Seasonal predictions of the Winter North Atlantic oscillation: variability, forecast skill and the signal-to-noise paradox in a large ensemble

Article

Published Version

Comer, R. E. ORCID: https://orcid.org/0009-0007-5336-0244, Scaife, A. A. ORCID: https://orcid.org/0000-0002-5189-7538, Kettleborough, J. A. ORCID: https://orcid.org/0000-0002-4698-6712, Sutton, R. T. ORCID: https://orcid.org/0000-0001-8345-8583 and Davis, P. J. ORCID: https://orcid.org/0009-0004-5008-5660 (2026) Seasonal predictions of the Winter North Atlantic oscillation: variability, forecast skill and the signal-to-noise paradox in a large ensemble. Atmospheric Science Letters, 27 (2). e70008. ISSN 1530-261X doi: 10.1002/asl.70008 Available at https://centaur.reading.ac.uk/128663/

To link to this article DOI: http://dx.doi.org/10.1002/asl.70008

Publisher: John Wiley & Sons

Atmospheric Science Letters

RMetS Royal Meteorological Society

RESEARCH ARTICLE OPEN ACCESS

Seasonal Predictions of the Winter North Atlantic Oscillation: Variability, Forecast Skill and the Signal-to-Noise Paradox in a Large Ensemble

Ruth E. Comer[1] | Adam A. Scaife[1,2] | Jamie A. Kettleborough[1] | Rowan T. Sutton[1,3] | Philip J. Davis[1]

[1]Met Office Hadley Centre, Exeter, UK | [2]Department of Mathematics and Statistics, University of Exeter, Exeter, UK | [3]National Centre for Atmospheric Science, Department of Meteorology, University of Reading, Reading, UK

Correspondence: Ruth E. Comer (ruth.comer@metoffice.gov.uk)

Received: 17 October 2024 | **Revised:** 20 November 2025 | **Accepted:** 11 December 2025

Keywords: NAO | predictability | seasonal forecast | signal-to-noise | variability

ABSTRACT

We investigate the variability of the North Atlantic Oscillation (NAO) in the Met Office Global Seasonal Forecasting System (GloSea) using a 132-member ensemble of coupled model forecasts, which is larger than has previously been available. Consistent with previous studies, we find that the signal-to-noise ratio is too small to match the correlation skill, and we additionally find that this result is statistically significant for years when the El Niño Southern Oscillation is active, and therefore skill is higher. We also show that correcting the signal-to-noise ratio by only increasing the signal would produce total variability that is still within real world estimates, removing the necessity for nonlinear mechanisms to increase the signal at the expense of the noise. Finally, we find an inverse relationship between yearly ensemble spread and ensemble mean, suggesting that the negative phase of the NAO may be less predictable than the positive, although the relationship is partly due to a longitudinal shift in the NAO pattern, which in positive NAO years moves the centre of variability away from the traditional Azores location.

1 | Introduction

The North Atlantic Oscillation (NAO) is the dominant mode of variability in the North Atlantic sector. At seasonal timescales, the NAO has the single most important influence on surface climate variability in Europe (e.g., Hurrell et al. 1995; Stephenson et al. 2003). Skilful seasonal predictions of the surface winter NAO have now been demonstrated in dynamical seasonal forecasting systems (Scaife et al. 2014; Athanasiadis et al. 2017). However, the NAO predictions in these systems also exhibit anomalously low signal-to-noise ratios (Scaife et al. 2014; Eade et al. 2014; Baker et al. 2018). This results in the so-called signal-to-noise paradox: the counter-intuitive result that for skilful systems, the model ensemble mean better predicts the observations than it predicts its own ensemble members (Scaife and Smith 2018).

Increasing ensemble size provides higher correlation skill for the NAO (Scaife et al. 2014) and reduces the uncertainty in estimates of skill and other statistical properties of the system (Weisheimer et al. 2019; Doi et al. 2019). In the current study, we combine several hindcasts of the GloSea Global Seasonal forecasting system to create a significantly larger single-system ensemble than has previously been available for studies of this kind. We then use our large ensemble to explore the total variability of the NAO in the system and we carefully compare this to observed NAO variance. We also investigate the signal-to-noise problem and how the predictable signal and noise fluctuate in forecasts for different years.

2 | Data and NAO Definition

The GloSea data used in this study come from hindcast sets as described by MacLachlan et al. (2015), updated to use the Global Coupled model version 3.0 (Williams et al. 2018). Several experiments have been carried out to check the sensitivity of the NAO representation to ocean model timestep, stochastic physics and choice of gravity wave scheme. Since these experiments showed no significant differences between their NAO predictions, we combined the members to form one large ensemble with 132 members per year for 23 years (44 initialised on each of 25th October, 1st and 9th November 1993–2015). Descriptions of the individual hindcast experiments and the method used to compare them are provided in the Supporting Information S1.

For comparison with observations, we use the ERA5 reanalysis (Hersbach et al. 2020). The NAO index is calculated from the December–February means of pressure at mean sea level (MSLP) from each data set. We use a simple difference of the grid cells nearest to 38°N, 27°W (Azores) and 65°N, 20°W (Iceland), consistent with Scaife et al. (2014). We also checked these results using the box-based NAO definition from Stephenson et al. (2006), which uses the difference of the averages over two boxes across the Euro-Atlantic region 90°W–60°E: 20°N–55°N for the south and 55°N–90°N for the north. Except where discussed, the results are not materially affected by the choice of NAO definition, as found by Baker et al. (2018). We therefore focus on results using the traditional point-based definition.

3 | NAO Variability

Figure 1 shows the distribution of our winter mean NAO index from observational reanalysis data from ERA5 and the ensemble seasonal forecasts from GloSea. Note that the apparent bimodality[1] in the ERA5 distribution can be attributed to the small sample size (86 winters vs. over 3000 winters in the GloSea model data). This is easily illustrated by taking a random subset of 86 GloSea values; for example, Figure 1 (centre) shows the first such sample obtained using an arbitrarily chosen random seed. Here we see a similar pattern, despite the true distribution (Figure 1 right) being clearly unimodal.

To further compare the modelled and observed NAO distributions, we follow the 'UNSEEN' approach of Thompson et al. (2017). We take one random member from each year and calculate characteristic distribution statistics. Repeating this many times produces a distribution for each statistic which is then compared with a single observed value. Figure 2 shows the distribution of sample means, standard deviations and skewness calculated in this way. The vertical dashed lines show the respective statistics calculated from observational reanalysis for the same years, in order to test the null hypothesis that data with these distribution characteristics could have been generated by the GloSea system. For all three statistics, the observations lie within the central 90% of the model distribution. So we conclude they are not significantly different from the modelled distribution at the 10% level according to a two-tailed test. Despite this lack of significant difference, we note that the observed estimate for standard deviation is close to the significance threshold at the upper end of the distribution, so it is possible that GloSea underestimates the total variability in the NAO index.

4 | Correlation Skill

The temporal correlation between the modelled ensemble mean and observed NAO for the full GloSea hindcast 1994–2016 is 0.44. By bootstrap resampling the years, we obtain a 95% confidence interval for this correlation of (0.06, 0.74), so there is a lot of uncertainty in the value but the skill is almost certainly positive. Siegert et al. (2016) calculated a 95% confidence interval for the correlation skill in the original GloSea5 system of (0.19, 0.68), and our new estimate falls within that. Siegert et al. (2016) also showed that the correlation uncertainty reduces when using a longer hindcast period, so the large uncertainty shown here is partly due to GloSea's relatively short hindcast period.

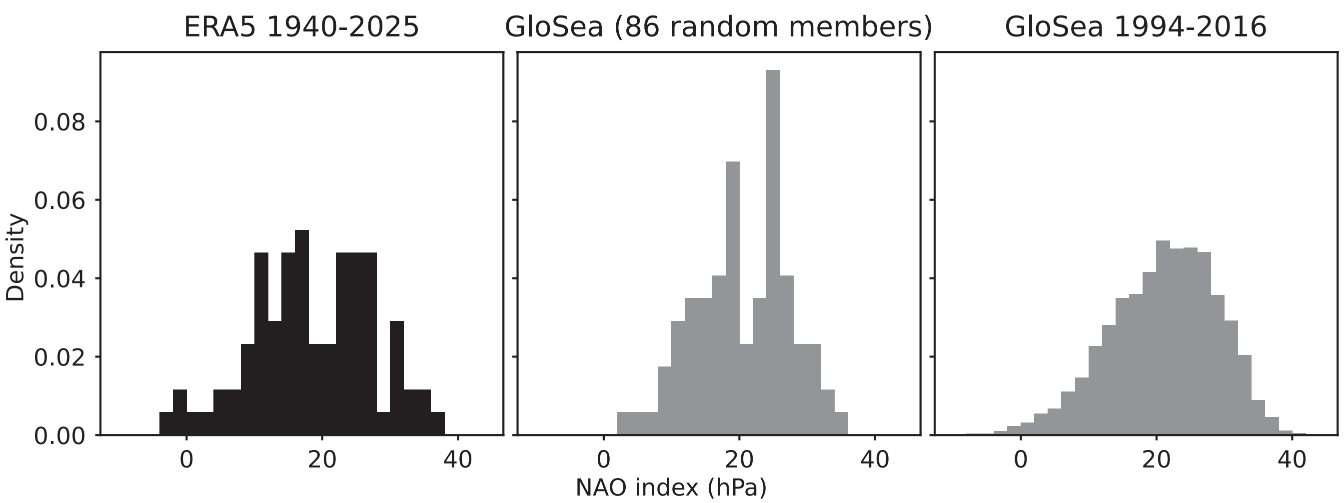

FIGURE 1 | NAO distribution in model and reanalysis: Distribution of NAO index calculated from all available data from ERA5 (left) and GloSea (right). Distribution of a random sample of 86 GloSea members (middle).

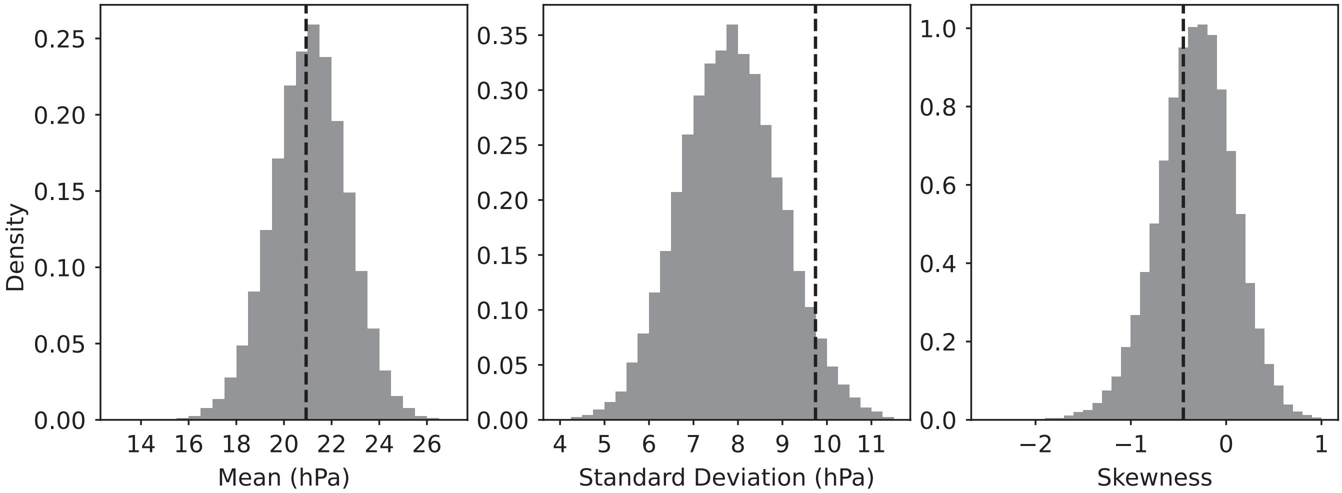

FIGURE 2 | Modelled NAO distribution is consistent with reanalysis. Distribution of statistics calculated by sampling one GloSea member per year (grey) for mean (left), standard deviation (middle) and skewness (right). Vertical dashed lines show observational reanalysis values from the same years (1994–2016), lying at the 45th percentile of the model distribution for the mean, the 94th percentile for standard deviation and the 35th percentile for skewness.

Correlation increases with ensemble size (e.g., Scaife et al. 2014), and Murphy (1990) provided a method to estimate the correlation $\langle r_M \rangle$ for a specific ensemble size M (his equation 2)

$$\langle r_M \rangle = \frac{\sqrt{M}\langle r_1 \rangle}{\sqrt{1 + (M-1)\langle r_{int} \rangle}}$$

where r_1 is the correlation calculated from one model member per year and observations, and r_{int} is calculated by correlating single model members with each other. The angle brackets denote the mean over many such samples. Taking the limit of Murphy's equation as ensemble size tends to infinity, we have

$$r_\infty = \frac{\langle r_1 \rangle}{\sqrt{\langle r_{int} \rangle}}$$

where r_∞ is the correlation between the mean of an arbitrarily large ensemble and observations. Applying this calculation to our large GloSea ensemble gives us an estimate of $r_\infty = 0.46$ (with a 95% confidence interval from resampling the years of (0.06, 0.77)). The small difference between our calculated correlation score and this theoretical limit suggests there is little to be gained by increasing the ensemble size beyond the 132 used in this study. By subsampling our ensemble to a more conventional size of 50 members per year, we obtain a central correlation estimate of 0.42, which is also not far below the theoretical limit. However, 95% of these subsamples have correlations ranging from 0.30 to 0.53, indicating a fairly large uncertainty due the choice of members at this size.

5 | Ensemble Spread is Inversely Related to Ensemble Mean

We further investigate ensemble forecast NAO variability by looking at year-by-year statistics. The left-hand panel of Figure 3 shows the standard deviation across all 132 members for each year plotted against the respective mean. We see a clear negative relationship with a correlation of −0.73 between ensemble spread and ensemble mean NAO. That is, more positive NAO years have smaller ensemble spread. This relationship is consistent with the fact that the overall distribution is negatively skewed (Figure 1, right), so samples taken from the lower end of the distribution can be expected to have a larger spread than samples taken from the upper end. The correlation is significant beyond the 0.1% level according to a *t*-test.

This negative relationship is also apparent if we use the box-based NAO index of Stephenson et al. (2006), however, it is less striking, with a correlation of −0.56. To investigate the difference, we look at the spatial pattern of MSLP variability. Having removed the ensemble mean for each year, the upper right panel of Figure 3 shows the standard deviation across all members for the five most *negative* years (the years for which GloSea's ensemble mean NAO are lowest). The middle right panel shows the same for the five most *positive years*, and the lower right panel shows the ratio of these. The choice of years here is not sensitive to the NAO definition. Much of the reduction in spread as we move from negative to positive NAO is in a region incorporating the Azores. There is also an increase in variability across central-eastern Europe. This decrease in variability to the west and increase to the east results from an eastward shift in the centre of variability in the southern node of the NAO. So the reduction in NAO spread is, in part, due to a shift in the region of variability away from the fixed NAO index location.

An eastward shift in the position of pressure variability with more positive NAO has been analysed in previous modelling studies (Peterson et al. 2002, 2003; Dong et al. 2011) and is consistent with the observational study of Hilmer and Jung (2000). The latter noted an eastward shift in the NAO location towards the end of the 20th century when it was also more positive than in earlier years.

The large-scale increase in ensemble spread for negative NAO years over the North Atlantic suggests there may be factors that make the negative phase of the NAO inherently more uncertain and less predictable. Indeed, the relative operating characteristic

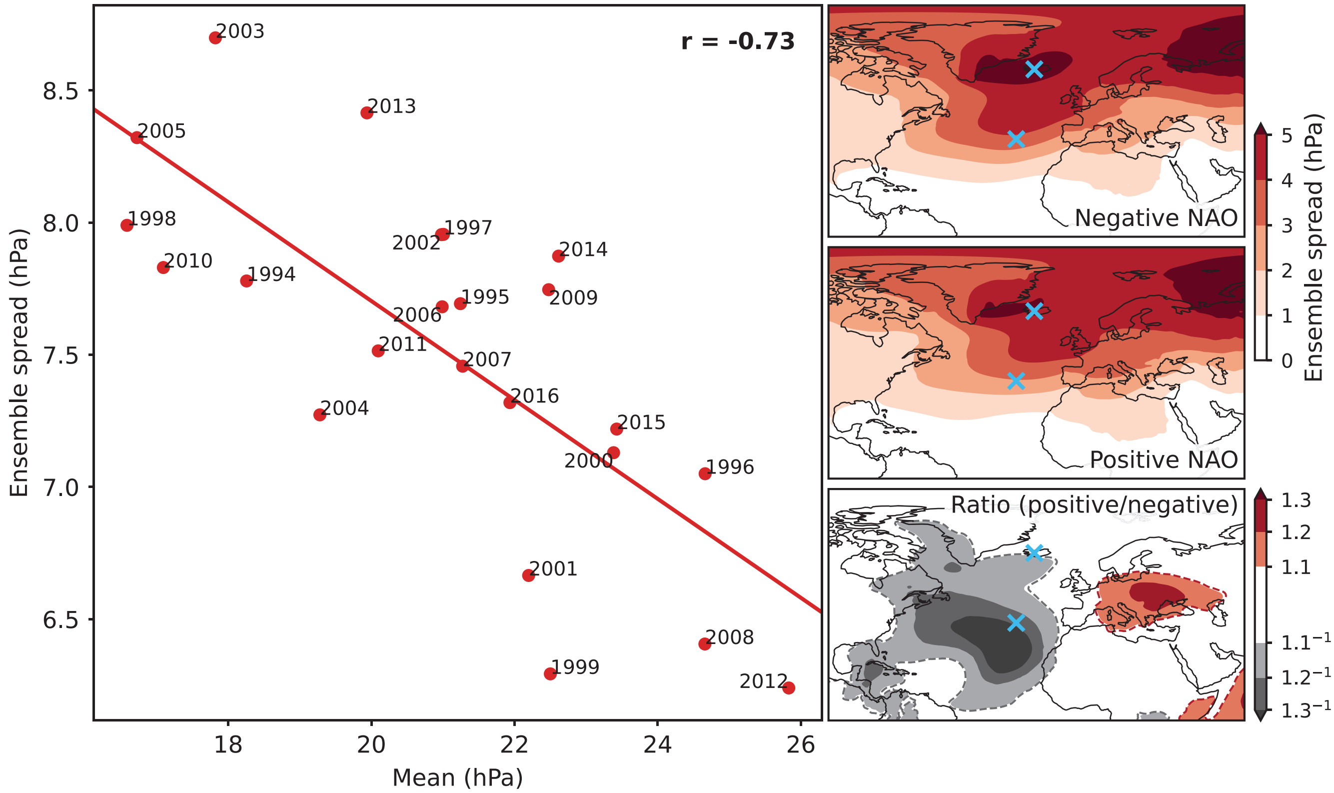

FIGURE 3 | NAO ensemble spread is inversely related to the ensemble mean. Left panel: GloSea hindcast point-based NAO standard deviation plotted against ensemble mean for each year together with least squares regression line. Right panels: MSLP member standard deviation around the ensemble mean for five most negative years (1998, 2005, 2010, 2003, 1994; upper) and five most positive years (2000, 2015, 2008, 1996, 2012; middle) using 660 members in each case; the ratio of positive and negative years' ensemble spread (lower), with dashed lines marking statistical significance at the 1% level according to a one-tailed *F*-test. Blue crosses indicate locations from which the point-based NAO index is calculated.

(ROC) score for the lower quintile is only 0.53, showing no skill, whereas the ROC score for the upper quintile is 0.82, which is significant at the 5% level according to a Mann–Whitney *U*-test (Mason and Graham 2002). This is consistent with Weisheimer et al. (2017) who showed that periods of high predictive NAO skill in the 20th century coincided with periods when the average NAO index was positive. They also showed that, towards the end of the 20th century, skill was higher for positive than for negative NAO years. Using the same 20th century ensemble, MacLeod et al. (2018) showed similar ensemble spread for positive and negative NAO years (their Figure 1b,c), but suggested that the predictions were underdispersive for the negative NAO years and overdispersive for the positive ones. Falkena et al. (2022) found that the regime frequency of positive NAO patterns is skilfully predicted by the forecasting system analysed in their study, but the regime frequency for negative NAO patterns was not.

6 | Effect of Signal-To-Noise Correction on Total Variability

Several previous studies have shown that the signal to noise ratio, or predictable component (PC), of the NAO in seasonal forecasting systems is low compared to observed estimates (e.g., Eade et al. 2014; Scaife et al. 2014; Siegert et al. 2016; Scaife and Smith 2018; Baker et al. 2018; Weisheimer et al. 2019). Using our large forecast ensemble, we calculate estimates of the signal and noise components of the NAO variability in this system. Eade et al. (2014) estimated the signal size simply using the temporal standard deviation of the ensemble mean, although they noted that this would be an over-estimate which would become more accurate with larger ensembles. In the current study, the temporal standard deviation of the ensemble mean is 2.6 hPa and the total standard deviation across all years and all members is 7.9 hPa.

Applying a standard analysis of variance consistent with Shukla et al. (2000) and Siegert et al. (2016), we have

$$\sigma^2_{total} = \sigma^2_{signal} + \sigma^2_{noise}$$

$$\sigma^2_{ens\ mean} = \sigma^2_{signal} + \frac{\sigma^2_{noise}}{N}$$

where N is the number of members per year. Rearranging these and eliminating the noise term, we estimate the signal size:

$$\sigma^2_{signal} = \frac{N\sigma^2_{ens\ mean} - \sigma^2_{total}}{N-1}$$

This new estimator for the signal variance represents an adjustment to the Eade et al. estimator and is robust to changes in ensemble size. For the current study, it gives an estimate of the signal standard deviation as 2.5 hPa and the average noise as 7.5 hPa, and therefore the PC ($\sigma_{signal}/\sigma_{total}$) for the forecast NAO is 0.31. Scaife and Smith (2018) estimated the model PC by correlating ensemble means with single member time series, and averaging over many such correlations. Applying their method to our ensemble yields an almost identical value to this new method (equal to three decimal places), although for smaller sample sizes it yields an underestimate.

Since the above calculation relies on using an ensemble, we obviously need a different approach to estimate the PC in the observations. One way, following Eade et al. (2014), is to use the correlation between observed NAO variability and the model ensemble mean, since this represents the proportion of the observed variability predicted by the model. This is likely an underestimate of the true observational PC due to model errors (Eade et al. 2014). Using our 0.46 estimate of the correlation limit as the observational PC we find, in agreement with the previous studies, that for this set of years the modelled PC is too low.

The uncertainty due to the choice of years is small for the model PC compared to the observational one: the 95% confidence interval obtained by resampling is (0.23, 0.37) for the model PC, compared with (0.06, 0.77) noted above for the correlation skill. 21% of these samples produce a larger PC estimate for the model than for the observations. So our result that the PC in the model is too small is not robust to all choices of years. The use of correlation skill as an estimate for the observational PC will be more appropriate in data sets where skill is high, because it relies on the assumption that the model describes the observational variability. Knight et al. (2022) argued that the low PC problem can be hidden in systems with low skill but become apparent as their skill improves. We therefore revisit this result in the next section, focusing on a skilful subset of years.

We now infer what this means for the comparison of modelled and observed total variance given above. This analysis gives insight into whether the signal to noise paradox is a result of a model PC that is too small, or overactive model unpredictable variability, or a combination of both. To increase the modelled PC ($\sigma_{signal} / \sigma_{total}$) to match the observed estimate we must either increase the modelled signal to 3.9 hPa or decrease the modelled noise to 4.9 hPa (or modify both by a smaller amount). Applying one or other correction would increase the modelled *total* variability to 8.5 hPa or decrease it to 5.5 hPa. Figure 4 illustrates how such an adjustment would affect our individual resampled realisations from Figure 2. In the upper panel, each standard deviation is simply adjusted to inflate its signal component to the value needed to correct the PC:

$$\sigma^2_{total*} = \sigma^2_{total} - \sigma^2_{signal} + \sigma^2_{signal*}$$

where the asterisk denotes the new values after adjustment. Although the signal size is almost doubled, the overall variability only sees a small shift, and the estimated observational value for this set of years is well within the new distribution. So, boosting the modelled signal to resolve the signal-to-noise paradox would retain an observationally consistent total variance. In contrast, the lower panel of Figure 4 shows the effect on the distribution of reducing the noise component of each standard deviation by the same factor:

$$\sigma^2_{total*} = \frac{\sigma^2_{noise*}}{\sigma^2_{noise}} \left(\sigma^2_{total} - \sigma^2_{signal} \right) + \sigma^2_{signal}$$

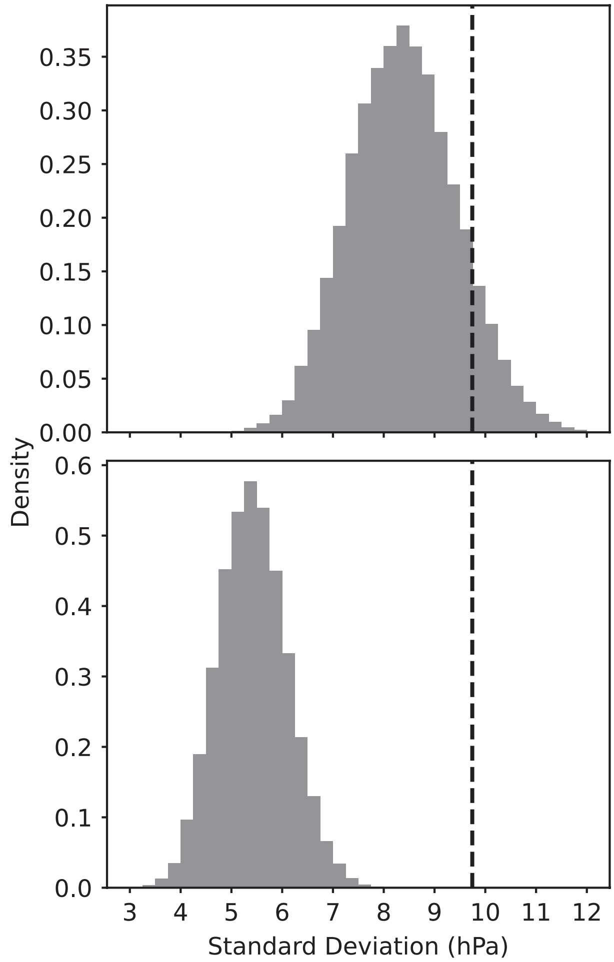

FIGURE 4 | Realistic ensemble variance results if the signal is increased but not if the noise is decreased. Distribution of standard deviations calculated by resampling GloSea to one member per year, but artificially increasing the signal (upper) or reducing the noise (lower) to make the model predictable component match the observed predictable component. Vertical dashed lines show the observational reanalysis value from the same years (1994–2016), lying at the 89th percentile of the model distribution in the top panel.

This reduction shifts the histogram so much that the observational estimate is completely outside the modelled distribution. So, only reducing the noise to rectify the PC would produce unrealistically low total variability in the ensemble. We therefore conclude that to address the signal to noise paradox without making the overall NAO distribution unrealistic, the modelled signal must be increased. Since the signal increase required to address the signal-to-noise paradox would not make the overall variability unrealistically large, it should be enough to *only* increase the signal. That is, we do not identify a need for mechanisms to reduce the noise at the same time as boosting the signal, such as the non-linear regime transition mechanisms discussed by Strommen and Palmer (2019).

This result differs from that of Eade et al. (2014) who proposed a calibration method that increased the modelled NAO signal to correct the PC and also reduced the modelled noise to match the total variability to their observational estimate. The difference comes partly from their use of the HadSLP observational data set (Allan and Ansell 2006), which is known to have low variance in some regions (Gillett et al. 2013, Supporting Information). For the years in the current study, we calculate an NAO standard deviation around 1 hPa lower from HadSLP than from ERA5. Eade et al. (2014) also reported a higher correlation estimate of 0.63, which is at the upper end of the confidence interval estimated by Siegert et al. (2016). This higher estimate for the observational PC would necessitate a larger signal increase which, without an accompanying reduction in noise, would bring the modelled variability above the low HadSLP estimate. To check the implications of these differences for the current study, we repeat the analysis in Figure 4 (upper) with the higher correlation estimate (not shown). We find the HadSLP standard deviation estimate lies at the 15th percentile of that shifted distribution. So the modelled total variability would be high compared with HadSLP, but not statistically significantly different, and our main result that it is enough to only increase the signal is unaffected.

7 | Prediction Skill and ENSO

It is well established that the NAO is influenced by the El Niño Southern Oscillation (ENSO; e.g., Brönnimann et al. 2007). It is therefore intuitive that both signal size and model skill should be larger for years when ENSO is active. Here we investigate the effect of ENSO on our forecast skill by splitting our hindcast set into two equal sets: one with active ENSO and one with neutral ENSO. To do this we calculate the Niño3.4 index from December–February means of the HadISST1.1 observational sea surface temperature dataset (Rayner et al. 2003) for our 1994–2016 hindcast period. We designate the 11 years with the largest magnitude Niño3.4 anomaly as ENSO years, and the 11 years with the smallest as neutral years. This approach is similar to early studies of potential seasonal predictability, which also assessed skill separately in ENSO and neutral years (Palmer et al. 2000; Branković and Palmer 2000; Shukla et al. 2000; Goddard and Dilley 2005).

We apply the procedure from the previous section to our two subsets and estimate that GloSea's signal size is 3.0 hPa for the ENSO years and 2.2 hPa for the neutral years (these estimates lie at the 92nd and 24th percentiles of a distribution created by calculating the signal from random 11-year GloSea samples). The difference in the correlation scores is illustrated in Figure 5, where the two black lines show correlation varying with ensemble size for the two data sets. The neutral years clearly show no skill by this measure. The ENSO years show increasing correlation with ensemble size, reaching 0.72 for the full 132 members. Although 11 years is a small sample, the correlation of 0.72 is significant at the 1% level according to a *t*-test. A similar result

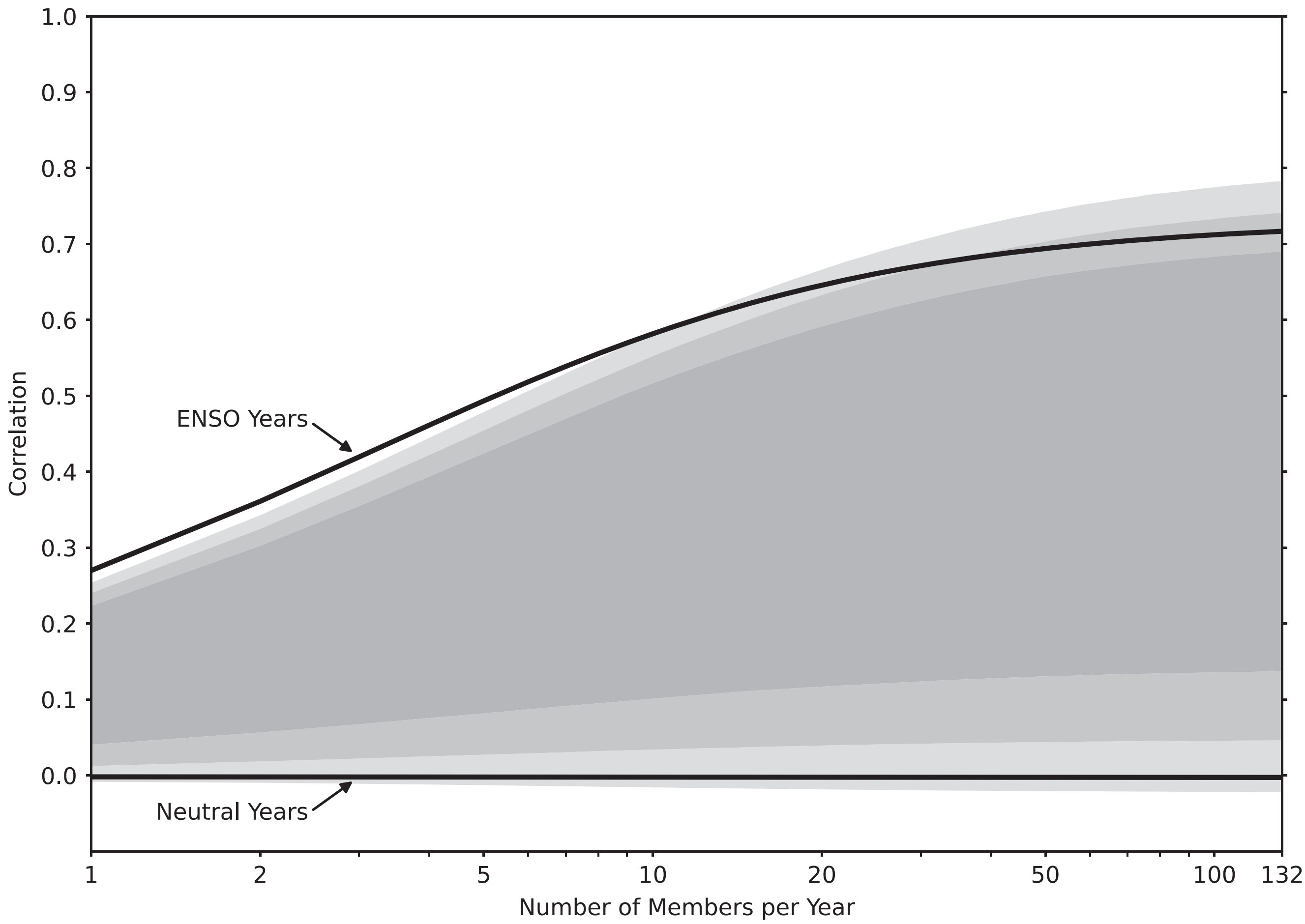

FIGURE 5 | ENSO years are much more skilful than non-ENSO years: Temporal correlation vs ensemble size for ENSO years and neutral years (black). Grey shading shows the 80%, 90% and 95% intervals obtained by using random sets of 11 years.

was found by Baker et al. (2024), who compared NAO skill in a 30-year multi-model ensemble. They noted an increased correlation (0.62) when discarding the 10 most neutral ENSO years and a decreased correlation (0.34) when discarding the 10 strongest ENSO years. Our result shows a larger difference because we do not have any common years in our two samples, and we illustrate the statistical significance of the result in Figure 5: the grey shading shows that the correlation for our neutral years is lower than that calculated from 95% of random 11-year subsets and the correlation for our ENSO years is higher than that calculated from 90% of random 11-year subsets.

Resampling the years within our 11-year ENSO active sample gives a 95% confidence interval of (0.58, 0.84) for the calculated correlation skill and (0.61, 0.86) for its theoretical limit, r_∞. So the skill uncertainty is lower for ENSO years than for the full hindcast. The confidence interval for the model PC is (0.28, 0.42). Using r_∞ as an estimate for the observational PC, we find that this is larger than the model PC in over 99% of samples. So, for ENSO years, the result that the model PC is too small is statistically significant at the 1% level.

ENSO teleconnection patterns are well simulated across seasonal prediction systems but the strength of those patterns is too weak (Garfinkel et al. 2022; Williams et al. 2023). Understanding the causes of these weak relationships should be a focus of future model development (Hardiman et al. 2022). Obviously, increasing the strength of the ENSO teleconnections would also increase the signal size for the ENSO years and hence the full hindcast set. Years with larger signals contribute more to the overall correlation (Weisheimer et al. 2019), so increasing the teleconnection strength should also increase the full hindcast correlation skill.

8 | Summary and Discussion

We analysed the interannual variability and ensemble spread of the NAO in the GloSea Global Seasonal forecasting system. By combining several hindcast sets to form a significantly larger ensemble than has been available for previous studies of this kind, we were able to give better estimates of the distribution of NAO characteristics and the quality of forecasts. We found that the correlation score calculated directly from our ensemble of 132 members per year is close to the theoretical limit obtained by applying the theory of Murphy (1990) so around 100 members is likely to be enough to maximise forecast skill.

Previous analyses (Eade et al. 2014; Scaife et al. 2014; Scaife and Smith 2018) found that the predictable component (the signal as a fraction of total variance) of the NAO in the forecast system is too small compared with observational estimates, and we have shown that this result is statistically significant for years when ENSO is active and therefore skill is high. We have further demonstrated that to correct the predictable component it is necessary to increase the signal size and that the implied linear increase in variance leads to total variance that remains consistent with observed variability, negating the need for nonlinear explanations that exchange variance between the ensemble mean and spread. Improving the strength of predictable teleconnections, for example from ENSO, is a plausible path forward for doing this.

The relationship shown in Figure 3 (left) implies that modifying the signal could also affect the ensemble spread: if we assume the relationship holds beyond the limits of the current data set then we would expect an increased signal to also increase the uncertainty in negative NAO years but decrease it in positive NAO years. However, this effect will be small: a doubling of the signal size would correspond to a mean NAO index for the most negative years changing from ~17 hPa to ~14 hPa. Extrapolating the relationship shown in Figure 3, this would correspond to a noise increase from ~8.3 hPa to ~9.0 hPa. So the signal-to-noise ratio would still be substantially increased.

Ongoing monitoring of smaller 21 member/year hindcast sets from the current operational GloSea6 system (Kettleborough et al. 2025) continues to show the low signal-to-noise problem, consistent with this study and the previous multimodel analysis of Baker et al. (2018). We therefore expect that the results found here are relevant to the current GloSea operational and other forecast systems.

In agreement with other recent studies (e.g., Hardiman et al. 2022; Williams et al. 2023; Baker et al. 2024) we conclude that future research should focus on understanding the weak teleconnections in the forecasting system. In the shorter term, operational systems may improve forecast skill by increasing ensemble size to around 100 members.

Author Contributions

Ruth E. Comer: methodology, software, data curation, investigation, formal analysis, writing – original draft, writing – review and editing, visualization. **Adam A. Scaife:** methodology, supervision, funding acquisition, writing – review and editing. **Jamie A. Kettleborough:** methodology, supervision, writing – review and editing. **Rowan T. Sutton:** methodology, writing – review and editing. **Philip J. Davis:** software, data curation, formal analysis.

Acknowledgements

The authors thank Craig MacLachlan, who originally proposed this project and contributed to early discussions.

Funding

This work was supported by the Met Office Hadley Centre Climate Programme (HCCP) and the UK Public Weather Service Programme, funded by the UK Department for Science, Innovation and Technology (DSIT).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

GloSea data used in this study can be downloaded from https://zenodo.org/records/17950763.

Endnotes

[1] Applying Hartigan's dip test (Hartigan and Hartigan 1985) to the ERA5 NAO data produces a p-value of 0.3, so we do not reject the null hypothesis that the distribution is unimodal.

References

Allan, R., and T. Ansell. 2006. "A New Globally Complete Monthly Historical Gridded Mean Sea Level Pressure Dataset (HadSLP2): 1850–2004." *Journal of Climate* 19, no. 22: 5816–5842. https://doi.org/10.1175/JCLI3937.1.

Athanasiadis, P. J., A. Bellucci, A. A. Scaife, et al. 2017. "A Multisystem View of Wintertime NAO Seasonal Predictions." *Journal of Climate* 30, no. 4: 1461–1475. https://doi.org/10.1175/JCLI-D-16-0153.1.

Baker, L. H., L. C. Shaffrey, S. J. Johnson, and A. Weisheimer. 2024. "Understanding the Intermittency of the Wintertime North Atlantic Oscillation and East Atlantic Pattern Seasonal Forecast Skill in the Copernicus C3S Multi-Model Ensemble." *Geophysical Research Letters* 51, no. 15: e2024GL108472. https://doi.org/10.1029/2024GL108472.

Baker, L. H., L. C. Shaffrey, R. T. Sutton, A. Weisheimer, and A. A. Scaife. 2018. "An Intercomparison of Skill and Overconfidence/Underconfidence of the Wintertime North Atlantic Oscillation in Multimodel Seasonal Forecasts." *Geophysical Research Letters* 45, no. 15: 7808–7817. https://doi.org/10.1029/2018GL078838.

Branković, Č., and T. N. Palmer. 2000. "Seasonal Skill and Predictability of ECMWF PROVOST Ensembles." *Quarterly Journal of the Royal Meteorological Society* 126, no. 567: 2035–2067. https://doi.org/10.1002/QJ.49712656704.

Brönnimann, S., E. Xoplaki, C. Casty, A. Pauling, and J. Luterbacher. 2007. "ENSO Influence on Europe During the Last Centuries." *Climate Dynamics* 28: 181–197. https://doi.org/10.1007/S00382-006-0175-Z.

Doi, T., S. K. Behera, and T. Yamagata. 2019. "Merits of a 108-Member Ensemble System in ENSO and IOD Predictions." *Journal of Climate* 32, no. 3: 957–972. https://doi.org/10.1175/JCLI-D-18-0193.1.

Dong, B., R. T. Sutton, and T. Woollings. 2011. "Changes of Interannual NAO Variability in Response to Greenhouse Gases Forcing." *Climate Dynamics* 37: 1621–1641. https://doi.org/10.1007/s00382-010-0936-6.

Eade, R., D. Smith, A. Scaife, et al. 2014. "Do Seasonal-To-Decadal Climate Predictions Underestimate the Predictability of the Real World?" *Geophysical Research Letters* 41, no. 15: 5620–5628. https://doi.org/10.1002/2014GL061146.

Falkena, S. K. J., J. de Wiljes, A. Weisheimer, and T. G. Shepherd. 2022. "Detection of Interannual Ensemble Forecast Signals Over the North Atlantic and Europe Using Atmospheric Circulation Regimes." *Quarterly Journal of the Royal Meteorological Society* 148, no. 742: 434–453. https://doi.org/10.1002/QJ.4213.

Garfinkel, C. I., W. Chen, Y. Li, C. Schwartz, P. Yadav, and D. Domeisen. 2022. "The Winter North Pacific Teleconnection in Response to ENSO and the MJO in Operational Subseasonal Forecasting Models Is Too Weak." *Journal of Climate* 35, no. 24: 8013–8030. https://doi.org/10.1175/JCLI-D-22-0179.1.

Gillett, N. P., J. C. Fyfe, and D. E. Parker. 2013. "Attribution of Observed Sea Level Pressure Trends to Greenhouse Gas, Aerosol, and Ozone Changes." *Geophysical Research Letters* 40, no. 10: 2302–2306. https://doi.org/10.1002/GRL.50500.

Goddard, L., and M. Dilley. 2005. "El Niño: Catastrophe or Opportunity." *Journal of Climate* 18, no. 5: 651–665. https://doi.org/10.1175/JCLI-3277.1.

Hardiman, S. C., N. J. Dunstone, A. A. Scaife, et al. 2022. "Missing Eddy Feedback May Explain Weak Signal-To-Noise Ratios in Climate Predictions." *npj Climate and Atmospheric Science* 5, no. 1: 1–8. https://doi.org/10.1038/s41612-022-00280-4.

Hartigan, J. A., and P. M. Hartigan. 1985. "The Dip Test of Unimodality." *Annals of Statistics* 13, no. 1: 70–84. https://www.rdocumentation.org/packages/diptest/versions/0.77-1/topics/dip.test.

Hersbach, H., B. Bell, P. Berrisford, et al. 2020. "The ERA5 Global Reanalysis." *Quarterly Journal of the Royal Meteorological Society* 146, no. 730: 1999–2049. https://doi.org/10.1002/QJ.3803.

Hilmer, M., and T. Jung. 2000. "Evidence for a Recent Change in the Link Between the North Atlantic Oscillation and Arctic Sea Ice Export." *Geophysical Research Letters* 27, no. 7: 989–992. https://doi.org/10.1029/1999GL010944.

Hurrell, J. W., G. D. Schmidt, A. N. Witt, et al. 1995. "Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation." *Science* 269, no. 5224: 676–679. https://doi.org/10.1126/SCIENCE.269.5224.676.

Kettleborough, J. A., P. Davis, R. E. Comer, et al. 2025. "Global Seasonal Forecasting System 6 (GloSea6): A Large Ensemble Seasonal Forecasting System." *Monthly Weather Review* 154: 39–57. https://doi.org/10.1175/MWR-D-25-0007.1.

Knight, J. R., A. A. Scaife, and A. Maidens. 2022. "An Extratropical Contribution to the Signal-To-Noise Paradox in Seasonal Climate Prediction." *Geophysical Research Letters* 49, no. 23: e2022GL100471. https://doi.org/10.1029/2022GL100471.

MacLachlan, C., A. Arribas, K. A. Peterson, et al. 2015. "Global Seasonal Forecast System Version 5 (GloSea5): A High-Resolution Seasonal Forecast System." *Quarterly Journal of the Royal Meteorological Society* 141, no. 689: 1072–1084. https://doi.org/10.1002/qj.2396.

MacLeod, D., C. O'Reilly, T. Palmer, and A. Weisheimer. 2018. "Flow Dependent Ensemble Spread in Seasonal Forecasts of the Boreal Winter Extratropics." *Atmospheric Science Letters* 19, no. 5. https://doi.org/10.1002/ASL.815.

Mason, S. J., and N. E. Graham. 2002. "Areas Beneath the Relative Operating Characteristics (ROC) and Relative Operating Levels (ROL) Curves: Statistical Significance and Interpretation." *Quarterly Journal of the Royal Meteorological Society* 128, no. 584 pt B: 2145–2166. https://doi.org/10.1256/003590002320603584.

Murphy, J. M. 1990. "Assessment of the Practical Utility of Extended Range Ensemble Forecasts." *Quarterly Journal of the Royal Meteorological Society* 116, no. 491: 89–125. https://doi.org/10.1002/QJ.49711649105.

Palmer, T. N., Č. Branković, and D. S. Richardson. 2000. "A Probability and Decision-Model Analysis of PROVOST Seasonal Multi-Model Ensemble Integrations." *Quarterly Journal of the Royal Meteorological Society* 126, no. 567: 2013–2033. https://doi.org/10.1002/QJ.49712656703.

Peterson, A. K., R. J. Greatbatch, J. Lu, H. Lin, and J. Derome. 2002. "Hindcasting the NAO Using Diabatic Forcing of a Simple AGCM." *Geophysical Research Letters* 29, no. 9: 50-1–50-4. https://doi.org/10.1029/2001gl014502.

Peterson, K. A., J. Lu, and R. J. Greatbatch. 2003. "Evidence of Nonlinear Dynamics in the Eastward Shift of the NAO." *Geophysical Research Letters* 30, no. 2: 36-1–36-4. https://doi.org/10.1029/2002GL015585.

Rayner, N. A., D. E. Parker, E. B. Horton, et al. 2003. "Global Analyses of Sea Surface Temperature, Sea Ice, and Night Marine Air Temperature Since the Late Nineteenth Century." *Journal of Geophysical Research: Atmospheres* 108, no. D14. https://agupubs.onlinelibrary.wiley.com/doi/epdf/10.1029/2002JD002670.

Scaife, A. A., A. Arribas, E. Blockley, et al. 2014. "Skillful Long-Range Prediction of European and North American Winters." *Geophysical Research Letters* 41, no. 7: 2514–2519. https://doi.org/10.1002/2014GL059637.

Scaife, A. A., and D. Smith. 2018. "A Signal-to-Noise Paradox in Climate Science." *npj Climate and Atmospheric Science* 1, no. 1: 1–8. https://doi.org/10.1038/s41612-018-0038-4.

Shukla, J., J. Anderson, D. Baumhefner, et al. 2000. "Dynamical Seasonal Prediction." *Bulletin of the American Meteorological Society* 81, no. 11: 2593–2606. https://doi.org/10.1175/1520-0477(2000)081<>2.3.CO;2.

Siegert, S., D. B. Stephenson, P. G. Sansom, A. A. Scaife, R. Eade, and A. Arribas. 2016. "A Bayesian Framework for Verification and Recalibration of Ensemble Forecasts: How Uncertain Is NAO Predictability?” *Journal of Climate* 29, no. 3: 995–1012. https://doi.org/10.1175/JCLI-D-15-0196.1.

Stephenson, D. B., V. Pavan, G. J. Boer, et al. 2003. “The North Atlantic Oscillation in Coupled Climate Models: A CMIP1 Evaluation.” *Climate Dynamics* 20, no. 4: 381–399. https://doi.org/10.1007/S00382-002-0281-5.

Stephenson, D. B., V. Pavan, M. Collins, M. M. Junge, and R. Quadrelli. 2006. “North Atlantic Oscillation Response to Transient Greenhouse Gas Forcing and the Impact on European Winter Climate: A CMIP2 Multi-Model Assessment.” *Climate Dynamics* 27, no. 4: 401–420. https://doi.org/10.1007/S00382-006-0140-X.

Strommen, K., and T. N. Palmer. 2019. “Signal and Noise in Regime Systems: A Hypothesis on the Predictability of the North Atlantic Oscillation.” *Quarterly Journal of the Royal Meteorological Society* 145, no. 718: 147–163. https://doi.org/10.1002/QJ.3414.

Thompson, V., N. J. Dunstone, A. A. Scaife, et al. 2017. “High Risk of Unprecedented UK Rainfall in the Current Climate.” *Nature Communications* 8, no. 1. https://doi.org/10.1038/s41467-017-00275-3.

Weisheimer, A., D. Decremer, D. MacLeod, et al. 2019. “How Confident Are Predictability Estimates of the Winter North Atlantic Oscillation?” *Quarterly Journal of the Royal Meteorological Society* 145, no. S1: 140–159. https://doi.org/10.1002/QJ.3446.

Weisheimer, A., N. Schaller, C. O'Reilly, D. A. MacLeod, and T. Palmer. 2017. “Atmospheric Seasonal Forecasts of the Twentieth Century: Multi-Decadal Variability in Predictive Skill of the Winter North Atlantic Oscillation (NAO) and Their Potential Value for Extreme Event Attribution.” *Quarterly Journal of the Royal Meteorological Society* 143, no. 703: 917–926. https://doi.org/10.1002/QJ.2976.

Williams, K. D., D. Copsey, E. W. Blockley, et al. 2018. “The Met Office Global Coupled Model 3.0 and 3.1 (GC3.0 and GC3.1) Configurations.” *Journal of Advances in Modeling Earth Systems* 10, no. 2: 357–380. https://doi.org/10.1002/2017MS001115.

Williams, N. C., A. A. Scaife, and J. A. Screen. 2023. “Underpredicted ENSO Teleconnections in Seasonal Forecasts.” *Geophysical Research Letters* 50, no. 5: e2022GL101689. https://doi.org/10.1029/2022GL101689.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** asl270008-sup-0001-Supinfo.docx.