Received: 2 May 2018 | Revised: 31 October 2018 | Accepted: 9 November 2018 | Published on: 22 February 2019

DOI: 10.1002/qj.3446

SUPPLEMENT ARTICLE

Quarterly Journal of the Royal Meteorological Society RMetS

# How confident are predictability estimates of the winter North Atlantic Oscillation?

Antje Weisheimer[1,2] | Damien Decremer[1] | David MacLeod[3] | Christopher O'Reilly[2,3] | Tim N. Stockdale[1] | Stephanie Johnson[1] | Tim N. Palmer[2,3]

[1] European Centre for Medium-Range Weather Forecasts (ECMWF), Research Department Reading, UK
[2]Department of Physics, National Centre for Atmospheric Science (NCAS), University of Oxford, Oxford, UK
[3]Department of Physics, University of Oxford, Oxford, UK

**Correspondence**
Antje Weisheimer, European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, RG2 9AX, UK.
Email: antje.weisheimer@ecmwf.int

**Funding information**
European Commission, 308378607085776613. Natural Environment Research Council, NE/M005887/1. EUCP, 776613. NERC NCAS, 607085. SPECS, 308378.

Atmospheric seasonal predictability in winter over the Euro-Atlantic region is studied with an emphasis on the signal-to-noise paradox of the North Atlantic Oscillation. Seasonal hindcasts of the ECMWF model for the recent period 1981–2009 show, in agreement with other studies, that correlation skill over Greenland and parts of the Arctic is higher than the signal-to-noise ratio implies. This leads to the paradoxical situation where the real world appears more predictable than the models suggest, with the forecast ensembles being overly dispersive (or underconfident). However, it is demonstrated that these conclusions are not supported by the diagnosed relationship between ensemble mean root-mean-square error (RMSE) and ensemble spread which indicates a slight under-dispersion (overconfidence). Furthermore, long atmospheric seasonal hindcasts suggest that over the 110-year period from 1900 to 2009 the ensemble system is well calibrated (neither over- nor under-dispersive). The observed skill changed drastically in the middle of the twentieth century and paradoxical regions during more recent hindcast periods were strongly under-dispersive during mid-century decades.

Due to non-stationarities of the climate system in the form of decadal variability, relatively short hindcasts are not sufficiently representative of longer-term behaviour. In addition, small hindcast sample size can lead to skill estimates, in particular of correlation measures, that are not robust. It is shown that the relative uncertainty due to small hindcast sample size is often larger for correlation-based than for RMSE-based diagnostics. Correlation-based measures like the RPC are shown to be highly sensitive to the strength of the predictable signal, implying that disentangling of physical deficiencies in the models on the one hand, and the effects of sampling uncertainty on the other hand, is difficult. Given the current lack of a causal physical mechanism to unravel the puzzle, our hypotheses of non-stationarity and sampling uncertainty provide simple yet plausible explanations for the paradox.

**KEYWORDS**
predictability of the NAO, seasonal forecasting

## 1 | INTRODUCTION

Understanding the predictability of extratropical circulation anomalies on seasonal time-scales is essential for making better decisions in societal sectors that rely crucially on weather and climate information for the seasons to come. Whereas our forecast models' abilities to predict the evolution of tropical sea-surface temperature (SST), especially in the El Niño–Southern Oscillation (ENSO) relevant Pacific Ocean, have improved over the recent decades and led to

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2018 The Authors. *Quarterly Journal of the Royal Meteorological Society* published by John Wiley & Sons Ltd on behalf of the Royal Meteorological Society.

demonstrated skill (Weisheimer *et al.,* 2009; Barnston *et al.,* 2012), forecasting the extratropical tropospheric circulation variability remains a major challenge in seasonal prediction systems (Weisheimer *et al.,* 2005; Lavers *et al.,* 2009; Kim *et al.,* 2012; Sigmond *et al.,* 2013a; Weisheimer and Palmer, 2014; Molteni *et al.,* 2015; Befort *et al.,* 2018; Beverley *et al.,* 2018; O'Reilly *et al.,* 2018a).

The main factor that constrains tropospheric seasonal forecast skill in the extratropics compared to the Tropics is related to the forecast signals being small while the noise levels of interannual variability are rather high, resulting in generally low signal-to-noise ratios. That is, the climatological distributions of seasonal climate anomalies are rather wide which makes skilful forecasts of interannual changes of such anomalies intrinsically difficult. Furthermore, our forecast models are not perfect and biases in their representation of the mean state, variability and regime structures can severely hamper predictability (Palmer and Weisheimer, 2011; Dawson *et al.,* 2012; Kumar *et al.,* 2012; Weisheimer *et al.,* 2014). In addition, since the main source of predictability on seasonal time-scales originates from ENSO as the dominant coupled mode of variability in the tropical Pacific, signals need to travel substantial distances around the globe, sometimes including interactions with the stratosphere (Bell *et al.,* 2009; Sigmond *et al.,* 2013b; Butler *et al.,* 2014; Scaife *et al.,* 2016), in order to impact the extratropics and especially Europe. Through interactions with the mean flow and wave disturbances on their way, the signals can become weakened and small displacements in terms of their phase relationship in the forecast models can easily result in lack of skill in remote areas.

The focus of this article is the predictability of the interannual variability of the winter North Atlantic Oscillation (NAO), the main mode of variability over the Euro-Atlantic region on a range of time-scales from days to seasons and longer. Teleconnections from ENSO, precursor temperatures of the North Atlantic Ocean, Arctic sea ice over the Barents-Kara Sea and the Quasi-Biennial Oscillation in the stratosphere are proposed physical mechanisms that can contribute to predictive NAO skill (Eade *et al.,* 2014; Scaife *et al.,* 2014; Dunstone *et al.,* 2016; Hansen *et al.,* 2017; Wang *et al.,* 2017; O'Reilly *et al.,* 2018b). Indeed, significant ensemble mean correlation skill of approximately $r \approx 0.6$ for the NAO (Scaife *et al.,* 2014) and its hemispheric counterpart of the Arctic Oscillation (AO: Derome *et al.,* 2005; Stockdale *et al.,* 2015; Kumar and Chen, 2018) were reported for some dynamical forecasting systems (see also Baker *et al.,* 2018).

Yet the predictive skill demonstrated in these studies comes with a paradox, or conundrum (Scaife and Smith, 2018): the level of actual forecast skill appears to be too high compared to the predictability one would expect for such forecasting systems, based on low signal-to-noise ratios. The expected correlation skill measure of a forecasting system is not independent of its signal-to-noise ratio, as demonstrated and discussed in Kumar (2009). In general, low signal-to-noise ratios are expected to also show low correlation skill. The implication of a situation with higher-than-expected skill is that the real world appears more predictable than the forecast model seems to suggest. What does this mean?

### 1.1 | Estimating predictability

The real-world predictability can be estimated by the correlation coefficient *r(ensmean,obs)* between ensemble-mean anomalies and observed (reanalysis) anomalies, computed over a retrospective forecast period. By definition, this means that the full ensemble is reduced to one deterministic measure that cannot represent the full spectrum of the forecast distribution. Estimates of the model predictability can be derived either by signal and noise analysis or by the so-called perfect model approach. The total variance $VAR_{\text{total}}$ of ensemble forecasts can be split into a signal variance $VAR_{\text{signal}}$ and a noise variance $VAR_{\text{noise}}$ with $VAR_{\text{total}} = VAR_{\text{signal}} + VAR_{\text{noise}}$. The model signal is simply the interannual variance of the ensemble mean ($VAR_{\text{signal}}$) and the model noise is the variance of the individual ensemble members about the ensemble mean ($VAR_{\text{noise}}$). The ratio of $VAR_{\text{signal}}/VAR_{\text{total}}$ can be interpreted as a measure of the intrinsic model predictability of the signal.

This notion led Eade *et al.* (2014) to define the so-called *Ratio of Predictable Components* (*RPC*) as the ratio between the predictable component of the real world and the predictable component of a model ensemble and estimated as follows:

$$RPC \geq \frac{r(ensmean, obs)}{\sqrt{VAR_{\text{signal}}/VAR_{\text{total}}}}. \quad (1)$$

If the model-based estimate of predictability is larger than the actual real-world predictability, the RPC will be smaller than one and indicative of overconfident forecasts. In such situations there is not enough variability in the ensemble, and the forecasts are said to be under-dispersive. Seasonal forecast models generally tend to be overconfident, especially in the Tropics, with not enough uncertainty included in their predictions (Weisheimer *et al.,* 2009; Ho *et al.,* 2013; Weisheimer and Palmer, 2014; Hao *et al.,* 2018). If, however, the model-based estimate of predictability is smaller than the real-world predictability, the RPC will become larger than one. Such a situation is described as underconfident with the model ensemble forecasts being overly dispersive, that is, exhibiting too much noise relative to the ensemble mean signal.

Another way of illustrating the signal-to-noise paradox involves the idea of constructing a hypothetical perfect model to estimate model predictability and interpreting the skill estimate from the perfect model as "potential skill" (Mehta *et al.,* 2000; Tang *et al.,* 2008; Kumar *et al.,* 2014; Jin *et al.,* 2017; L'Heureux *et al.,* 2017). Here the perfect model approach relies on the interchangeability of ensemble members with the observations if the underlying probability distribution of the true state is perfectly sampled by the model. Over a large number of forecasts, the statistical properties of the truth

are then identical to the statistical properties of any member of the ensemble. The perfect model uses the ensemble members as observations and a perfect model skill (potential skill) can be derived that serves as a guide for intrinsic model predictability.

The properties of a perfect model have several direct implications that allow us to test for the validity of the perfect model assumption in realistic forecasting systems. The RPC for a perfect model is exactly one (Eade *et al.,* 2014), that is, the squared correlation skill of the perfect model equals the variance ratio $VAR_{\text{signal}}/VAR_{\text{total}}$. Another consequence of a perfect model which is routinely used in operational numerical weather prediction to monitor the behaviour of the forecast ensemble is the fact that for a perfect model ensemble the root-mean-squared forecast error (*RMSE*) equals the mean ensemble standard deviation, or *spread* (Palmer *et al.,* 2006; Rodwell and Doblas-Reyes, 2006). We will make explicit use of these properties in the discussion of the results presented in this article.

Skill estimates based on seasonal retrospective forecasts, or hindcasts, are subject to various sampling uncertainties. The finite size of the ensemble impacts not only the estimated probabilities but also the ensemble-mean characteristic and associated skill measures (Richardson, 2001; Kumar, 2009; Scaife *et al.,* 2014; Siegert *et al.,* 2016). However, the largest sampling uncertainty for seasonal hindcasts is related to short hindcast period, typically only a few decades in length. Using NAO hindcasts from several seasonal prediction models performed over a 42-year period, Shi *et al.* (2015) were able to show that the RPC estimates and conclusions about a potential underconfidence of the forecasts strongly depend on the length of the hindcast period. The above-cited studies on the signal-to-noise paradox reported correlation skill of $r \approx 0.6$ for the NAO and AO based on hindcasts over recent periods of 20–35 years (that is, the verification sample size is 35 winter seasons at maximum); and even while the reported correlations are found to be statistically significant, the error bars around correlation estimates for small samples can be substantial. Anscombe's quartet (Anscombe, 1973) famously demonstrated the risk of misinterpreting descriptive statistics for small sample sizes. Similarly, the relationship between expected correlation skill and signal-to-noise will only be realised in the limit of very large sample sizes, or long hindcast periods (Kumar, 2009). In addition, non-stationarity in the form of decadal variability of the atmospheric dynamics and related longer-term variations in forecast skill (Derome *et al.,* 2005; O'Reilly *et al.,* 2017; Weisheimer *et al.,* 2017; Kumar and Chen, 2018) is an additional factor that can further contribute to sampling uncertainty and make robust statements about the over- or underconfidence of a seasonal forecasting system challenging.

In this article, we study the atmospheric predictability in the Euro-Atlantic region in different versions of the ECMWF seasonal forecast model, specifically focussing on the above-mentioned signal-to-noise paradox of the NAO. The issue of over-dispersion is revisited in different forecasting systems using different forecasting metrics. Non-stationarity of the time series, sampling uncertainties and the impact of weak signals are also analysed.

Section 2 discusses the performance of the model for hindcasts over a recent period. It is found that the correlation skill measures indicate underconfidence in parts of the North Atlantic and over Greenland, whereas the *RMSE* and ensemble spread do not confirm this result and show a slight overconfidence instead. In section 3 we investigate the model behaviour for independent hindcast periods in the past, based on a set of hindcasts that cover the entire twentieth century. We find evidence for a non-stationary character of the seasonal predictability estimates, including the signal-to-noise paradox, of the North Atlantic atmospheric flow on time-scales of several decades. Section 4 describes the contributions from individual winters in the 110-year-long hindcast set to the NAO paradox and discusses the characteristics of the atmospheric flow that lead to the unexpected behaviour during recent forecast periods. In section 5 we demonstrate the impact on the predictability estimates of small sample sizes due to the number of ensemble members and hindcast years. We argue that the relative uncertainty due to small hindcast sample sizes is, under most circumstances, larger for the correlation measures than it is for the *RMSE*. The hypotheses of non-stationarity and sampling uncertainty provide a simple yet plausible explanation for the apparently conflicting findings of under- and overconfidence over parts of the North Atlantic. Section 6 summarises the results and presents some conclusions.

## 2 | PREDICTABILITY ESTIMATES OF THE EURO-ATLANTIC SECTOR IN RECENT DECADES

In this section we examine how well the observed winter circulation over the Euro-Atlantic region in the recent past was predicted in seasonal forecasts performed with three different configurations of the ECMWF model. We will assess the evidence for the signal-to-noise paradox in these simulations and whether the real world might be more predictable than the model suggests. We consider both metrics of dispersion, *RPC* and spread vs. *RMSE*, after first assessing the forecast skill as correlations with observations and in the perfect model scenario.

### 2.1 | Model simulations

ECMWF is a world-leading numerical weather prediction centre and as part of its seamless forecasting strategy also runs operational seasonal forecasts. Their latest seasonal forecasting system (SEAS5: Johnson *et al.,* January 2018; Stockdale *et al.,* 2018) became operational in November 2017 and replaced the previous system (System 4: Molteni

*et al.,* 2011) which had been the operational system during the last six years. Both System 4 and SEAS5 are based on the Integrated Forecast System (IFS) atmospheric component coupled to the Nucleus for European Modelling of the Ocean (NEMO) ocean model. The atmospheric resolutions are T255L91 and Tco319L91, respectively, which correspond to approx. 80 and 36 km horizontally. The resolution of the ocean model increased from 1° and 42 layers in System 4 to $^1/_4$° and 75 layers in SEAS5. In addition to several advances in the atmospheric and ocean model components compared to System 4, the new SEAS5 is also coupled to the dynamic Louvain-la-Neuve sea-Ice Model (LIM2).

The third configuration of the seasonal forecasts which we are going to analyse is based on an IFS atmospheric model cycle between System 4 and SEAS5 and is run at the same resolution as System 4. It is called ASF-20C (Atmospheric Seasonal Forecasts of the 20th Century: Weisheimer *et al.,* 2017) and is an atmospheric seasonal hindcast experiment covering the 110-year period from 1900 to 2010. It uses prescribed sea-surface temperatures (SST) as a lower boundary condition over the ocean. The atmospheric initial conditions for ASF-20C were derived from ECMWF's atmospheric reanalysis of the twentieth century ERA-20C (Poli *et al.,* 2016). As lower boundary conditions, SSTs from the Hadley Centre global sea-Ice and Sea-Surface Temperature coverage (HadISST2) dataset were used, similarly to ERA-20C. The experiment comprised 4-month long seasonal forecasts initialised on 1 November 1900 to 2009 using an ensemble of 51 perturbed members.

Here we compare the performance of System 4, SEAS5 and ASF-20C over the common hindcast period 1981–2009. Here the year corresponds to the start date of the forecasts for that winter, for example, 1981 denotes the December–February (DJF) season of 1981/1982. We focus our analysis on the mid-tropospheric geopotential height field anomalies at 500 hPa (Z500) for DJF means of all 29 starts dates of 1 November during the hindcast period. For verification the ERA-Interim reanalysis has been used. All forecasts are issued as ensembles and we analyse 25 ensemble members for each forecasting system (note that System 4 and ASF-20C have a total of 51 hindcast members and we have used only the first 25). Anomalies are defined with respect to the climatological mean state over the hindcast period for the reanalysis and the model ensemble hindcasts separately which implies a linear bias correction of the model data.

### 2.2 | Forecast performance

Figure 1 shows maps of the anomaly correlation coefficient ACC (left) and the anomaly correlation coefficient for the perfect model ACP (right). The ACC is the temporal correlation of the ensemble mean anomalies and observed anomalies, while the ACP correlates the ensemble mean anomalies with anomalies of each ensemble member assuming it is an interchangeable substitute for observations. Here, the ACP has been estimated from averaging the correlations obtained by correlating all individual continuous ensemble members with the ensemble mean of all members (including the verifying member, see discussion in section 5.1). The results are very robust against details in the randomisation process of all possible combinations of ensemble members across start dates. Correlation skill results are shown for System 4 (top), SEAS5 (middle) and ASF-20C (bottom). The white lines indicate where the correlation skill becomes significant at the 95% confidence level. Here, statistical confidence is evaluated using a Monte-Carlo technique, the so-called counting norm bootstrap (Livezey and Chen, 1983; Zwiers, 1987; Wilks, 1996).

As a rough guide as to where the centres of action of the NAO are, the two black dots show the geographical location of the dipole of the first empirical orthogonal function (1st EOF) of Z500 over the North Atlantic sector which will be used in section 3 to define the NAO index. While the northern centre of action near the tip of Greenland can be well defined as a point in the 1st EOF, characterizing the subtropical part of the dipole with a single point is, however, less meaningful due to the elongated structure bending across large parts of the North Atlantic of the southern centre of action.

In all simulations the Euro-Atlantic area extending from the central North Atlantic to 40°E and from the Mediterranean Sea to Scandinavia is characterized by lack of skill when correlated with observations. The simulations with ASF-20C in Figure 1e also see a ridge of higher skill between 40°N and 50°N extending from the western and central North Atlantic into the eastern North Atlantic to the Bay of Biscay which might be due to the prescribed SSTs over the Gulf Stream area. Parts of Siberia suffer from a lack of skill. North America sees in general higher levels of skill than Europe and Asia, which are statistically significant.

Greenland and the Arctic further to the east are areas of high and significant skill with a strong gradient and decline in skill from the east coast of Greenland through the Denmark Strait and Iceland to the British Isles. This part of the North Atlantic is, of course, the geographic area of one of the two centres of action of the NAO.

In contrast to these apparent longitudinal variations in real-world skill, the skill estimates from the perfect model approach show a more uniform and smooth pattern of moderate skill across the North Atlantic, Europe and Asia. This means that the model estimate of predictability across the eastern North Atlantic–European region is higher than the real-world predictability in that region. For Greenland and surrounding areas, however, the perfect model skill is lower than the rather high level of observed skill leading to the paradox described in the Introduction where the real world appears more predictable than the models suggest. In ASF-20C a narrow ridge of observed skill across the subtropical North Atlantic is not matched in the potential skill. Similar to the observed correlation maps, the perfect model also indicates higher intrinsic levels of predictability for North America,

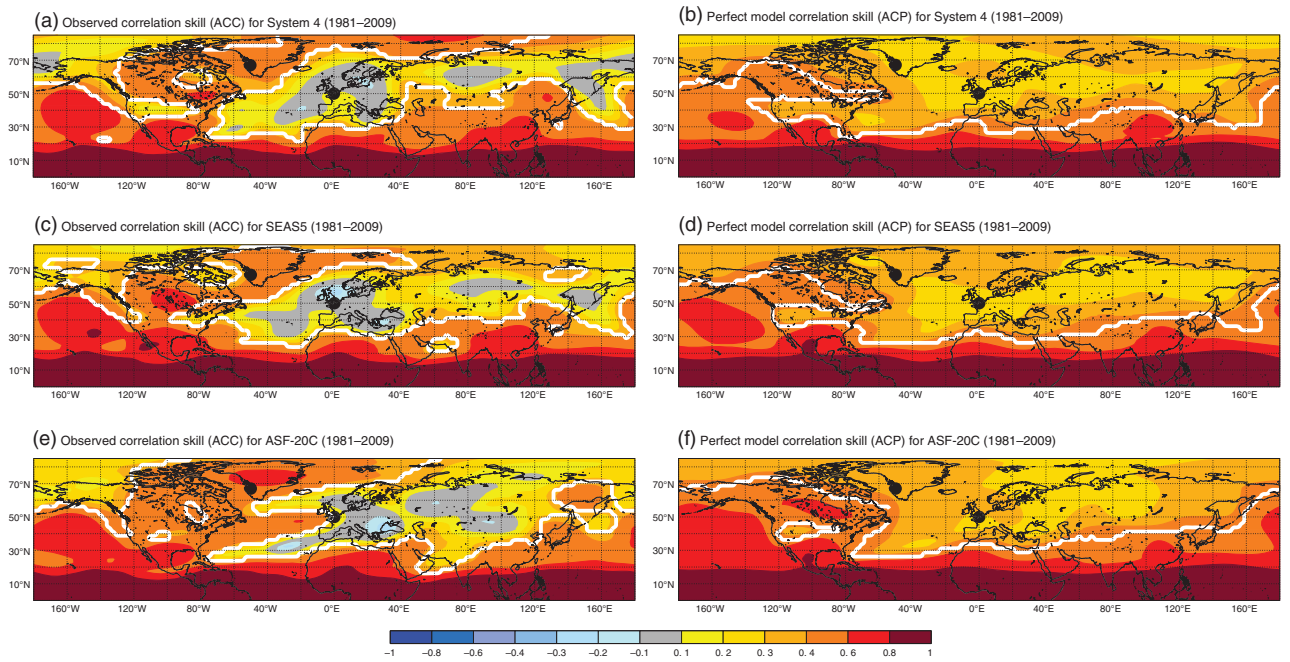


**FIGURE 1** Anomaly correlation between the ensemble mean and observations (a,c,e), and perfect model anomaly correlation between the ensemble mean and each ensemble member (b,d,f). System 4 (a,b), SEAS5 (c,d) and ASF-20C (e,f). The white lines indicate where the correlation skill becomes significant at the 95% confidence level. The two black dots annotate the centres of action of the NAO based on the 1st EOF of Z500 in the Euro-Atlantic region, see text for details. Valid for Z500 forecasts of DJF initialised on 1 November over the period 1981–2009 using ERA-Interim as verification

due to stronger tropical Pacific teleconnections in these areas (L'Heureux *et al.,* 2017; O'Reilly *et al.,* 2017). A comparison of the observed skill to the perfect model potential skill is presented in the right column of Figure 2 as the ratio between the observed and potential skill. It is worth mentioning that qualitatively similar results are found nearer to the surface for sea-level pressure and also higher up in the atmosphere.

The overall spatial structures of actual and potential skill are rather similar in the two versions of the operational coupled model and the uncoupled atmospheric model, lending support to the hypothesis that the uncoupled hindcasts do not perform substantially differently from the coupled hindcasts in terms of Z500 predictability of the Northern Hemisphere (NH) extratropics. The fact that all three configurations of the ECMWF model agree in their general patterns of observed and perfect model skill indicates little sensitivity to atmospheric and oceanic model resolution and whether or not the system uses an interactive ocean and sea-ice model.

The Ratio of Predictable Components (*RPC*) diagnostics shown at the left of Figure 2 reflect the main findings of Figure 1. The *RPC* diagnostics and the observed versus perfect model correlation analysis only differ in their estimation of the model-based predictability as the real-world predictability estimates come in both cases from the correlation of the ensemble mean with observations, *r(ensmean,obs)*. The model-based predictability in the case of the *RPC* relies on the signal-to-noise variance ratios in the model, see Equation (1), and for the perfect model approach it is the expected value of the correlation of the ensemble mean with any ensemble member. The *RPC* for most of the NH extratropics is below or close to one for all three forecasting configurations, indicating overconfident (under-dispersive) forecasts. A pronounced minimum is noticeable over Europe and the eastern North Atlantic, in agreement with areas of no or even negative skill (the white lines indicate areas of significant correlation skill). In contrast, parts of the Arctic and especially Greenland show *RPC* values larger than 1 and up to 2, reflecting the high levels of observed skill in this region.

The ratios between observed and potential skill in the right column of Figure 2 give a very similar picture to the *RPC* diagnostics in the left column. These plots clearly show the consistency in all three simulations of the predictability paradox being limited to mainly over Greenland. Figure 2 resembles in its general structure over the North Atlantic the *RPC* for similar seasonal forecasts with the Met Office model GloSea5 shown in figure 1b of Eade *et al.* (2014), even though the forecasting systems described here tend to be a little less skilful in predicting the NAO index over the common hindcast period.

The correlation measures of interannual variability for a perfect model versus observed skill and the *RPC* in Figures 1 and 2 indicate that the forecasting systems are underconfident over Greenland and parts of the Arctic Ocean and also that the models underestimate the predictability in these regions. As described in the Introduction, the mathematical property of a perfect model that the mean ensemble-mean forecast *RMSE* equals the mean ensemble spread as measured by the ensemble standard deviation (Palmer *et al.,* 2006; Rodwell and Doblas-Reyes, 2006) can be used as an alternative to the correlation-based measures discussed above in order to diagnose any under- or over-dispersive behaviour.

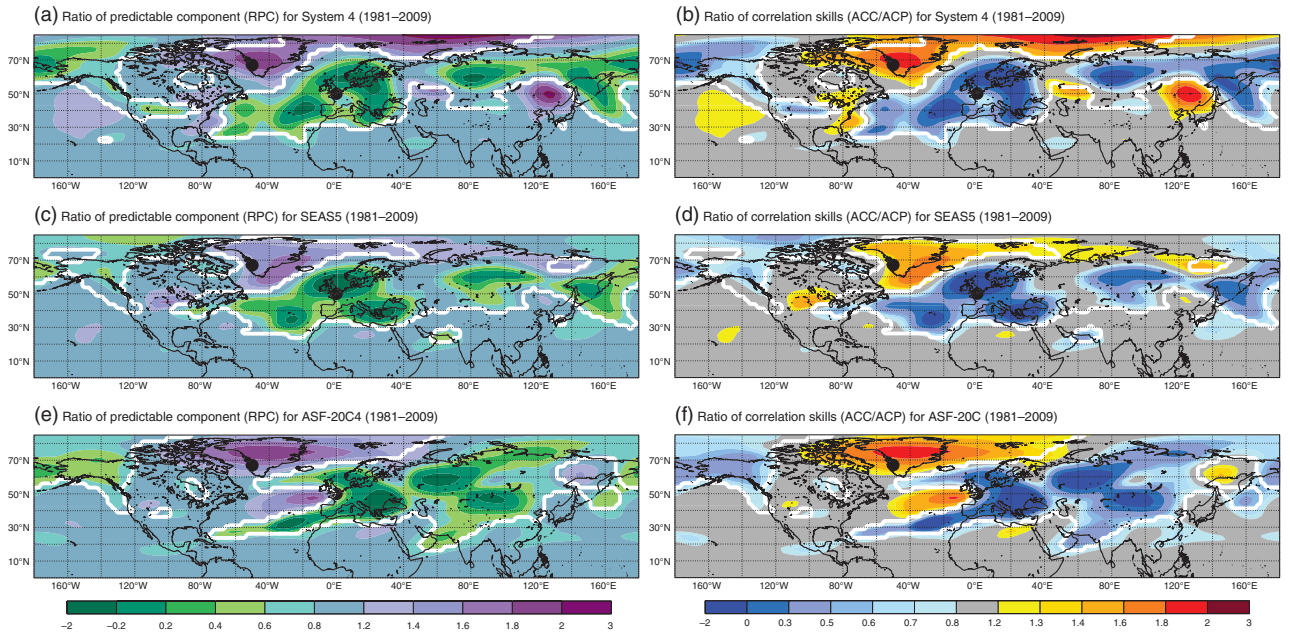


**FIGURE 2** Ratio of predictable components (*RPC*) (a,c,e), and ratio between observed and potential skill (b,d,f), for System 4 (a,b), SEAS5 (c,d) and ASF-20C (e,f). The white lines indicate where the observed correlation skill becomes significant at the 95% confidence level, see Figure 1. Valid for Z500 forecasts of DJF initialised on 1 November over the period 1981–2009 using ERA-Interim as verification

In Figure 3 we show the spatial structure of the mean ensemble spread (left) and the *RMSE* (right) for System 4, SEAS5 and ASF-20C. As can be seen, the two quantities agree very well in general terms of their geographical structure for all three simulations. Figure 4 displays the ratio between ensemble spread and RMSE. With values around 1 it clearly indicates that for ASF-20C the spread matches the *RMSE* very well over the North Atlantic. The behaviour for System 4 and SEAS5 is similar. An area of increased errors over the North Atlantic can be seen which is accompanied by larger spread of the ensemble, giving a correct indication of the flow-dependent uncertainty in these areas. It should be noted that the *RMSE* tends to be slightly larger than the spread in regions to the east and south of Iceland. Contrary to the findings for the correlation-based measures, such a behaviour is characteristic of an under-dispersive (overconfident) forecasting system that does not produce enough ensemble spread to balance the forecast errors. This is in very good agreement with the typical behaviour of the ECMWF ensemble for short- and medium-range forecasts (Rodwell *et al.,* 2018). Recent analyses of sub-seasonal forecasts of up to 45 days from a variety of modelling systems across the world lend further support for the argument that long-range ensemble forecasts over the North Atlantic tend to be under-dispersive, or overconfident, if measured by the *RMSE* vs. spread behaviour (L. Ferranti, personal communication 2017). A more detailed analysis of the flow dependence of ensemble spread in the ASF-20C seasonal hindcasts also concludes that the spread and *RMSE* over the North Atlantic match reasonably well (MacLeod *et al.,* 2018).

These findings raise a number of questions: What are the characteristics of the atmospheric flow that lead to the unexpected underconfident behaviour found when diagnosing the prediction of interannual variability using correlation-based measures? What could be the reasons for the apparently conflicting findings of under- and overconfidence over parts of the North Atlantic, depending on whether the problem is looked at by interannual correlation-based measures or by the statistical relation between spread and forecast error?

Our working hypothesis here is twofold: (a) due to the non-stationary character of the North Atlantic atmospheric flow on time-scales of several decades (e.g. Woollings *et al.,* 2014), short hindcast periods are not necessarily representative of the longer-term behaviour of the climate system with distinct patterns of multi-decadal variability, and (b) small sample sizes of the ensemble members and hindcast length can lead to estimates, particularly in correlation, that are not robust. The resulting uncertainties from (a) and (b) pose a serious limitation on the interpretation of predictability estimates of the real world versus the model world and statements of over- or underconfidence of seasonal forecasting systems. In the following we are going to discuss (a) and (b) in sections 3 and 5, respectively.

## 3 | DECADAL FLUCTUATIONS OF NAO PREDICTABILITY

The North Atlantic Oscillation as the main mode of atmospheric variability over the North Atlantic varies on a spectrum of time-scales reaching from days to decades and longer (Woollings *et al.,* 2014). For example, the mid-twentieth century was characterized predominantly by the negative phase of the NAO, while a strong positive trend starting in the 1960s and lasting for approx. three decades led to a shift towards positive NAO for the majority of

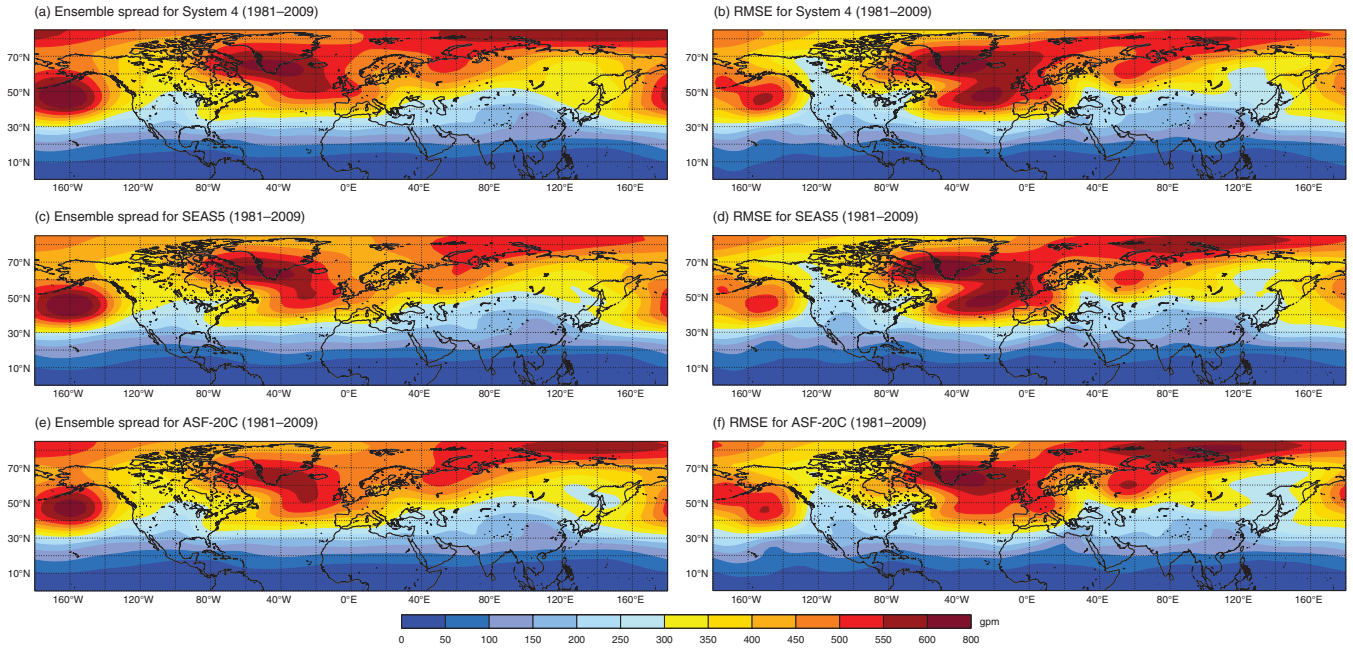


**FIGURE 3** Mean ensemble standard deviation (a,c,e), and mean ensemble-mean forecast *RMSE* (b,d,f). System 4 (a,b), SEAS5 (c,d) and ASF-20C (e,f). Valid for Z500 forecasts of DJF initialised on 1 November over the period 1981–2009 using ERA-Interim as verification

winters during the more recent decades. What are the implications of such low-frequency variations for seasonal NAO predictability estimates?

Traditionally, seasonal hindcasts are performed for the previous 20–30 years with the aim to help calibration and derive skill and predictability estimates as guidelines for the performance of the forecasting system in operational conditions in the future. As such, the standard hindcasts are, by design, not able to capture any fluctuations of the large-scale atmospheric circulation that occurred on longer time-scales than the hindcast period, and thus cannot provide guidance about the performance of the forecasting system if the climate system underwent significant decadal fluctuations in the future. Specifically, hindcasts from the 1980s and later provide skill information during a period of mainly positive NAO winters. They cannot, however, robustly estimate the predictability of the NAO when dominated by its negative phase (e.g. during the mid-twentieth century). Arguably, the ultimate aim of any skilful climate prediction system on seasonal and longer time-scales should include the ability to correctly represent potential sensitivities of the predictability estimates to changes in the background state that occur on longer time-scales. Whether these changes can be linked to a response to external forcings or are purely a result of internal variations of the climate system, is not of direct importance for this question. It is, however, important to test seasonal forecast models for conditions that differ substantially from the current state of large-scale atmospheric circulation in order to improve our confidence in the underlying physical mechanisms that lead to predictability on seasonal lead times in the models.

In a first approach to explore how seasonal forecast skill varies over much longer hindcast periods, Weisheimer *et al.* (2017) performed the ASF-20C (Atmospheric Seasonal Forecasts of the 20th Century) experiment covering the 110-year period from 1900 to 2010. As demonstrated in section 2, the ASF-20C experiments performs comparably to System 4 and SEAS5 with regards to winter mid-tropospheric flow over the Atlantic–European region.

Figure 5a shows the temporal evolution of forecast correlation skill of ASF-20C to predict the DJF mean NAO index over the 110-year hindcast period. Here, the NAO index is defined as the time series of projecting the geopotential height of the 500 hPa level (Z500) on the leading empirical orthogonal function over the Atlantic sector (for details see Weisheimer *et al.,* 2017). The correlation skill in Figure 5a is computed for 30-year windows which are moved by 1 year. That is, the last data point plotted at 1996 shows the ensemble mean correlation with ERA-20C during the period of hindcasts started in 1980 to 2009. The black solid curves show the estimated correlation coefficients for the real world whereas the dotted grey curves show the perfect model correlation skill where the ensemble members have been used as verification data. The thin lines around the bold curves indicate the 90% confidence range and indicate sampling uncertainty. As was discussed in detail in Weisheimer *et al.* (2017), ERA-20C exhibits multi-decadal variability of predictive skill with relatively high levels of skill during recent decades and also during the earlier decades of the twentieth century. However, the mid-century period 1950s–1970s was characterised by levels of lower skill. Whilst these inter-decadal differences in skill are, by themselves, only marginally statistically significant, the variations in skill strongly co-vary with statistics of the general circulation itself (examples of which can be found, e.g. Minobe, 1997; Hoerling *et al.,* 2001; Fletcher and Saunders, 2006; Greatbatch and Jung, 2007; Douville *et al.,*

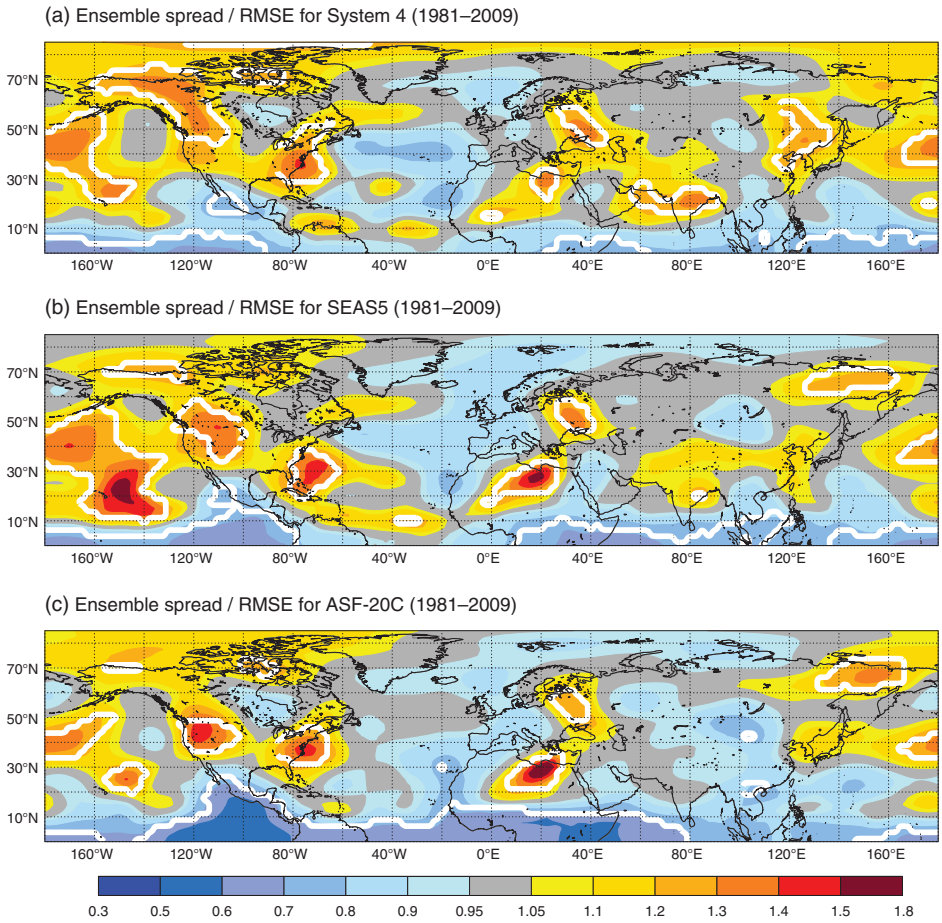


**FIGURE 4** Ratio of ensemble spread to *RMSE*. System 4 (a), SEAS5 (b) and ASF-20C (c). Valid for Z500 forecasts of DJF initialised on 1 November over the period 1981–2009 using ERA-Interim as verification. The white lines indicate where the ratio is significantly different from 1 at the 95% confidence level

2017; Hegerl *et al.,* 2018; Huang *et al.,* 2018) suggesting that such differences are indeed physically based. In particular, the temporal skill evolution of the NAO co-varies with changes in the skill of the Pacific–North America (PNA) pattern that exhibit a high statistical significance and relate to changes in the ENSO–North Pacific teleconnections (O'Reilly *et al.,* 2017; O'Reilly, 2018).

The variance ratio (see denominator in Equation (1)) is relatively constant over time during the second half of the century (not shown) so that the *RPC* variability is dominated by changes in the anomaly correlation. Figure 5b shows a time series of moving window *RPC* values (solid black). The *RPC* for the perfect model (dotted grey) is exactly one for all times, by definition and empirically confirmed. The *RPC* for the latest 30-year window is close to 1.5, in agreement with the results from the correlation analysis of Z500 over the North Atlantic region as shown in Figures 1 and 2. Similar *RPC* values above one occur for hindcasts that cover periods after approx. 1960. During the mid-century period when the correlation skill is minimal, the *RPC* is on average below one. In the early half of the century *RPC* fluctuates somewhat around one. Figure 5b demonstrates that the *RPC*, similar to the correlation skill, undergoes multi-decadal variations with periods indicative of underconfidence during the more recent decades and periods with evidence for overconfidence in the middle of the century. The *RPC* computed over the entire 110-year hindcast period is 1.02.

In order to characterize the geographical variations of skill and *RPC*, we define three 29-year non-overlapping periods based on the time series in Figure 5 which represent the first half of the century (1912–1940), the mid-century period (1942–1970) and the more recent decades (1981–2009). These periods were chosen to represent epochs of relatively high and low skill. Figure 6 shows the spatial structure of the observed and perfect model correlation skill for Z500 during DJF for these three characteristic periods and the full hindcast period, using ERA-20C as verification. The correlations during the latest period are very similar to the results shown in Figure 1e,f (which use ERA-Interim as verification) with larger observed skill than in the perfect model over Greenland and parts of the Arctic, while the observed skill over most of Europe is zero or negative and smaller than the perfect model estimate. During the mid-century period the picture of the observed skill has completely changed. Greenland and parts of the North Atlantic show negative correlation skill while the perfect model indicates a much smaller reduction of skill over these areas. That means that regions which appear more predictable in the real world than in the model over the most recent few decades (i.e. paradoxical or underconfident), were strongly overconfident during the mid-century decades. The

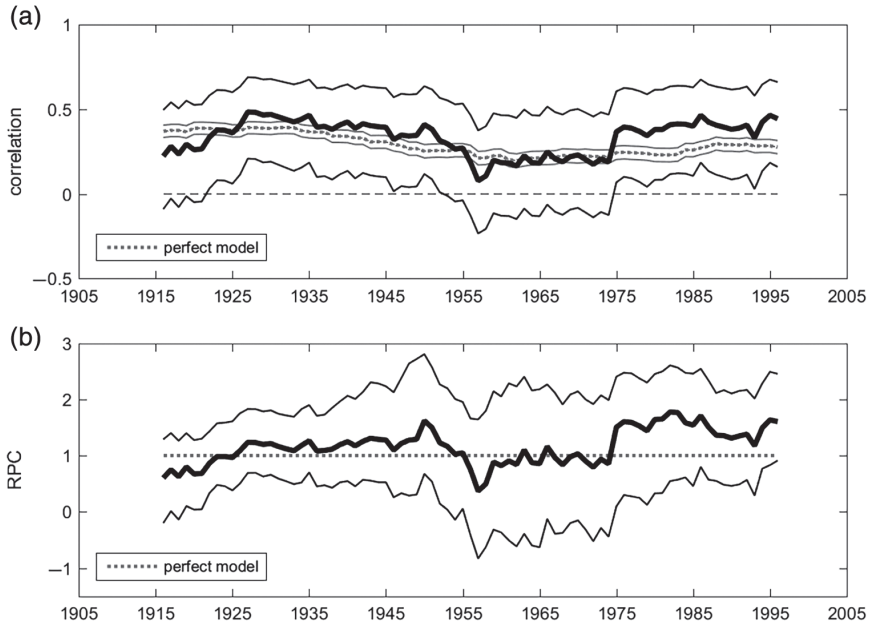


**FIGURE 5** (a) Anomaly correlation coefficient of the DJF NAO index between the ensemble mean ASF-20C and ERA-20C (solid black) and for a perfect model ensemble where the ensemble members are used as verification (dotted grey) over the period 1900–2009 computed for moving 30-year windows by 1 year. Thinner lines indicate the 90% confidence intervals. (b) Ratio of predictable components (*RPC*, solid black) and *RPC* for a perfect model (dotted grey). Thinner lines indicate the 90% confidence intervals

suggested potential skill level of the model was much higher than in the real world (Figure 6e,f). In the earlier parts of the twentieth century the situation appears less extreme in either direction; slightly higher values of perfect model skill are found over Greenland and the northern North Atlantic and moderate observed skill levels over most of the North Atlantic except for the northeastern parts. The last row of Figure 6 shows the correlation maps for the full 110-year period. Observed skill over the North Atlantic is mostly significant. The perfect model skill displays a smooth pattern of high skill in low latitudes and lower skill in higher latitudes with a ridge of relatively high skill over the Pacific and North America and a trough of relatively low skill over a wide area of the North Atlantic, Europe and northern Asia. For almost all areas, ACP > ACC except for some small-scale regions over the southern part of Greenland and southwest of the British Isles.

The corresponding geographical structure of the *RPC* is displayed in the left column of Figure 7 and confirms a substantial difference in the behaviour of the predictable components over the North Atlantic and Greenland between the more recent decades and the mid-century period. Almost all of the Northern Hemisphere during the mid-century period has *RPC* values substantially below one (see also O'Reilly, 2018). In the earlier period of the century the Arctic sticks out as a region with small *RPCs* (overconfidence), with large areas of the North Atlantic being in the range of a perfect value around one. Large *RPC* values during that period are observed over parts of Siberia in regions that were characterized by rather small *RPC* values during the subsequent two periods. Over the full hindcast period the *RPC* is dominated by values around one or smaller, indicative of under-dispersive ensembles.

The right column in Figure 7 shows the ratio between the mean ensemble spread and the mean *RMSE* for the three climate periods as an alternative measure of the model's dispersive behaviour. The statistical significance of the ratio being different from 1 at the 95% confidence level has been estimated using a counting norm bootstrap approach (Livezey and Chen, 1983; Zwiers, 1987; Wilks, 1996). As discussed in Figure 3, the ensemble spread and *RMSE* forecast error match, on average, very well over the Euro-Atlantic region including Greenland for the latest period. Some variation in the ratio can also be detected during the other representative periods but these variations are generally smaller than for the correlation-based measures (see also section 5). These temporal variations are in agreement with the findings of MacLeod *et al.* (2018) who show in their Fig. 2a the temporal evolution from 1900 to 2009 of the RMSE and ensemble spread of the NAO index in the ASF-20C simulations. The latest period clearly shows no sign of underconfidence in the North Atlantic region. For the early period there are some patches across the wider Euro-Atlantic region where the *RMSE* is smaller than the ensemble spread, for example, over the Iberian Peninsula and Scandinavia. In agreement with the *RPC* for the full period (Figure 7g), the relationship between ensemble spread and *RMSE* over the 110 years indicates an overall well balanced ensemble (Figure 7h).

## 4 | SKILL CONTRIBUTIONS OF INDIVIDUAL EXTREME WINTERS

To better understand the atmospheric flow patterns associated with the reported unexpectedly high levels of correlation

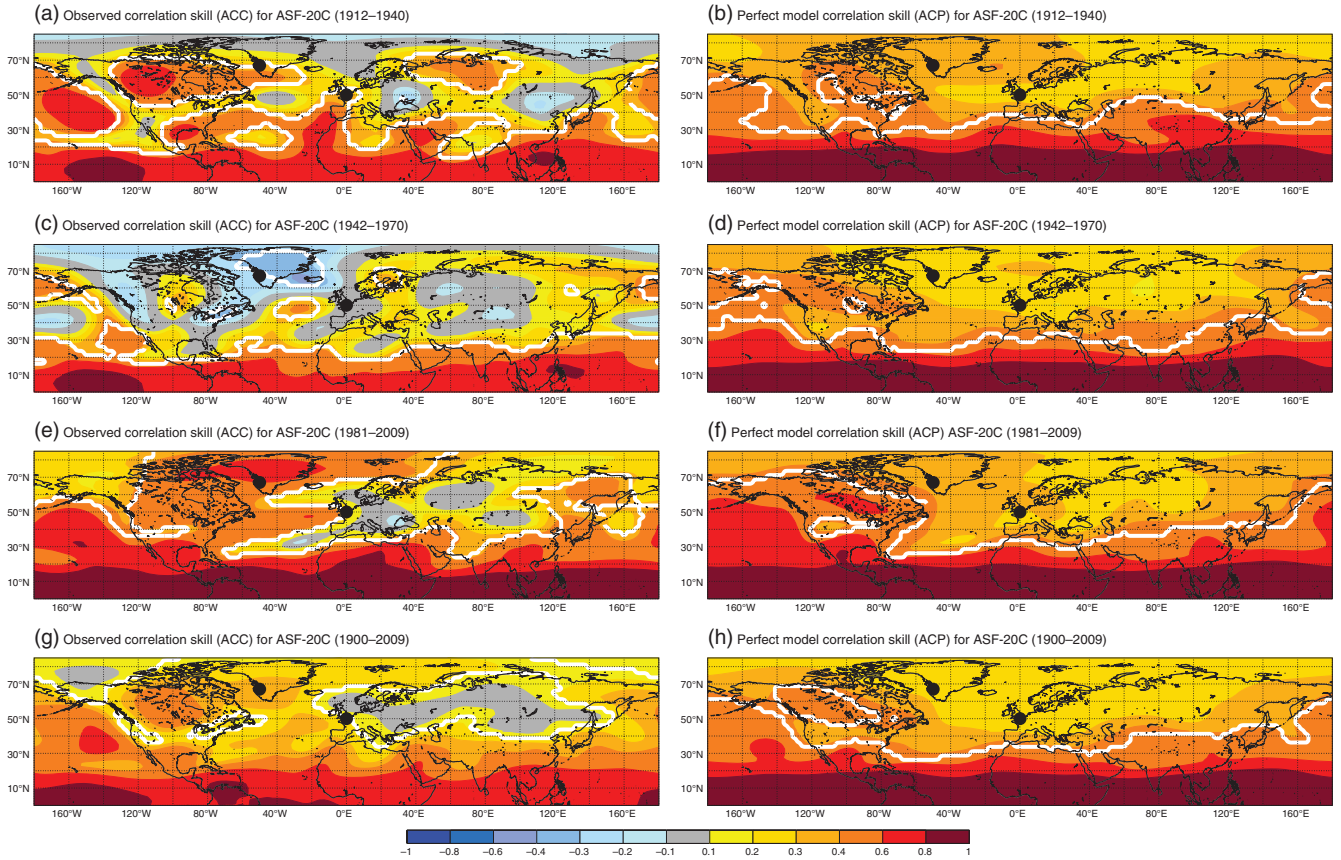


**FIGURE 6** Anomaly correlation between the ensemble mean and observations (a,c,e,g), and perfect model anomaly correlation between the ensemble mean and each ensemble member (b,d,f,h), in ASF-20C for different time periods. The white lines indicate where the correlation skill becomes significant at the 95% confidence level. Valid for Z500 forecasts of DJF initialised on 1 November over the periods 1912–1940 (a,b), 1942–1970 (c,d), 1981–2009 (e,f) and 1900–2009 (g,h), using ERA-20C as verification

skill, we analyse how individual winters contribute differently to the overall correlation. The Pearson correlation coefficient $r$ used here is defined as the covariance between the ensemble mean and the verification divided by the product of the standard deviations of the ensemble mean and the verification for a time period $t = 1 \ \ldots \ N$:

$$r = \frac{cov(ensmean, obs)}{\sigma_{ensmean}\sigma_{obs}} = \frac{\sum_{t=1}^{N} ensmean_t' \cdot \Delta obs_t'}{\sigma_{ensmean}\sigma_{obs}}, \quad (2)$$

where $ensmean_t'$ and $obs_t'$ are the ensemble mean and observed anomalies for the individual winter of time step (year) $t$. That is, the sample covariance is estimated by aggregating the product of the ensemble mean anomaly and the observed anomaly for each winter over all forecast years. Figure 8b shows time series of the yearly contributions to the normalised cov(*ensmean,obs*) for the observed correlations (black) and the perfect model correlations (purple line and coloured shadings).

Several things can be noted from Figure 8. Firstly, there is substantial variability in the yearly contributions to skill for both the observed correlations and the perfect model. It is also noticeable that the period in the middle of the twentieth century has a reduced activity with smaller overall contributions leading to overall reduced skill levels (cf. Figure 5a and Weisheimer *et al.,* 2017). The contribution of the perfect model mean is positive throughout. The five record winters with the largest contributions to observed skill are the years 1939, 1940, 1976, 1988 and 2009, as indicated by the vertical lines and red dots. There is a tendency for more frequent extreme winters towards the latest decades of the time series. The winter with the strongest negative NAO index (Figure 8a) is 2009 which is also one of the five record winters in terms of correlation contribution. The years 1939 and 1940 are both also characterised by strongly negative NAO states (perhaps as a consequence of a series of El Niño years), as is 1976. It is interesting to note, however, that the winter of 1988, which gives the strongest overall individual (positive) contribution, is the most positive NAO winter in the 110-year period.

In general, there is a good correspondence between positive contributions to the observed skill and positive contributions to the perfect model skill. The relationship is not perfect and the perfect model contribution for the five extreme winters is smaller than the contribution to the observed skill. While most of the time the distribution of perfect model contributions (orange shades) includes the observed skill contribution, it is noticeable that for the record winters the observed contribution falls nearer the upper tails of this distribution or even slightly outside (e.g. 1988).

The observed and ensemble-mean model anomalies of Z500 for the five record winters are displayed in Figure 9.

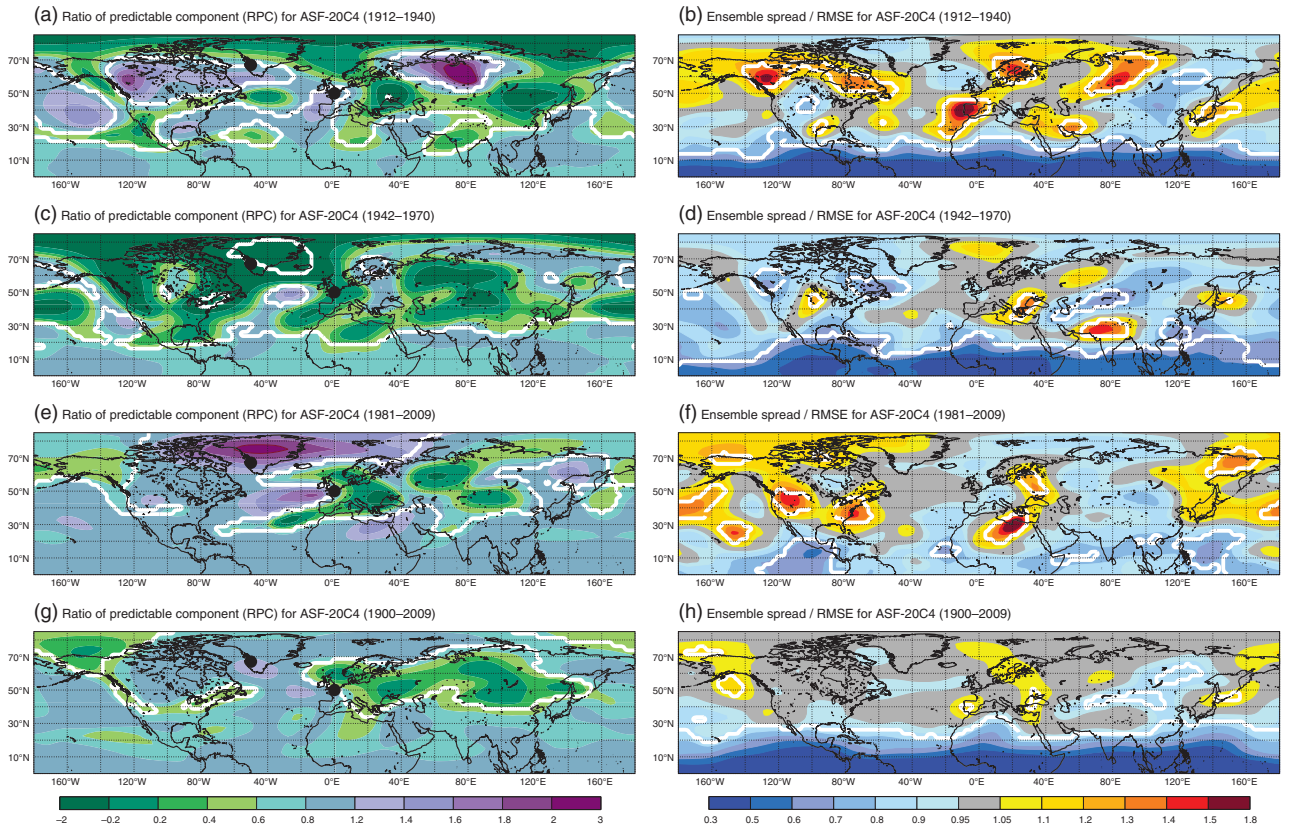


**FIGURE 7** Ratio of predictable components (*RPC*) (a,c,e,g), and ratio of ensemble spread to *RMSE* (b,d,f,h), in ASF-20C for different time periods. The white lines in the left column indicate where the correlation skill becomes significant at the 95% confidence level (see Figure 6). In the right column the white lines indicate where the ratio is significantly different from 1 at the 95% confidence level. Valid for Z500 forecasts of DJF initialised on 1 November over the periods 1912–1940 (a,b), 1942–1970 (c,d), 1981–2009 (e,f) and 1900–2009 (g,h), using ERA-20C as verification

They show the negative NAO flow pattern over the North Atlantic for the four winters 1939, 1940, 1976 and 2009 with strong positive Z500 anomalies centred over Greenland. The model is largely able to reproduce these large-scale patterns. The finding that it is mostly the negative NAO winters that contribute to the overall skill in predicting the NAO is in agreement with Weisheimer *et al.* (2017) who stratified the winters according to NAO phase and also looked at probabilistic forecast skill as a function of the strength of the NAO. In contrast to the negative NAO pattern for those four winters, the extreme NAO positive winter 1988 with the strongest individual contribution to the correlation skill was dominated by a pronounced zonal flow across the North Atlantic that led to an unusually stormy season over Scotland and northwest of the British Isles (Murray, 1991). The winter followed the warmest year on record at the time (Ratcliffe, 1989).

## 5 | UNCERTAINTIES OF SKILL ESTIMATES

Estimations of skill in seasonal hindcasts are typically exposed to sampling uncertainties due to finite ensemble size and hindcast length, the degree of which varies with regions, seasons and variables. The extratropical atmospheric flow in winter with its substantial interannual variability (predictable and unpredictable) suffers from rather large sampling uncertainties compared to quantities with lower variability as, for example, tropical SSTs. Here we explore the impact of both ensemble size and hindcast size on estimates of *RPC*, *RMSE* and ensemble spread.

### 5.1 | Ensemble size

One source of sampling uncertainty is related to the finite and often small size of the ensembles used to construct the forecasts. The ensemble size not only impacts the estimation of forecast probabilities but also has an effect on the estimation of the ensemble mean and related skill measures (Déqué, 1997; Kumar *et al.,* 2001; Richardson, 2001; Müller and Appenzeller, 2005; Scaife *et al.,* 2014).

The ASF-20C hindcast ensemble, with its large ensemble size of 51 ensemble members and 110 hindcast years, provides an ideal dataset to estimate the effect of ensemble size on the predictability of the NAO. Figure 10a illustrates the dependence of the ratio between *RMSE* and ensemble spread as a function of the size of the ensemble, while Figure 10b shows a similar dependence for the *RPC*. The grey solid lines denote the behaviour of the seasonal forecasts when verified against observations. The ratio of *RMSE* and ensemble spread shows a strong overestimation for small ensemble sizes compared to its asymptotic value of 1.05, which is mostly due to uncertainties in the estimation of

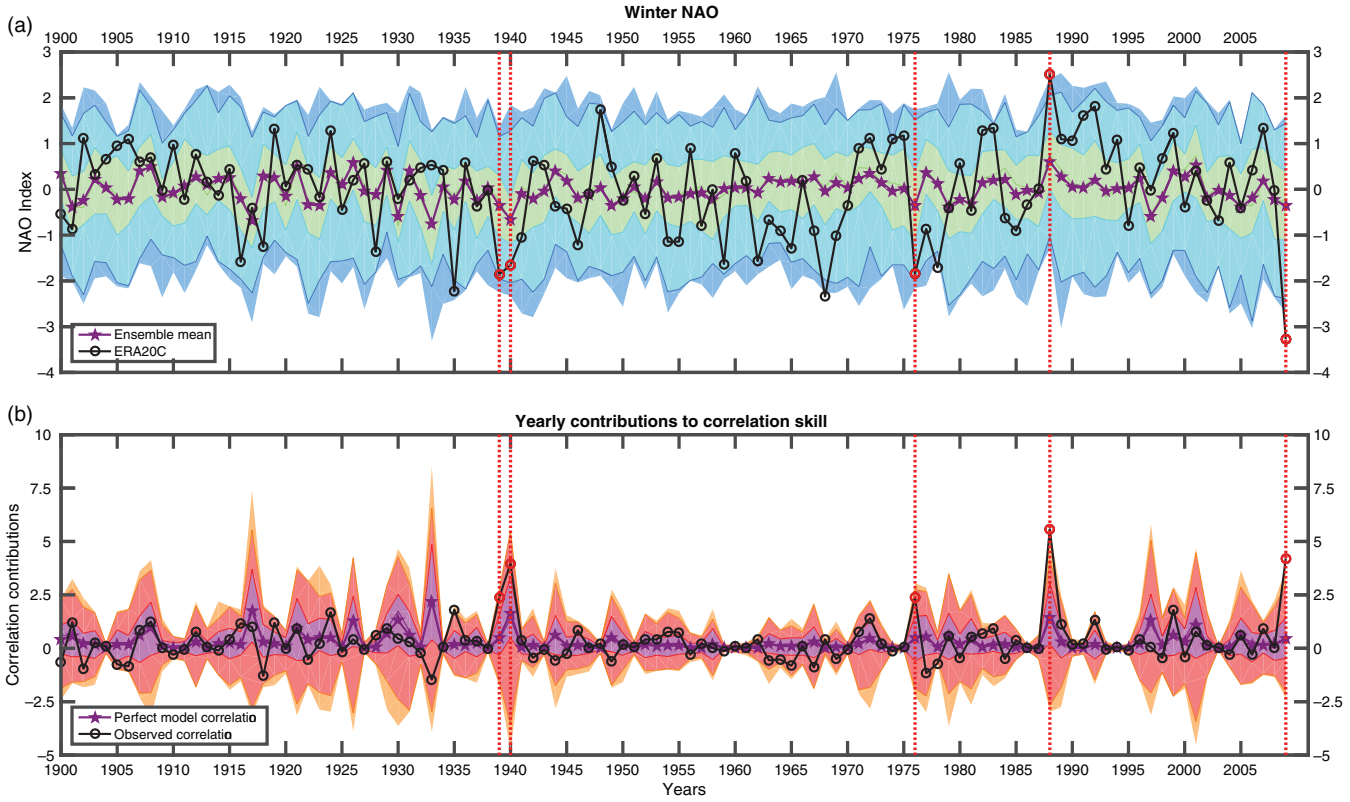


**FIGURE 8** (a) Comparison of NAO DJF index from 1900 to 2009 calculated for ERA-20C (black) and for the model ensemble mean, ASF-20C (purple). (b) Yearly contributions to correlation skill for observed skill (black) and perfect model skill (purple). The 5 years with the highest correlation skill contributions to the observed skill (1939, 1940, 1976, 1988 and 2009) are highlighted with vertical lines and red dots in both panels. Coloured shadings show the minimum, 2.5%, 25%, median, 75%, 97.5% and maximum of model ensemble probability density functions for the NAO index in (a) and the contributions to the perfect model skill in (b)

the *RMSE* (not shown). An ensemble of *circa* 30 members would be needed to approximate the asymptotic value reasonably well.

The black dashed lines in Figure 10 represent the perfect model behaviour. Here, the perfect model is constructed in such a way that the respective verifying ensemble member has been excluded from the computation of the ensemble mean. This approach mimics real-life forecast situations where the verification cannot be part of the forecast. As can be seen, the perfect model has a ratio *RMSE/spread* closer to one for almost all ensemble sizes than in the real-world forecasts, yet its value of 1.02 for the largest ensemble size of 51 still differs somewhat noticeably from the theoretical value of one. A simple adjustment for the finite size of the ensemble spread has been introduced (see also Rodwell *et al.*, 2018) in the form of a multiplicative factor of $(m+1)/(m-1)$ for the ensemble variance, with $m$ being the ensemble size. The resulting behaviour of the ratio *RMSE/spread* is shown by the grey dashed and black dotted lines. It can be seen that the adjustment is very effective in reducing the overestimation of the ratio for small ensemble sizes. It reduces the asymptotic value for the real-world estimate from 1.05 to approx. 1.03. The adjusted estimate for the perfect model is much closer to its theoretical value of one for all ensembles sizes and is not distinguishable in the plot from one for ensembles with more than 40 members.

It is worth mentioning that the theoretical value of the ratio between *RMSE* and ensemble spread of one for a perfect model can be achieved if the perfect model ensemble mean is constructed from all ensemble members, including the verifying member. The black solid lines in Figure 10 show the empirical value of the perfect model if the perfect model ensemble mean is computed in this way. Defining the perfect model by including all ensemble members yields for the full range of ensemble sizes, also for very small ones, the correct ratio of one, within very small uncertainties. As discussed, even with 51 ensemble members and a hindcast length of 110 years (which is currently the most we can realistically expect from seasonal forecasts) the perfect model estimate based on the ensemble mean that *excludes* the verifying member has still not reached one, within small uncertainties, see black dashed line in Figure 10. Although the approach to include the verifying ensemble member is counter-intuitive to any forecaster, it does give the expected theoretical results for a perfect model. Similarly to the observed curves (grey lines), a partial compensation of the perfect model deficiency when the verifying member is excluded in the ensemble mean can be obtained by the simple adjustment of the ensemble spread computation for finite ensembles (dotted black line).

The anomaly correlation coefficient of the NAO index computed over the entire 110 hindcast years is $r = 0.31$ ($RPC = 1.02$) and the dependence of the correlation on the

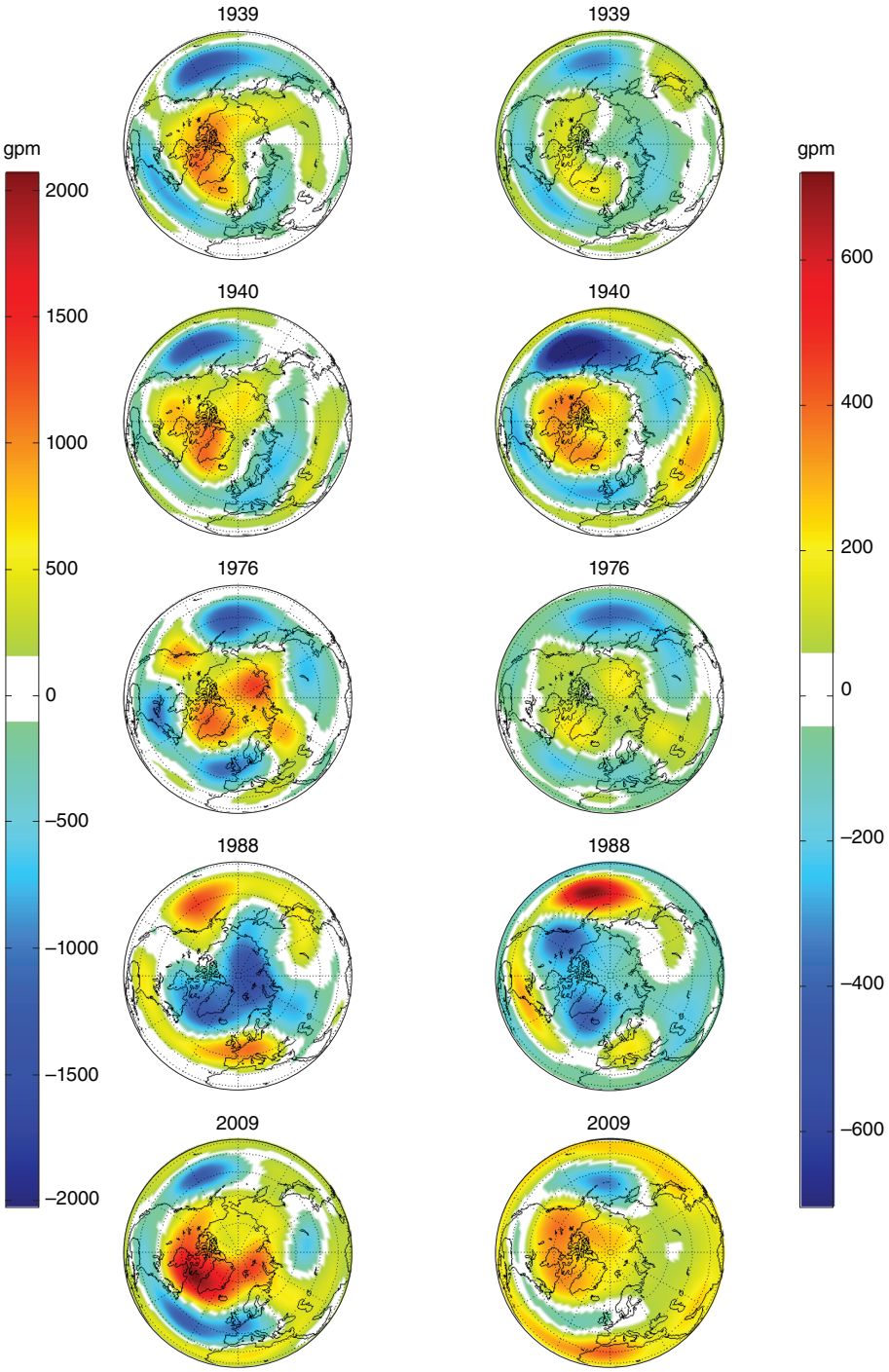


**FIGURE 9** Anomalies corresponding to the five winters with the highest contributions to observed correlation skill as highlighted in Figure 8. ERA-20C (left) and ASF-20C ensemble-mean anomalies (right). Note the difference in colour scale

size of the ensemble asymptotes this value rather quickly in our model (not shown) compared to the UK Met Office model which has a larger sensitivity (Scaife *et al.,* 2014*)*. For small ensemble sizes the correlation is underestimated. From a size of the ensemble of approx. 15 members and larger, the correlation stays constant. The perfect model correlation (excluding the verifying member) has an overall correlation skill of $r_{\mathrm{pm}} = 0.25$.

While the *RPC* and the ratio ensemble *spread/RMSE* for the perfect model have to be exactly 1, no such theoretical relationship exists for the correlation coefficient of the perfect model. Thus the effect of how to define the perfect model cannot easily be studied for correlation skill. However, for the *RPC* which is closely related to the correlation (see Equation (1), the value for a perfect model has to be one. In Figure 10b we test how well this condition is fulfilled for the above-discussed two versions of defining the perfect model by either excluding (dashed black) or including (solid black) the verifying member in the computation of the ensemble mean. It becomes clear that the perfect model which excludes the verifying member strongly underestimates the theoretical value of one for all sizes of the ensemble and especially

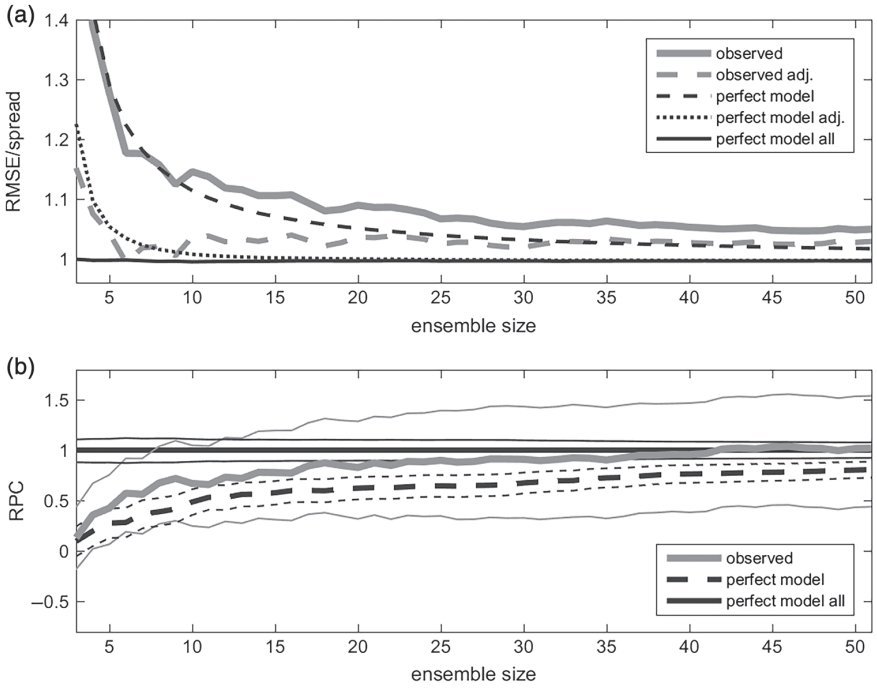


**FIGURE 10** Dependence on ensemble size of the ratio *RMSE/spread* (a) and the *RPC* (b) of the NAO index, based on the 110-year hindcasts of ASF-20C. The grey solid lines show the diagnostics of the real world (i.e. using observations as verification). The black lines show versions of the perfect model: The verifying ensemble member is excluded (dashed) or included (solid) in the estimation of the corresponding ensemble mean. The grey dashed and the black dotted lines indicate that an adjustment for the finite ensemble size in the estimation of the ensemble spread is used. Thin lines around the *RPC* curves indicate 90% confidence levels

for smaller ensemble sizes. For the largest available ensemble size it still is only $RPC = 0.81$. Note that the uncertainties around the perfect model *RPC* are much reduced due to the larger sample sizes when compared to the real-world *RPC*.

The empirically derived *RPC* for the perfect model where the verifying ensemble member is retained in the computation of the ensemble mean shows an excellent agreement with the theoretical value of 1 throughout the range of ensemble sizes (black solid line in Figure 10b). The corresponding perfect model correlation coefficient over the entire period and based on 51 ensemble members is increased to $r_{\mathrm{pm}} = 0.31$ and is thus at a very similar level to the observed correlation. This analysis demonstrates that if a definition of the perfect model is used which fulfils the theoretical conditions, there is no discrepancy between the skill of the real world and the skill estimate from the model. Note that no simple adjustment to the finite ensemble size, as for the ensemble spread in Figure 10a, can be applied to the correlation and *RPC* estimates.

## 5.2 | Length of hindcast period

The largest sampling uncertainty for seasonal hindcasts is related to the typically very short length of the hindcast period, which often consists of just 20–30 data points. The impact of additional hindcast years does not, in general, lead to an asymptotic increase in skill and reduction of uncertainty because of the flow-dependent aspect of skill which is closely linked to interannual and low-frequency variability (see e.g. Weisheimer *et al.,* 2017). The impact of a bigger size of the ensemble is, however, rather monotonic in the sense that it will, on average, always increase the skill and reduce the uncertainty. A quantitative example of the NAO uncertainties due to ensemble size and hindcast length can be found in the study by Siegert *et al.* (2016) using a Bayesian framework.

We will now analyse the robustness of the *RMSE* and correlation skill estimates due to finite hindcast periods. As discussed in section 2, our hypothesis is that these skill estimates, and in particular the correlation measures, lack robustness when only a small number of data is considered. Here, a synthetic long dataset of the "truth" (i.e. observations) and ensemble mean (i.e. model) was generated from random draws of standardised correlated Gaussian data. These data with their "true" correlations were used to sub-sample shorter periods and then study the effect of sampling on uncertainties of the RMSE and correlation measures.

"True" correlations between the observations and the model in the range from −1 to +1 were prescribed. We use a sample size for the long "truth" data of 30,000. The non-linear relationship between the *RMSE* and the correlation is displayed in Figure 11a and follows the analytical expression of Barnston (1992). In order to analyse the uncertainties related to the sample size, we have randomly sub-sampled the 30,000 data points with lengths between 30 and 300 data points. As a result of the Monte-Carlo sampling we obtain a distribution of skill measures (*RMSE* and correlation) for each hindcast length. The robustness of the skill measure is related to the width of this distribution with sharper distributions

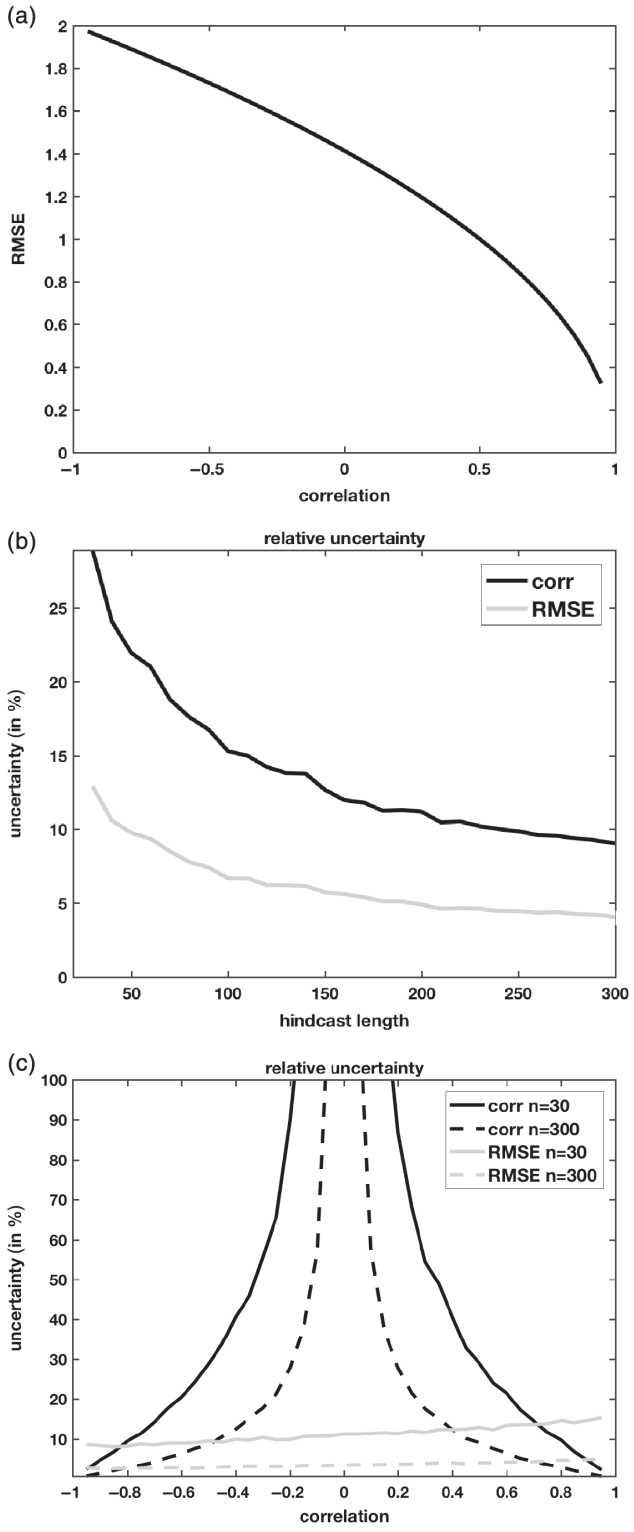


**FIGURE 11** (a) Relationship between *RMSE* and correlation for standardised data. (b) Relative uncertainty (%) of the correlation and *RMSE* as a function of hindcast length for a "true" correlation of 0.5. (c) Relative uncertainty (%) of the correlation and *RMSE* as a function of true correlation for sub-sample sizes of 30 (solid lines) and 300 (dashed lines)

(smaller variance) having more robust estimates. We define the relative uncertainty measure of skill as the standard deviation of the skill distribution normalised with its "true" skill obtained from the full-length data. Figure 11b illustrates as an example (for an underlying "true" correlation of 0.5) how the relative uncertainty decreases when the hindcast sample size increases. The uncertainty decline follows a nearly exponential decay for both measures and can be reduced by a factor of 3 when the sample size is increased from 30 to 300. It is very clear from Figure 11b that the relative uncertainty of the correlation measure is larger than for the *RMSE* measure, in support of our working hypothesis (b) discussed in the last paragraph of section 2.

In order to illustrate the robustness of the skill estimates for a wider range of underlying "true" correlations, Figure 11c shows for two example hindcast sample sizes (solid lines for $N = 30$ and dashed lines for $N = 300$) how the uncertainty in estimating the true skill varies as a function of the "true" correlation. For small sample sizes the uncertainty estimates for both skill measures are larger than for larger sample sizes across all underlying correlations (comparing solid and dashed curves). The uncertainty of the correlation measure shows a much stronger dependence on the "true" correlation than the *RMSE* (comparing black and grey lines). Correlation uncertainty increases exponentially for smaller correlations with a singularity at zero and vanishing uncertainty for perfect correlations, whereas the *RMSE* uncertainty is nearly constant across the correlations. As a result, the *RMSE* estimates are more robust than the estimates of correlation for a very wide range of "true" correlations between −0.8 and +0.7. This result is independent of the hindcast sample size.

## 6 | SUMMARY AND CONCLUSIONS

In this article, we have studied the atmospheric predictability in the Euro-Atlantic region in different versions of the ECMWF seasonal forecast model with a specific view on the so-called signal-to-noise paradox of the winter NAO. One of the main factors that constrains tropospheric seasonal forecast skill in the extratropics compared to the Tropics is related to the forecast signals being small while the noise levels of interannual variability are rather high, resulting in low signal-to-noise ratios. In general, low signal-to-noise ratios are expected to also show low correlation skill. The implication of a situation with higher-than-expected skill is that the real world appears more predictable than the forecast model seems to suggest. High predictive skill as reported, for example, in Eade *et al.* (2014), Scaife *et al.* (2014), Stockdale *et al.* (2015) and Baker *et al.* (2018) comes with a paradox, or conundrum (Scaife and Smith, 2018): the level of actual forecast skill appears to be too high compared to the intrinsic predictability one would expect for such forecasting systems, given their low signal-to-noise ratios.

In order to examine what the evidence for the signal-to-noise NAO paradox is in the ECMWF seasonal forecast model, we have analysed seasonal hindcasts over the period 1981–2009 with different configurations of the model (varying horizontal resolutions, atmosphere only versus coupled ocean/sea ice). In all simulations most of the Northern

Hemisphere extratropics are under-dispersive (overconfident) in the ensemble-mean correlation-based *RPC* skill measures. Model estimates of predictive skill across the eastern North Atlantic–European region are positive and higher than the real-world predictability in that region which is largely characterized by a lack of skill. However, Greenland and the Arctic further to the east are areas of high and significant observed skill, with the perfect model skill being lower than the observed skill, leading to the paradoxical situation where the real world appears more predictable than the models suggest (over-dispersion). Here, *RPC* values larger than 1 and up to 2 reflect the high levels of observed skill in this region. These results resemble in its general structure over the North Atlantic the *RPC* for similar seasonal forecasts with the UK Met Office model, even though the forecasting systems described here tend to be a little less skilful in predicting the NAO index over the common hindcast period.

On the contrary, the diagnosed relationships between the ensemble mean *RMSE* and ensemble spread clearly do not show an over-dispersion problem over the North Atlantic (Figure 7). Rather, the forecasts reflect well-calibrated ensemble systems where the *RMSE* is slightly larger than the ensemble spread (minor under-dispersion). Such a behaviour is in very good agreement with the typical performance of the ECMWF ensemble for short and medium-range forecasts (Rodwell *et al.,* 2018) and long-range sub-seasonal ensemble forecasts over the North Atlantic (L. Ferranti, personal communication).

One of our proposed hypotheses to explain the predictability is that the recent short hindcast period from after 1980 onwards is not sufficiently representative for the longer-term behaviour of the climate system with its non-stationary character, especially over the North Atlantic. For example, the mid-twentieth century was characterised predominantly by the negative phase of the NAO, while a strong positive trend starting in the 1960s and lasting for approx. three decades led to a shift towards positive NAO for the majority of winters during the more recent decades. What are the implications of such low-frequency climatic variations for seasonal NAO predictability estimates?

We have used the Atmospheric Seasonal Forecasts of the 20th Century (ASF-20C: Weisheimer *et al.,* 2017) that cover the period 1900 to 2010 to analyse multi-decadal variability in forecast skill and predictability estimates. In the middle of the twentieth century the picture of the observed skill drastically changed compared to more recent decades. Greenland and parts of the North Atlantic show negative observed correlation skill while the perfect model indicates a much smaller reduction of skill over these areas. That means that regions which appear more predictable in the real world than in the model (i.e. paradoxical or underconfident) during recent decades, were strongly overconfident during the mid-century decades (see also O'Reilly, 2018). In the earlier parts of the twentieth century the situation appears less extreme in either direction with slightly higher values of perfect model skill over Greenland and the northern North Atlantic and moderate observed skill levels over most of the North Atlantic except for the northeastern parts. Similar to the correlation skill, the *RPC* undergoes multi-decadal variations with periods indicative of underconfidence during the more recent decades and periods with evidence for overconfidence in the middle of the century. However, the *RPC* of 1.02 computed over the entire 110-year hindcast period is indicative of a near-perfect system.

We would like to emphasise that whilst the inter-decadal differences in NAO skill are, by themselves, only marginally statistically significant (Weisheimer *et al.,* 2017), the variations in skill strongly co-vary with statistics of the general circulation itself (examples of which can be found in e.g. Minobe, 1997; Hoerling *et al.,* 2001; Derome *et al.,* 2005; Fletcher and Saunders, 2006; Greatbatch and Jung, 2007; Douville *et al.,* 2017; Hegerl *et al.,* 2018; Huang *et al.,* 2018) suggesting that such differences are indeed physically based. In particular, the temporal skill evolution of the NAO co-varies with changes in the skill of the Pacific–North America (PNA) pattern that are highly statistically significant and relate to changes in the ENSO–North Pacific teleconnections (O'Reilly *et al.,* 2017; O'Reilly, 2018). Further studies to understand the physical mechanisms of these inter-decadal general circulation changes in the atmosphere and ocean both in observed data as well as in models (initialised and non-initialised: Kumar and Chen, 2018) are clearly needed.

In an attempt to better understand the atmospheric flow that leads to the unexpected behaviour of underconfidence when diagnosed using correlation-based measures, we have analysed the contributions to skill from individual years. There is substantial variability in these yearly contributions to skill for both the observed correlations and the perfect model. The period in the middle of the twentieth century has a reduced activity with smaller overall contributions leading to overall reduced skill levels. It is perhaps not too surprising that years with largest observed NAO anomalies (both signs) also have the largest skill contributions but the covariance contribution in Equation (2) consists of the product of both the observed and modelled anomalies and thus, in principle, the covariance contribution can be small for observed extreme anomalies if the corresponding model anomalies are for example close to zero. The observed and ensemble-mean model anomalies of Z500 for the five record winters show that strong negative NAO flow patterns over the North Atlantic with large positive Z500 anomalies centred over Greenland dominate during 4 of the 5 winters. However, it is the winter 1988 with the absolute largest positive NAO index that produced the strongest individual positive contribution to correlation skill.

Skill estimates based on seasonal hindcasts are subject to various sampling uncertainties, for example, due to the finite size of the ensemble or the hindcast length. The ECMWF seasonal forecasts are somewhat less sensitive to ensemble size than other reported systems (Scaife *et al.,* 2014; Baker *et al.,* 2018). It was shown that the largest sampling uncertainty

for seasonal hindcasts is related to the typically rather short hindcast period (see also Shi *et al.,* 2015). The relationship between expected correlation skill and signal-to-noise will only be realised in the limit of very large sample sizes, or long hindcast periods (Kumar, 2009). Using synthetic data, we have demonstrated how the relative uncertainties in the estimation of skill decrease with longer hindcast length and how the uncertainties differ systematically between correlation and *RMSE* measures. It is argued that the relative uncertainty due to small hindcast sample sizes is, under most circumstances, larger for the correlation measures than it is for the RMSE (implying more robust estimates for the *RMSE*).

Furthermore, it has been demonstrated that the so-called perfect model approach needs to be carefully defined. While it is commonly agreed that in a perfect model approach observations and ensemble members should be interchangeable, there is less agreement about how the ensemble mean of a perfect ensemble should be constructed. From a forecasting perspective one would want to exclude the verifying ensemble member from the ensemble mean. However, as has been shown, the theoretical and analytical perfect model properties of matching *RMSE* and ensemble spread and an *RPC* of exactly one can only be achieved when the verifying member is included in the computation of the ensemble mean. The effect of omitting the verifying member is still noticeable even with an ensemble size of 51 members and 110 hindcast years. It was demonstrated that if the inclusive definition of the perfect model is used (which fulfils the theoretical conditions), there is no discrepancy between the skill of the real world and the skill estimate from the model. Adjusting the estimation of ensemble spread for small ensembles helps reduce the dependence on ensemble size.

Conjectures about the possible reasons for the signal-to-noise paradox form an active on-going debate within the scientific community. For example, Strommen and Palmer (2018) have recently proposed an alternative hypothesis for the NAO paradox: model deficiencies in atmospheric regime behaviour may manifest themselves as increased spread in the forecast distributions. They have used a bimodal toy model to demonstrate that underestimating the regime persistence can lead to both high levels of skill and low signal-to-noise ratios (*RPC*). Some evidence for underestimating of regime persistence in medium-range weather forecasts were provided in Matsueda and Palmer (2018).

The fact that most of the signal-to-noise paradox in the ECMWF model is confined to Greenland and adjacent areas to the northeast points at the possibility that the representation of orography in the model and related problems (see e.g. Pithan *et al.,* 2016) might contribute to the paradox. While this is a question that will potentially receive more attention within the community soon, any explanation would need to be able to account for the multi-decadal variations in the model behaviour.

Siegert *et al.* (2016) use a Bayesian inferential framework for a statistical signal–plus–noise model to analyse the NAO seasonal hindcast data of Scaife *et al.* (2014) generated by the Met Office model GloSea5. They conclude that the observed skill of $r = 0.62$ in the UK Met Office model has large sampling uncertainty and falls into the upper tail of the posterior distribution, suggesting a high chance of a decrease in correlation skill if the model were evaluated over different periods. These results suggest that the particular 20-year hindcast period of Scaife *et al.* (2014) is unusual and produces higher-than-normal correlation skill. Due to the nature of their statistical model and analysis, any multi-decadal fluctuations within the atmospheric dynamics cannot be considered.

A further conclusion from the Siegert *et al.* (2016) study is that the predictable signal in the model is too weak. We have thus far emphasised that decadal variability and insufficient sample size can affect the robustness of skill and confidence estimates of hindcasts that are only a few decades long. However, it is also possible that hindcasts can display under-confidence in the *RPC* measure whilst seemingly having the correct *RMSE/spread* relationship; such a situation can occur through the model having a weak predictable signal (see also O'Reilly *et al.,* 2018b). This can be demonstrated using the ASF-20C NAO hindcast, which actually has *RPC* $\sim 1$. We first split the NAO hindcast of each ensemble member into $NAO_{\mathrm{ens}} = SIGNAL + NOISE$, where the *SIGNAL* is the ensemble mean NAO and the *NOISE* is the anomaly from the ensemble mean in each ensemble member. If we then suppose that the *SIGNAL* could be weaker or stronger in the model for some reason, we can artificially scale the *SIGNAL* in the hindcast to reflect this possibility. The NAO for each ensemble members is then $NAO_{\mathrm{ens}} = f{*}SIGNAL + NOISE$, where $f$ is signal scaling factor. We performed this scaling on the hindcast NAO from the ASF-20C ensemble using signal scaling factors between 0.01 and 2. The recalculated *RPC* and *RMSE/spread* values for these various scaling factors are shown in Figure 12. The first thing to note is that because the ensemble mean *SIGNAL* is linearly scaled, the ensemble mean correlation skill is the same for all scaling factors. The scaling factor, however, does have a large influence on the *RPC*. As the scaling factor is reduced – and the *SIGNAL* is weakened – the *RPC* increases dramatically. For example, the *RPC* approximately doubles when the signal is scaled by 0.25 (Figure 12). In contrast, changes in the *RMSE/spread* values are very small when the signal is reduced. The relatively small changes in the *RMSE/spread* reflects the fact that the *SIGNAL* makes up relatively little of the variance in these NAO hindcasts, such that $RMSE \sim \mathrm{std}(NOISE)$ regardless of whether the scaling factor is 0 or 2. For more skilful forecasts, such as on the medium-range or in the Tropics, where the *SIGNAL* accounts for a greater fraction of the total variance, the *RMSE/spread* will be more sensitive to changes in the size of the *SIGNAL*.

It is interesting to note that all tested configurations of the ECMWF model agree in their general patterns of observed and perfect model skill for the common recent hindcast period, indicating little sensitivity to atmospheric and oceanic

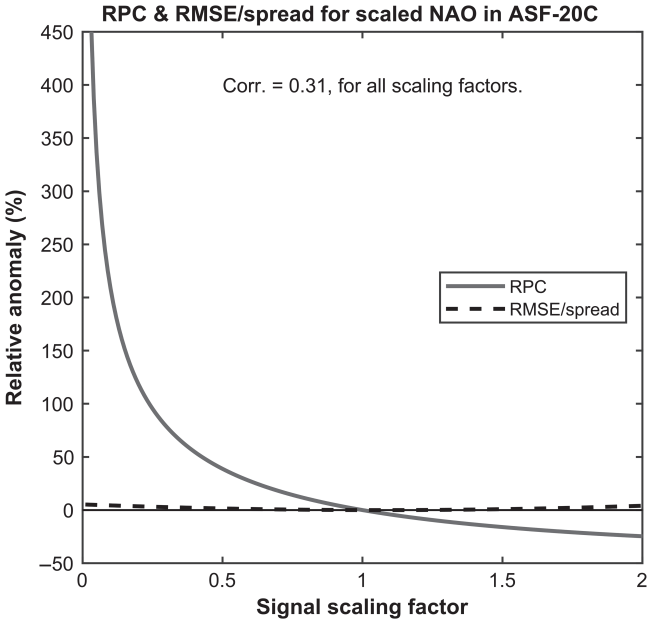


**FIGURE 12** Relative anomalies (%) of *RPC* and the ratio *RMSE/spread* of the NAO hindcasts in the ASF-20C ensemble. The NAO index in each ensemble member has been scaled as $NAO_{ens} = f * SIGNAL + NOISE$, where *SIGNAL* is the ensemble mean NAO, *NOISE* is the anomaly from the ensemble mean in each ensemble member and $f$ is the signal scaling factor. Anomalies relative to $f = 1$ (i.e. the raw hindcast data) are plotted. The ensemble mean NAO skill is equal to 0.31 for all scaling factors

model resolution and whether or not the system uses an interactive ocean and sea-ice model. The high sensitivity of correlation-based skill measures to weak predictable signals (Figure 12) makes disentangling of physical deficiencies and sampling uncertainty challenging. In order to separate the effect on the NAO predictability due to using prescribed SSTs in the presented long seasonal hindcasts, it is planned for the near future to study the behaviour of a fully coupled atmosphere–ocean–sea-ice system from the beginning of the twentieth century.

## ACKNOWLEDGEMENTS

The authors thank James Heatley, Mio Matsueda, Kristian Strommen, Laura Ferranti, Magdalena Balmaseda, Dan Rowlands, Stefan Siegert, David Stephenson, Bart van den Hurk and Jochen Bröcker for fruitful discussions. The careful comments by two reviewers have helped improve the manuscript. We acknowledge the provision of seasonal hindcast data from System 4 and SEAS5 and the reanalysis data ERA-Interim and ERA-20C by ECMWF. The computing resources for ASF-20C were kindly provided through the ECMWF special project "Seasonal forecasts of the 20th Century: reliability, attribution and the impact of stochastic perturbations." AW, DD and DM acknowledge supported from the EU FP7 project SPECS (grant agreement number 308378), AW had support through NERC NCAS, from the EU FP7 project EUCLEIA (grant agreement number 607085) and the EU H2020 project EUCP (grant agreement number 776613), and AW and COR acknowledge funding from the Natural Environmental Research Council (project SummerTIME, NE/M005887/1).

## ORCID

*Antje Weisheimer* https://orcid.org/0000-0002-7231-6974
*David MacLeod* https://orcid.org/0000-0001-5504-6450
*Christopher O'Reilly* https://orcid.org/0000-0002-8630-1650
*Tim N. Palmer* https://orcid.org/0000-0002-7121-2196

## REFERENCES

Anscombe, F.J. (1973) Graphs in statistical analysis. *The American Statistician*, 27(1), 17–21. https://en.wikipedia.org/wiki/Anscombe%27s_quartet.

Baker, L.H., Shaffrey, L.C., Sutton, R.T., Weisheimer, A. and Scaife, A.A. (2018) An intercomparison of skill and over/underconfidence of the wintertime North Atlantic Oscillation in multi-model seasonal forecasts. *Geophysical Research Letters*, 45, 7808–7817. https://doi.org/10.1029/2018GL078838.

Barnston, A.G. (1992) Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Weather and Forecasting*, 7, 699–709.

Barnston, A.G., M. Tippett, M. L'Heureux, S. Li and D. DeWitt. (2012) Skill of real-time seasonal ENSO model predictions during 2002–11: is our capability increasing? *Bulletin of the American Meteorological Society*, 93, 631–651. doi:https://doi.org/10.1175/BAMS-D_11-00111.1.

Befort, D.J., Wild, S., Knight, J.R., Lockwood, J.F., Thornton, H.E., Hermanson, L., Bett, P.E., Weisheimer, A. and Leckebusch, G.C. (2018) Seasonal forecast skill for extratropical cyclones and windstorms. *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.3406.

Bell, C.J., Gray, L.J., Charlton-Perez, A.J., Joshi, M. and Scaife, A.A. (2009) Stratospheric communication of El Niño teleconnections to European winter. *Journal of Climate*, 22, 4083–4096.

Beverley, J.D., Woolnough, S.J., Baker, L.H., Johnson, S.J. and Weisheimer, A. (2018) The Northern Hemisphere circumglobal teleconnection in a seasonal forecast model and its relationship to European summer forecast skill. *Climate Dynamics*. https://doi.org/10.1007/s00382-018-4371-4.

Butler, A.H., Polvani, L.M. and Deser, C. (2014) Separating the stratospheric and tropospheric pathways of El Niño-Southern Oscillation teleconnections. *Environmental Research Letters*, 9, 024014. https://doi.org/10.1088/1748-9326/9/2/024014.

Dawson, A., Palmer, T.N. and Corti, S. (2012) Simulating regime structures in weather and climate prediction models. *Geophysical Research Letters*, 39, L21805. https://doi.org/10.1029/2012GL053284.

Déqué, M. (1997) Ensemble size for numerical seasonal forecasts. *Tellus*, 49A, 74–86.

Derome, J., Lin, H. and Brunet, G. (2005) Seasonal forecasting with a simple general circulation model: predictive skill in the AO and PNA. *Journal of Climate*, 18, 597–609.

Douville, H., Peings, Y. and Saint Martin, D. (2017) Snow-(N)AO relationship revisited over the whole twentieth century. *Geophysical Research Letters*, 44, 569–577. https://doi.org/10.1002/2016GL071584.

Dunstone, N., Smith, D., Scaife, A.A., Hermanson, L., Eade, R., Robinson, N., Andrews, M. and Knight, J. (2016) Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nature Geoscience*, 9, 809–814. https://doi.org/10.1038/ngeo2824.

Eade, R., Smith, D., Scaife, A.A., Wallace, E., Dunstone, N., Hermanson, L. and Robinson, N. (2014) Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, 41, 5620–5628. https://doi.org/10.1002/2014GL061146.

Fletcher, C. and Saunders, M.A. (2006) Winter North Atlantic hindcast skill: 1900–2001. *Journal of Climate*, 19, 5762–5776. https://doi.org/10.1175/JCLI3949.1.

Greatbatch, R.T. and Jung, T. (2007) Local versus tropical diabatic heating and the winter North Atlantic Oscillation. *Journal of Climate*, 20, 2058–2075.

Hansen, F., Greatbatch, R.J., Gollan, G., Jung, T. and Weisheimer, A. (2017) Remote control of NAO predictability via the stratosphere. *Quarterly Journal*

*of the Royal Meteorological Society*, 143, 706–719. https://doi.org/10.1002/qj.2958.

Hao, Z., Singh, V.P. and Xia, Y. (2018) Seasonal drought prediction: advances, challenges, and future prospects. *Reviews of Geophysics*, 56, 108–141. https://doi.org/10.1002/2016RG000549.

Hegerl, G.C., S. Brönnimann, A. Schurer and T. Cowan. (2018) The early 20th Century warming: anomalies, causes, and consequences. *WIREs Climate Change*, 9, e522. doi:https://doi.org/10.1002/wcc.522.

Ho, C.K., Hawkins, E., Shaffrey, L., Bröcker, J., Hermanson, L., Murphy, J.M., Smith, D.M. and Eade, R. (2013) Examining reliability of seasonal to decadal sea surface temperature forecasts: the role of ensemble dispersion. *Geophysical Research Letters*, 40, 5770–5775. https://doi.org/10.1002/2013GL057630.

Hoerling, M.P., Hurrell, J.W. and Xu, T. (2001) Tropical origins for recent North Atlantic climate change. *Science*, 292, 90–92.

Huang, W.R., Simon Wang, S.-Y. and Guan, B.T. (2018) Decadal fluctuations in the western Pacific recorded by long precipitation records in Taiwan. *Climate Dynamics*, 50, 1597–1608. https://doi.org/10.1007/s00382-017-3707-9.

Jin, Y., X. Rong and Z. Liu. (2017) Potential predictability and forecast skill in ensemble climate forecast: a skill-persistence rule. *Climate Dynamics*, 51, 2725–2742. https://doi.org/10.1007/s00382-017-4040-z.

Johnson, S., Stockdale, T., Ferranti, L., Balmaseda, M., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D. and Weisheimer, A. (2018) SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, 1–44. https://doi.org/10.5194/gmd-2018-228.

Kim, H.M., Webster, P.J. and Curry, J.A. (2012) Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere winter. *Climate Dynamics*, 39, 2957–2973. https://doi.org/10.1007/s00382-012-1364-6.

Kumar, A. (2009) Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Monthly Weather Review*, 137, 2622–2631.

Kumar, A. and Chen, M. (2018) Causes of skill in seasonal predictions of the Arctic Oscillation. *Climate Dynamics*, 51, 2397–2411.

Kumar, A., Barnston, A.G. and Hoerling, M.P. (2001) Seasonal predictions, probabilistic verifications, and ensemble size. *Journal of Climate*, 14, 1671–1676.

Kumar, A., Chen, M., Zhang, L., Wang, W., Xue, Y., Wen, C., Marx, L. and Huang, B. (2012) An analysis of the nonstationarity in the bias of sea surface temperature forecasts for the NCEP Climate Forecast System (CFS) Version 2. *Monthly Weather Review*, 140, 3003–3016. https://doi.org/10.1175/MWR-D-11-00335.1.

Kumar, A., Peng, P. and Chen, M. (2014) Is there a relationship between potential and actual skill? *Monthly Weather Review*, 142, 2220–2227.

Lavers, D., Luo, L. and Wood, E.F. (2009) A multiple model assessment of seasonal climate forecast skill for applications. *Geophysical Research Letters*, 36, L23711. https://doi.org/10.1029/2009GL041365.

L'Heureux, M.L., Tippett, M.K., Kumar, A., Butler, A.H., Ciasto, L.M., Ding, Q., Harnos, K.J. and Johnson, N.C. (2017) Strong relations between ENSO and the Arctic Oscillation in the North American Multimodel Ensemble. *Geophysical Research Letters*, 44, 11654–11662. https://doi.org/10.1002/2017GL074854.

Livezey, R.E. and Chen, W.Y. (1983) Statistical field significance and its determination by Monte Carlo techniques. *Monthly Weather Review*, 111, 46–59. https://doi.org/10.1175/1520-0493.

MacLeod, D., O'Reilly, C., Palmer, T.N. and Weisheimer, A. (2018) Flow dependent ensemble spread in seasonal forecasts of the boreal winter extratropics. *Atmospheric Science Letters*, 19, 1–19, 7. https://doi.org/10.1002/asl.815.

Matsueda, M. and T.N. Palmer. (2018) Estimates of flow-dependent predictability of wintertime Euro-Atlantic weather regimes in medium-range forecasts. *Quarterly Journal of the Royal Meteorological Society*, 144, 1012–1027. doi:https://doi.org/10.1002/qj.3265.

Mehta, V.M., Suarez, M.J., Manganello, J. and Delworth, T.L. (2000) Oceanic influence on the North Atlantic Oscillation and associate Northern Hemisphere climate variations: 1959–1993. *Geophysical Research Letters*, 27, 121–124.

Minobe, S. (1997) A 50–70 year climate oscillation over the North Pacific and North America. *Geophysical Research Letters*, 24, 683–686. https://doi.org/10.1029/97GL00504.

Molteni, F., T. Stockdale, M. Alonso Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, T.N. Palmer and F. Vitart. (2011) *The new ECMWF seasonal forecast system (System 4)*. ECMWF Tech. Memo., 656.

Molteni, F., Stockdale, T. and Vitart, F. (2015) Understanding and modelling extra-tropical teleconnections with the Indo-Pacific region during the northern winter. *Climate Dynamics*, 45, 3119–3140. https://doi.org/10.1007/s00382-0115-2528-y.

Müller, W.A. and Appenzeller, C. (2005) A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *Journal of Climate*, 18, 1513–1523. https://doi.org/10.1175/JCLI3361.1.

Murray, R. (1991) A note on the remarkable winter of 1988/89 over the United Kingdom. *Weather*, 46, 51–53.

O'Reilly, C.H. (2018) Interdecadal variability of the ENSO teleconnection to the wintertime North Pacific. *Climate Dynamics*, 51, 3333–3350. doi:https://doi.org/10.1007/s00382-018-4081-y.

O'Reilly, C.H., Heatley, J., MacLeod, D., Weisheimer, A., Palmer, T.N., Schaller, N. and Woollings, T. (2017) Variability in seasonal forecast skill of Northern Hemisphere winters over the twentieth century. *Geophysical Research Letters*, 44, 5729–5738. https://doi.org/10.1002/2017GL073736.

O'Reilly, C.H., T. Woollings, L. Zanna and A. Weisheimer. (2018a) The impact of tropical precipitation on summertime Euro-Atlantic circulation via a circumglobal wave-train. *Journal of Climate*, 31, 6481–6504. doi:https://doi.org/10.1175/JCLI-D-17-0451.1.

O'Reilly, C.H., Weisheimer, A., Woollings, T., Gray, L. and MacLeod, D. (2018b) The importance of stratospheric initial conditions for winter North Atlantic Oscillation predictability and implications for the signal-to-noise paradox. *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.3413.

Palmer, T.N. and Weisheimer, A. (2011) Diagnosing the causes of bias in climate models – why it is so hard*? Geophysical & Astrophysical Fluid Dynamics*, 105, 351–365. https://doi.org/10.1080/03091929.2010.547194.

Palmer, T.N., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M. and Smith, L.A. (2006) *Ensemble prediction: a pedagogical perspective*. ECMWF Newsletter, 106, 10–17.

Pithan, F., Shepherd, T.G., Zappa, G. and Sandu, I. (2016) Climate model biases in jet streams, blocking and storm tracks resulting from missing orographic drag. *Geophysical Research Letters*, 43, 7231–7240. https://doi.org/10.1002/2016GL069551.

Poli, P., Hersbach, H., Dee, D.P., Berrisford, P., Simmons, A.J., Vitart, F., Laloyaux, P., Tan, D.G.H., Peubey, C., Thépaut, J.-N., Tremolet, Y., Hólm, E.V., Bonavita, M., Isaksen, L. and Fisher, M. (2016) ERA-20C: an atmospheric reanalysis of the 20th Century. *Journal of Climate*, 29, 4083–4097. https://doi.org/10.1175/JCLI-D-15-0556.1.

Ratcliffe, R.A.S. (1989) Review of winter 1988–89 in the Northern Hemisphere. *Weather*, 44, 226–228.

Richardson, D.S. (2001) Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 127, 2473–2489. https://doi.org/10.1002/qj.49712757715.

Rodwell, M. and Doblas-Reyes, F. (2006) Medium-range, monthly and seasonal prediction for Europe and the use of forecast information. *Journal of Climate*, 19, 6025–6046. https://doi.org/10.1175/JCLI3944.1.

Rodwell, M. J., D.S. Richardson, D.B. Parson and H. Wernli. (2018) Flow-dependent reliability: a path to more skilful ensemble forecasts. *Bulletin of the American Meteorological Society*, 99, 1015–1026. doi:https://doi.org/10.1175/BAMS-D-17-0027.1.

Scaife, A.A. and Smith, D. (2018) A signal-to-noise paradox in climate science. *npj Climate and Atmospheric Science*, 1, 28. https://doi.org/10.1038/s41612-018-0038-4.

Scaife, A.A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R.T., Dunstone, N., Eade, R., Fereday, D., Folland, C.K., Gordon, M., Hermanson, L., Knight, J.R., Lea, D.J., MacLachlan, C., Maidens, A., Martin, M., Peterson, A.K., Smith, D., Vellinga, M., Wallace, E., Waters, J. and Williams, A. (2014) Skilful long-range prediction of European and North American winters. *Geophysical Research Letters*, 41, 2514–2519. https://doi.org/10.1002/2014GL059637.

Scaife, A.A., Karpechko, A.Y., Baldwin, M.P., Brookshaw, A., Butler, A.H., Eade, R., Gordon, M., MacLachlan, C., Martin, N., Dunstone, N. and Smith, D. (2016) Seasonal winter forecasts and the stratosphere. *Atmospheric Science Letters*, 17, 51–56. https://doi.org/10.1002/asl.598.

Shi, W., N. Schaller, D. MacLeod, T.N. Palmer and A. Weisheimer. (2015) Impact of hindcast length on estimates of seasonal climate predictability. *Geophysical Research Letters*, 42, 1554–1559. doi:https://doi.org/10.1002/2014GL062829.

Siegert, S., Stephenson, D.B. and Sansom, P.G. (2016) A Bayesian framework for verification and recalibration of ensemble forecasts: how uncertain is NAO predictability? *Journal of Climate*, 29, 995–1012. https://doi.org/10.1175/JCLI-D-15-0196.1.

Sigmond, M., Fyfe, J.C., Flato, G.M., Kharin, V.V. and Merryfield, W.J. (2013a) Seasonal forecast skill of Arctic sea ice area in a dynamical forecast system. *Geophysical Research Letters*, 40, 529–534. https://doi.org/10.1002/grl.50129.

Sigmond, M., Scinocca, J.F., Kharin, V.V. and Shepherd, T.G. (2013b) Enhanced seasonal forecast skill following stratospheric sudden warmings. *Nature Geoscience*, 6, 98–102. https://doi.org/10.1038/ngeo1698.

Stockdale, T.N., Molteni, F. and Ferranti, L. (2015) Atmospheric initial conditions and the predictability of the Arctic Oscillation. *Geophysical Research Letters*, 42, 1173–1179. https://doi.org/10.1002/2014GL062681.

Stockdale, T., Johnson, S., Ferranti, L., Balmaseda, M. and Briceag, S. (January 2018) *ECMWF's new long-range forecasting system SEAS5*. ECMWF Newsletter, 154, 15–20.

Strommen, K. and Palmer, T.N. (2018) Signal and noise in regime systems: a hypothesis on the predictability of the North Atlantic Oscillation. *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.3414.

Tang, Y., Lin, H. and Moore, A.M. (2008) Measuring the potential predictability of ensemble climate predictions. *Journal of Geophysical Research*, 113(D4), D04108. https://doi.org/10.1029/2007JD008804.

Wang, L., Ting, M. and Kushner, P.J. (2017) A robust empirical seasonal prediction of winter NAO and surface climate. *Scientific Reports*, 7, 279. https://doi.org/10.1038/s41598-017-00353-y.

Weisheimer, A. and Palmer, T.N. (2014) On the reliability of seasonal climate forecasts. *Journal of the Royal Society Interface*, 11(96), 20131162. https://doi.org/10.1098/rsif.2013.1162.

Weisheimer, A., Judd, K. and Smith, L.A. (2005) A new view of forecast skill: bounding boxes from the DEMETER ensemble seasonal forecasts. *Tellus A*, 57, 265–279.

Weisheimer, A., Doblas-Reyes, F.J., Palmer, T.N., Alessandri, A., Arribas, A., Deque, M., Keenlyside, N., MacVean, M., Navarra, A. and Rogel, P. (2009) ENSEMBLES – a new multi-model ensemble for seasonal-to-annual predictions: skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophysical Research Letters*, 36, L21711. https://doi.org/10.1029/2009GL040896.

Weisheimer, A., Corti, S., Palmer, T.N. and Vitart, F. (2014) Addressing model error through atmospheric stochastic physical parametrizations: impact on the coupled ECMWF seasonal forecasting system. *Philosophical Transactions of the Royal Society A*, 372(2018), 20130290. https://doi.org/10.1098/rsta.2013.0290.

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D. and Palmer, T.N. (2017) Atmospheric seasonal forecasts of the 20th Century: multi-decadal variability in predictive skill of the winter North Atlantic Oscillation and their potential value for extreme event attribution. *Quarterly Journal of the Royal Meteorological Society*, 143, 917–926. https://doi.org/10.1002/qj.2976.

Wilks, D.S. (1996) Statistical significance of long-range "optimal climate normal" temperature and precipitation forecasts. *Journal of Climate*, 9, 827–839. https://doi.org/10.1175/1520-0442.

Woollings, T., Franzke, C., Hodson, D.L.R., Dong, B., Barnes, E.A., Raible, C.C. and Pinto, J.G. (2014) Contrasting interannual and multidecadal NAO variability. *Climate Dynamics*, 45, 539–556. https://doi.org/10.1007/s00382-014-2237-y.

Zwiers, F.W. (1987) Statistical considerations for climate experiments. Part II: Multivariate tests. *Journal of Climate and Applied Meteorology*, 26, 477–487. https://doi.org/10.1175/1520-0450.

**How to cite this article:** Weisheimer A, Decremer D, MacLeod D, *et al*. How confident are predictability estimates of the winter North Atlantic Oscillation?. *Q J R Meteorol Soc*. 2019;145 (Suppl. 1):140–159. https://doi.org/10.1002/qj.3446