference CTRL-ERA5 is negative near the pole and positive further equatorwards, while a blanket reduction would give negative everywhere. This is now explicitly mentioned in the text.

**RC1:** *Figure 1 and 3: I think that it would be more useful to show plots for OCE-ERA5 as well to make the improvements easier to see.*

**Response:** The differences between OCE-ERA5 and CTRL-ERA5 are quite challenging to see by eye: a linear colorbar can't be picked which emphasises these differences ithout causing massive saturation effects in CTRL-ERA5. We have therefore left the Figure as is, but have added a comment in the text to avoid confusion.

**RC1:** *Figure 2: What does the sea ice variability look like in the Barents-Kara sea in CTRL? There is substantially less variability connected with the EOF1, but is that because it is in other EOFs or because there is substantially less variability? I don't think it is latter based on Figure 1.*

**Response:** Your guess is correct: the standard deviations of the CTRL and OCE sea ice time series are quite close to each other. We interpret this to mean that the typical spatial pattern of Arctic sea ice evolution differs between CTRL and OCE, which is potentially important for understanding why CTRL and OCE behave differently. The relevant discussion has been extensively revised based on previous comments.

**RC1:** *L218: sea surface temperatures*

**Response:** Fixed.

**RC1:** *L243: Blackport et al. 2019 did not do this and has little to do with the NAO.*

**Response:** Sorry, we got our Blackport et al's mixed up, this should have been 2021, not 2019. We note that strictly speaking that paper did not ever regress November ice onto DJF pressure anomalies, but they did regress DJF ice onto DJF pressure anomalies, which we hope the reviewer agrees is sufficiently similar to warrant a citation.

**RC1:** *L279-281: I don't understand this. The Bering sea is a completely different region which will have different impacts on the circulation, so I don't see how it can be the equivalent to the BK Sea.*

**Response:** Our text was not very clear here. However, the point we were attempting to make was related to the patch of significant correlations in the Bering sea in the CTRL ensemble. With the increased ensemble size, this patch has vanished, making this redundant, so we've simply removed the relevant text.

**RC1:** *L281-283: There has been a lot more work looking at the response/correlation to sea ice in different regions than what is portrayed here (e.g. Screen 2017 doi :10.1175/JCLI-D-16-0197.1, McKenna et al .2017 doi:10.1002/2017GL076433, Blackport et al. 2019). The reason there has been more on the Barents-Kara is because there are stronger links in both observations and models.*

**Response:** We thank the reviewer for the references which we clearly should have included, and which we now will. The obvious blooper of not citing Blackport et al 2019 here was because we were focused on the NAO at this point, while ibid focuses on surface temperature response (ignoring the fact that we have elsewhere acted as if Blackport 2019 *does* deal with the NAO…). In any case, the discussion here has changed because of previous revisions: the new citations are still included elsewhere.

**RC1:** *L307: I don't think any study, including Koenigk and Brodeau (2017), state that the observed signal is a spurious signal. This study, and others like it, express caution that it could be. There is a lot internal variability and spurious signals can arise in model simulations of similar length to the observed record even when there is no/weak signal overall. It also the case that the recent observed correlation appears to be unusually high compared to the longer record(Kolstad and Screen 2019).*

**Response:** We have reworded from "[...] is in fact just a spurious signal" to "[...] may just be a spurious signal". Kolstad and Screen was already commented on earlier.

**RC1:** *Figure 5a: The fact that all simulations start off with a higher correlations than over the whole period intrigues me. Because all simulations start of the same ocean state, is it*

*possible that they happened to be initialized in particular state of low frequency variability that contributes to a stronger correlation?*

**Response:** Yes, that's possible. The fact that the difference between OCE and CTRL could be a result of this was explicitly stated in the Discussion (line 543). Another possibility is that the teleconnection depends on the mean sea ice state, which is very different at the start vs at the end (cf discussion starting line 270 of original draft). There is also a robust trend in the NAO in the models (commented on later here). These possibilities are now discussed further in Section 4.1. The line in the Discussion has also been edited to be even more explicit.

**RC1:** *L317-319: I don't understand why that would suggest it is coincidental. You wouldn't be able to rule it out, but that is very different from suggesting that it is.*

**Response:** This is no longer relevant due to the doubled ensemble size.

**RC1:** *L322-323: Is it actually the case that each 30 year period is statistically significant from 0? I doubt that this is the case given that some 30 year periods show correlations close to 0.*

**Response:** The statement was that the time-series of concatenated ensemble members has statistically significant correlations, not individual ensemble members. The latter statement is false as you say, while the former is true. This is again less relevant with the increased ensemble size so we have cut this.

**RC1:** *L328: How often do they attain correlations that exceed the observed correlation?*

**Response:** The Figure has been changed in light of your next comment about it being misleading, which we agree with. The answer to your question as written is that 14 out of the 255 30-year chunks in the SPHINX ensemble had correlations exceeding 0.39 (with a maximum of 0.48), made up of two periods of ~7 `consecutive' 30-year chunks (i.e highly overlapping). In the less misleading changed Figure 5, there are precisely 2 model simulations out of 79 independent 1980-2015 simulations for which the correlation exceeds 0.39, one of which is a SPHINX ensemble member.

**RC1:** *Figure 5b: I think it is misleading to plot it this way because the overlapping 30 year periods are obviously not independent. There are really only about 6 independent data points in the OCE distribution. I don't doubt that the differences are statistically significant, but this plot likely exaggerates the perceived significance.*

**Response:** We agree, and have changed the plot to rather show entirely independent 1980-2015 samples by using the coupled CMIP6 ensemble + HighResMIP + SPHINX and our CTRL simulations.

**RC1:** *L350-352: Isn't it more relevant to know whether or not these correlations are statistically different from the correlations in OCE or CTRL?*

**Response:** Perhaps, but it is problematic to say something like "AMIP is significantly different from OCE", because this is a statement about comparing two distributions (the ensemble members of CTRL/OCE and AMIP cannot be compared like for like). In the submitted manuscript, both these distributions would be estimated using only 3 points making such a statement hard to justify. In our revisions, CTRL and OCE now have 6 members, but AMIP still only has 3, so the same problem remains here.

Our choice to rather just argue that there is no teleconnection in AMIP was a pragmatic one based on this.

**RC1:** *L360-368: The regressions of November zg500 on November sea ice is likely not the response to the sea ice anomalies(at least not entirely). Instead, a large part of it is the atmospheric circulation that forces the sea ice anomalies. The sign of the NAO is opposite to what would be expected if it was the response. Unless the authors are arguing that the initial response to reduced sea ice is a positive NAO, but that contradicts what is shown in Figure 7.*

**Response:** True, this was also pointed out by another reviewer. The pattern for November there will be a combination of atmospheric forcing on the ice and vice versa, which our text didn't address. This is now discussed in the revisions.

**RC1:** *L380-385:This negative feedback between the sea ice and NAO was identified in a number of studies including Strong et al. 2009,doi:10.1175/2009JCLI3100.1 .*

**Response:** This is an excellent reference, thank you. To be clear, we were not claiming originality here, though we should have stated this clearly and made a citation. Incidentally, the technique of Strong et al is relevant to your Major Objection 1, since they use an EOF based sea ice index which clearly captures sea ice anomalies more broadly than just in Barents-Kara. As shown in Figure 2 of our work, EOFs will likely be different in models vs observations, so their technique would pick out somewhat different regions. This will therefore be cited for this discussion as well.

**RC1:** *L435: This is not reproducing the result of Blackport et al. 2019. They examined the regression between winter circulation and winter sea ice, not November sea ice.*

**Response:** We corrected this, as discussed also earlier.

**RC1:** *L425-456/Figure 9. I am not sure I understand the point of this analysis. The authors have already established that feedback between sea ice and the NAO, so I don't see how the NAO forcing of the sea ice could explain the difference between OCE and CTRL. There could potentially by a stratospheric pathway where there are causality issues, as suggested by Peings 2019, but the authors have effectively argued against this being the reason for the*

*improvement by showing that difference can entirely be explain based on the daily coupling. The authors should more clearly explain the motivation for it, or remove it.*

**Response:** The point is that other atmospheric variability that isn't the NAO might affect the results. For example, if there is another common driver of both the ice and the NAO (whether systematic or purely due to random decadal variability) then this might give the appearance of ice-NAO coupling (and associated correlations) even if there is no such coupling. The technique of Blackport et al 2019 is an elegant way to test for the influence of atmospheric variability in a very generic manner (i.e., without prescribing what the other atmospheric variability actually is) which is why we think it is a relevant technique to use to address this. This motivation has now been made clearer.

**RC1:** *L463:Figure 9->Figure 10*

**Response:** Fixed.

**RC1:** *L516: How would the varying model biases contribute to the inconsistencies within long simulations from a single model? Note that there also appears to be large inconsistencies between short periods in observations as well (Kolstad and Screen 2019).*

**Response:** Our point was unclear. We have rewritten to clarify: the variations of the correlation coefficient, both between models and across long fixed-forcing simulations, is consistent with a hypothesis that most models fail to simulate the teleconnection (due to e.g. inadequate coupling, as we hypothesise here). As mentioned in the Methods, a basic AR1 null hypothesis has a 95% confidence interval of 0.35, consistent with the spread of both CMIP6 models and, e.g., the results of Koenigk and Brodeau (2017), and indeed our own results using CTRL+SPHINX.

Kolstad and Screen 2019 has already been discussed.

**RC1:** *What do the trends in NAO look like? If the improved correlations represent a response to sea ice loss, it may be expected that there is more negative NAO trends in the OCE simulations. This could have implications for the midlatitude response to sea ice loss and global warming, not only for seasonal predictions. This may be a bit beyond the scope of the study, and a larger ensemble may be needed to find robust differences, but it would really simple to check.*

**Response:** We include a Figure showing the trends for the benefit of the reviewer (see below). All ensemble members show a negative trend. On average the OCE members have slightly steeper trends than CTRL, but the difference appears small. We consider this beyond the scope to look into further, but we have included a line mentioning the possible importance of these trends in Section 4.

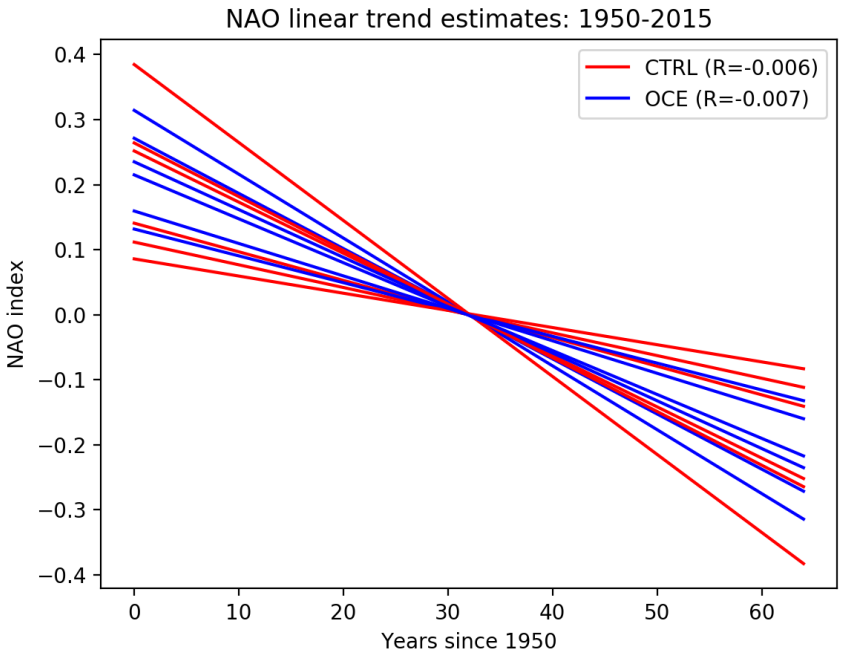


**Figure 2:** *Linear fits to the DJF NAO timeseries of each ensemble member of CTRL and OCE. The values R in the legend are the mean slope of all 6 ensemble members.*

# REVIEWER #2

**Major comments:**

**RC2:** *1. The authors argue with respect to Figure 9 that OCE and ERA5 have similar daily timescale forcing, suggesting that OCE is getting things right for the right reasons. I'm not entirely convinced of this given that the b coefficient in OCE is larger than ERA5. It would be interesting to see the coupling between ice and other variables using the LIM to provide a bit more evidence that OCE is getting things right, for example the relationship between ice and a variable that is more thermodynamically connected to ice. The authors also note that the difference seen in Figure 9 could be due to chance. If so, can you show similar plots as Figure 9 for each ensemble member of OCE? If chance plays a role maybe there is some evidence of this if all ensemble members are examined individually.*

**Response:** After doubling the ensemble size the mismatch of OCE with observational data in Figure 9 has been notably reduced. The improved teleconnection in OCE still appears more driven by the forcing of the ice on the atmosphere, but a clear NAO signal is now also seen for years where the atmosphere drives the ice. We hope this will help reassure the reviewer.

As for coupling with thermodynamic variables, a new figure looking at this has been added to the appendix and discussion has been added in the LIM section. In short, our analysis suggests there is nothing obviously wrong with the heatflux response to sea ice anomalies in CTRL, but there is some hint that ice adjustments to heatflux anomalies are not done well.

**RC2:** *2. Figure 9h and 9i seem to suggest something is quite unrealistic about how this model represents fall sea ice variability. In the Blackport et al. (2019) paper, they examine a version of EC-Earth, EC-EarthV2.3, I believe. Are you able to reproduce their findings with EC-Earth3P used here for the CTRL runs (it would be great to see plots similar to their Fig. 4c, f, and i? It seems that you are getting very different patterns (Fig. 9h), which makes me concerned about the suitability of this model for this study.*

**Response:** As pointed out, the mismatch between ERA5 and OCE is now much less notable with the doubled ensemble size.

It is perhaps also worth pointing out that we are either way still suggesting that there is "something quite unrealistic" about the CTRL model, to paraphrase the reviewer. We are suggesting that the lack of a teleconnection is unrealistic, and that its improvement in OCE is a genuine improvement. The point being that this is an important result *even if CTRL is unrealistic* in some other ways, because it implies that the considerable intermodel spread in reproducing the observed teleconnection may to a large extent be due to model biases rather than internal variability. If that is the case, then the teleconnection may be much more robust than many studies suggest it is. But in any case, EC-Earth3 does not seem to be a

particularly poor model, and it is very likely that many CMIP6 models behave similarly: see the response to RC1 for more on that, or see the revised Section 4.

Note that the EC-Earth figures from Blackport et al. 2019 are not reproducible with our data. While the model used is closely related, the EC-Earth experiments considered in Blackport et al essentially use fixed forcings (they consider 400 5-year simulations each covering the same period), while our experiments are 65 successive years with historical forcings. Identical diagnostics would not be expected as a result, so we don't see any discrepancies here as a point of concern.

**RC2:** *3. Could the direct effect of mean state changes be quantified using AMIP-style runs with monthly sea ice and SSTs from the coupled OCE runs? I think it is important to get a better sense of what is going on - is it the stocasticity itself or the effect of the stocasticity on the mean state. Untangling this has implications in terms of how this study informs model development.*

**Response:** When it comes to elucidating the mechanisms more clearly, we produced some additional lag correlation/regression plots between sea ice and heatfluxes (this also being suggested by RC1) as well as some other diagnostics to help clarify. While these do hint at some small improvements in OCE to the daily time-scale local coupling between ice and heatfluxes, our analysis generally suggests that the flaws in CTRL are not clearly visible in the local, short timescale thermodynamic coupling. Instead, the errors in CTRL appear to be primarily due to errors in the subsequent adjustment and growth of the initial pressure anomaly across the North Atlantic and ice edge more broadly. In fact, this is already what the LIM results suggest, but this was not really made clear in the submitted manuscript. All this is discussed (and the relevant new plots included) in the revised paper. Unfortunately, a thorough analysis of errors in the more remote response is not going to be possible to include in this already lengthy paper and will have to be left for future work.

Regretfully no time or resources are available to carry out experiments of the sort you describe at present, though we agree they would help. The role of the mean state (also raised by the other reviewers) is discussed in more detail in the revised manuscript in any case, but it has not proven possible to decisively nail down the contribution of mean state vs coupling in our analysis. Besides the complication of local vs remote adjustments raised above, it is likely that the inherently non-linear component to ice/heatflux coupling plays a role which our analysis, entirely based on anomalies, cannot detect. Possible non-linear diagnostics that could be explored in follow-up work are discussed in, e.g. Caian et al. *An interannual link between Arctic sea-ice cover and the North Atlantic Oscillation* (2018), Clim Dyn.

It is perhaps also worth stating that in almost all earlier papers looking at the impact of stochasticity that we are familiar with, it has proven extremely challenging to determine exact mechanistic pathways. This is effectively because when turned on, the stochastic schemes typically alter both the variability and the mean state within hours/days in a highly coupled manner. Untangling cause and effect therefore becomes very difficult without highly targeted experiments. We hope that the extra diagnostics and discussion, including of potential future work, will satisfy the reviewer anyway.

**Minor comments:**

**RC2:** *1. lines 25-30: lots of issues with parentheses that need to be tidied up.*

**Response:** We fixed these.

**RC2:** *2. line 27: You may want to say "negative NAO" rather than just "NAO" for clarity.*

**Response:** Done.

**RC2:** *3. Section 2: there are many different abbreviations/acronyms for the model used in this section. After you finish describing the various configurations, can you tell the author which name you are going to stick with throughout the paper? Something like, "Hereafter, the model will be referred to as...".*

**Response:** We added something to this effect in Section 2.1 and 2.2.

**RC2:** *4. Line 153: What prescribed SSTs and sea ice?*

**Response:** Daily HadISST2 data. This is now explicitly stated.

**RC2:** *5. line 162: extra parentheses*

**Response:** Fixed.

**RC2:** *6. line 218: Figures -> Figure and ssea -> sea*

**Response:** Fixed.

**RC2:** *7. Table 1 caption: there is a missing section number - just shows ??*

**Response:** Fixed.

**RC2:** *8. line 405-406: I don't think this is the correlation you are showing. It's sea ice and NAO, correct?*

**Response:** No, it was correct as stated, but it is now clear that this is not quite the correct correlation to look at. Our description and analysis of the LIM in this section was not well done. We have substantially revised this to make things hopefully much clearer.

In brief, our confusion arose from the fact in our infinite-ensemble-mean LIM reconstructions, the LIM ice and NAO are perfectly correlated with each other (being perfectly determined by each other), so the correlation between the LIM DJF NAO and the True DJF NAO is the same as the correlation between the LIM Nov ice and the True DJF NAO. It is this latter

correlation which can be sensibly interpreted as a `forecast' using the LIM and its close match to the observed teleconnection correlations is a sign of the skill of the LIM. Correlations between the LIM ice and LIM NAO are now also discussed.

**RC2:** *9. FIgure 9i does not really look like Figure 9g to me. And it seems a bit strange that Fig. 9h does not look anything at all like Fig. 9g.*

**Response:** This Figure is now different given the doubled ensemble size, and Figure 9i and 9g are more easily comparable as a result. Fig 9g still looks different, which is interesting yes. It strongly suggests that even in years where the atmospheric forcing dominates the heatfluxes, the actual circulation response still depends essentially on the coupling with the sea ice. Otherwise you would expect the CTRL model to do fine here, unless it were the case that the CTRL model has biases in its atmospheric dynamics that are substantially improved by OCE. Given that the atmospheric components of CTRL and OCE are identical, this does not seem obvious, though of course we can't rule out some mean state change being crucial. We will comment on this in the revised.

**RC2:** *10. line 463: FIg. 9 -> Fig. 10*

**Response:** Fixed.

# REVIEWER #3

**Major comments:**

**RC3:** *1) Interpretation of Fig. 6*

*I'm not sure I fully agree with the interpretation of the lagged relationships between Z500 and sea ice – or I may have misunderstood the authors. The text L360–368 seems to imply that the Z500 anomalies are a "response" to the sea ice at all lag times. This makes sense at positive lags (December onwards, when Z500 lags the sea ice), but for the November anomalies (1st row of Fig. 6) we also need to consider the possibility that it is the circulation driving the sea ice, rather than the other way around. I think this is indeed what is happening: the Z500 anomalies are consistent with northerly flow into the Barents sea area, which would drive enhanced sea ice concentration. I believe this also explains why the November Z500 anomalies are so consistent among ERA5, CTRL and OCE. In any case, the possible two-way interaction between Z500 and the sea ice needs to be discussed in the context of Fig. 6.*

**Response:** Yes, you're absolutely right that there is a 2-way interaction there which we totally failed to comment on. This will be discussed in the revised.

**RC3:** *2) AMIP results*

*I am still unclear as to why the AMIP simulations show no midlatitude response to the sea ice anomalies. I understand the result in Fig. 7 that there is two-way coupling, and the NAO → ice effect is absent from AMIP. But the ice → NAO effect should be in AMIP, so why don't we see that? Also, is this result consistent with any prior work looking at AMIP runs with other climate models?*

**Response:** Yes, there is evidence in prior literature that this teleconnection is weaker in AMIP models. This was mentioned in line 520, citing Blackport and Screen (2021), though I believe earlier studies (cited in their paper) had pointed to this as well. For EC-Earth in particular, the study Caian et al. *An interannual link between Arctic sea-ice cover and the North Atlantic Oscillation* (2018), Clim Dyn, showed that ice/NAO links are weaker in an AMIP simulation than a coupled simulation, something they attributed to the missing coupling. Our paper provides further evidence to the importance of coupling to get a good teleconnection, though several questions remain about exact mechanisms. We show that while the initial, local ice->heatflux response appears fairly similar for both CTRL and OCE, the subsequent growth and evolution of the anomaly is significantly better in OCE. Presumably, as you point out, the initial local anomaly would be highly realistic in the AMIP simulations, but the failure to propagate the anomaly would likely be even worse given the total lack of coupling. Probably the propagation of the anomaly depends not just on the local response but the response in neighbouring regions (both the ocean and neighbouring ice) that are missing in AMIP. Caian et al. includes some discussion on possible mechanisms

here. This will be pointed to in the revised paper.

**RC3:** *3) Coupling timescales*

*I feel some clarification is needed on the timescales at play in the sea ice–NAO coupling. Figure 7 suggests the coupling happens on daily timescales; but it's not obvious how to reconcile this with the finding that the NAO responds to November sea ice anomalies on the timescale of a *season* (DJF). My interpretation would be that the sea ice anomalies are relatively persistent (Fig. B5), so the November anomalies are a skillful predictor of those occurring later in the winter season – and these anomalies continue forcing the NAO through the winter. Is this consistent with the authors' thinking? Please clarify in the paper.*

**Response:** Yes, exactly: the initial anomaly is long-lasting due to the persistence of sea ice, but is ultimately damped away by the opposing response of the NAO. We have substantially revised Section 5 to make this clearer. More discussion about the initial local response vs more remote adjustments are also included, as per point 2 above.

**RC3:** *4) Coupling in CTRL*

*Figure 8b suggests the BG sea ice in CTRL does have a measurable impact on the NAO, which appears at odds with the lack of an ice → NAO relationship in Fig. 7. Is this because the BG sea ice varies so little in CTRL – so that even though the effect is there, the impact is minimal because there's almost no forcing?*

**Response:** This question is related to our somewhat flawed interpretation/discussion of the LIM, especially as it relates to CTRL. This should hopefully be dealt with by our thorough rewrite of the section on the LIM, which also answers the reviewer's specific question about sea ice variation/initial conditions.

**RC3:** *5) NAO definition*

*I was unclear as to the NAO metric as defined L166, and since this is key to the result, the definition seems important. I don't understand the subtraction of the daily climatology after the calculation of the PC. Why not deseasonalize the data beforehand? If using non-deseasonalized data, there is a risk that the EOFs are capturing the seasonal cycle (an externally forced signal), rather than the true internal atmospheric variability. It was also unclear to me whether the EOFs were calculated for each CTRL and OCN realization separately, or whether these realizations were concatenated prior to computing the EOFs. While it probably makes little difference, I'd favor the latter, which should give more robust EOFs – and ensures any differences among the realizations aren't due to differences in the EOF basis.*

**Response:** The NAO EOF was computed separately for each dataset, to allow the centers of NAO action to shift between each dataset according to differences in the mean state: this will be made clearer in revisions. We believe it is important to allow for some shifts between

models to not obscure signals or overly penalise models (i.e. penalising both for mean state biases and changes to modes of variability). That being said, in this case there is very little difference between the NAO pattern in CTRL, OCE, and ERA5, with a pattern correlation between any two of around 0.97. The results are therefore highly unlikely to change if using the exact same NAO pattern for all three. This will be mentioned in revisions.

**Minor comments:**

**RC3:** *1) Please fix the citation format – the parentheses are often in the wrong places. I suspect this may be due to mixing the Natbib commands \citet and \citep in LaTeX. One example is L25, where it should be "(Hoskins and Karoly 1981)", "(Garcia-Serrano et al. 2015)".*

**Response:** We have now streamlined and corrected the use of citet and citep.

**RC3:** *2) Consider clarifying the definition of the word "deterministic" – not being a stochastic parameterization expert, I initially thought this might mean "prescribed SST" as opposed to coupled, when actually this means "not stochastic".*

**Response:** Thanks for pointing out possible ambiguity: we have clarified this in both the abstract and the introduction.

**RC3:** *Typos etc:*

*L52: "are a manifestation"*

*L169: "are computed"*

*L208: "to reduce"*

*L218: "sea surface"*

*L229: "Examination… supports"*

**Response:** Fixed!

**RC3:** *L297–300: This text is a repetition of L179–183, so I suggest deleting.*

**Response:** We deleted the repetition.

**RC3:** *L405: Strictly speaking, Table 1 shows the correlations between the LIM NAO and LIM sea ice – not LIM NAO with true NAO. The latter is shown in Fig. B6.*

**Response:** Actually, the line was correct as stated. However, the reason for only looking at this was a result of our flawed interpretation of the LIM in general, as discussed in earlier points. In brief, the correlation between the LIM DJF NAO and the True DJF NAO can be viewed as an infinite ensemble mean forecast of the true NAO, and the fact that these correlations match the ice-NAO teleconnection correlations suggest that the skill of these LIM forecasts is coming from the correct propagation of the ice initial conditions. Expected correlations between the LIM Nov ice and the LIM DJF NAO have also now been included and discussed, since as the Reviewer clearly notes, these are of obvious importance. The entire LIM section has been rewritten in a way which should hopefully clarify all this.

**RC3:** *L423: "may have changed" → I think you mean "between CTRL and OCN", but it's not entirely obvious from the phrasing.*

**Response:** Yes that's what we meant: we clarified this in the revised.

**RC3:** *Caption of Table 1, L3: broken link to section 5.2*

**Response:** Fixed.

**RC3:** *Figures 4 and 6: Suggest highlighting the BK and BG regions with boxes in the maps*

**Response:** Good idea, we have done this.