Received: 17 December 2019 | Revised: 5 August 2020 | Accepted: 6 August 2020 | Published on: 15 September 2020

DOI: 10.1002/qj.3890

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Assessing the robustness of multidecadal variability in Northern Hemisphere wintertime seasonal forecast skill

Christopher H. O'Reilly[1,2] | Antje Weisheimer[2,3] | David MacLeod[1] | Daniel J. Befort[1] | Tim Palmer[1]

[1]Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, Oxford, UK

[2]National Centre for Atmospheric Science, University of Oxford, Oxford, UK

[3]European Centre for Medium-Range Weather Forecasts, Reading, UK

Correspondence
C.H. O'Reilly, Atmospheric, Oceanic and Planetary Physics, Clarendon Laboratory, University of Oxford, Sherrington Road, Oxford, OX1 3PU, UK.
Email: christopher.oreilly@physics.ox.ac.uk

Funding information
H2020 European Research Council, Grant/Award Number: 741112; Horizon 2020 Framework Programme, Grant/Award Number: 776613; National Centre for Atmospheric Science

Abstract

Recent studies have found evidence of multidecadal variability in Northern Hemisphere wintertime seasonal forecast skill. Here we assess the robustness of this finding by extending the analysis to a diverse set of ensemble atmospheric model simulations. These simulations differ in either the numerical model or type of initialisation and include atmospheric model experiments initialised with reanalysis data and free-running atmospheric model ensembles. All ensembles are forced with observed sea-surface temperatures (SSTs) and sea-ice boundary conditions. Analysis of large-scale Northern Hemisphere circulation indices over the Northern Hemisphere (namely the North Atlantic Oscillation, the Pacific/North American pattern and the Arctic Oscillation) reveals that in all ensembles there is larger correlation skill in late-century periods than in mid-century periods. Similar multidecadal variability in skill is found in a measure of total skill integrated over the whole extratropical region. Most of the differences in large-scale circulation skill between the skilful late period (as well as the early period) and the less skilful mid-century period seem to be due to a reduction in skill over the North Pacific and a disappearance of skill over North America and the North Atlantic. The results are robust across different models and different types of initialisation, indicating that the multidecadal variability in Northern Hemisphere winter skill is a robust feature of twentieth-century climate variability. Multidecadal variability in skill therefore arises from the evolution of the observed SSTs, likely related to a weakened influence of the El Niño–Southern Oscillation on the predictable extratropical circulation signal during the middle of the twentieth century, and is evident in the signal-to-noise ratio of the different ensembles, particularly the larger ensembles.

KEYWORDS

Arctic Oscillation, large-scale circulation, multidecadal variability, North Atlantic Oscillation, Pacific/North American pattern, predictability, seasonal forecasts

1 | INTRODUCTION

Recent studies have shown that modern seasonal forecasting systems exhibit substantial skill in predicting mean Northern Hemisphere winter circulation, as measured by reforecasts (or hindcasts) of atmospheric circulation anomalies. Perhaps most notable is the skill for the North Atlantic Oscillation (NAO) index, the dominant mode of large-scale circulation variability over the Euro-Atlantic sector, which has been clearly demonstrated in the Met Office's coupled forecasting systems (Scaife *et al.*, 2014; Dunstone *et al.*, 2016). High levels of extratropical skill have also been found for other large-scale circulation indices in other seasonal forecasting systems, such as the Arctic Oscillation index (Riddle *et al.*, 2013; Stockdale *et al.*, 2015). However, these examples of high levels of skill have all been demonstrated over a relatively short hindcast period, typically of 30 years or less.

Recent work has sought to put these results into a broader context by assessing skill in experiments performed over longer periods. Weisheimer *et al.* (2017) presented results from initialised wintertime ensemble atmospheric model simulations over the whole of the twentieth century, using an atmospheric model with prescribed sea-surface temperatures (SSTs) and sea-ice boundary conditions. The 30-year running correlation skill of the NAO in these experiments varies significantly over the twentieth century. In particular, there is significant ensemble mean correlation skill for the NAO in the early twentieth century and the most recent period, but there is an apparent reduction in skill during the mid-twentieth century. In a study of the same dataset, O'Reilly *et al.* (2017) showed that the mid-twentieth-century drop in skill is present, and more dramatic, for the Pacific/North American pattern (PNA). MacLeod *et al.* (2018) also found a high root-mean-square error for the PNA in the mid-century period. The drop in PNA skill – which is closely linked to NAO skill during this period – seems to be due to a weakening of the observed wintertime El Niño–Southern Oscillation (ENSO) teleconnection to the extratropics during this mid-twentieth-century period (O'Reilly, 2018).

Whilst these studies have presented some evidence of multidecadal variability in winter seasonal forecast skill over the twentieth century, these were based on a single initialised atmospheric model ensemble, referred to hereafter as ASF-20C (i.e., "Atmospheric Seasonal Forecasts of the Twentieth Century"). One limitation of these experiments is that they were initialised from a long reanalysis dataset, ERA-20C (Poli *et al.*, 2016), which only assimilates surface observations and is therefore inferior to modern reanalyses available over the satellite era. The upper-atmospheric data that is assimilated into these more comprehensive (but shorter) reanalysis products has been shown to improve wintertime forecast skill (Stockdale *et al.*, 2015; Nie *et al.*, 2019; O'Reilly *et al.*, 2019), so it is plausible that variations in the quality of the atmospheric initial conditions (from ERA-20C) used in the ASF-20C simulations may be a source of the multidecadal variability in skill. Moreover, the variability in skill across the twentieth century is not clearly statistically significant for some fields in the ASF-20C simulations, so it is interesting to examine whether similar features are observed in different ensemble simulations. In addition, some specific doubt has been cast over the robustness of the multidecadal variability in wintertime seasonal skill by Kumar and Chen (2018), who find no evidence of variability in the ensemble mean correlation skill of the Arctic Oscillation using a different atmospheric model, although this analysis was over a much shorter period (from 1957 onwards).

In this article we analyse the robustness of the multidecadal variability in seasonal forecast skill for Northern Hemisphere winters across different ensemble simulations, and in particular how this depends on ensemble initialisation. We compare seasonal atmospheric model ensemble simulations – initialised in different ways – with several free-running atmospheric model simulations run over the whole of the twentieth century, all performed with prescribed SSTs and sea-ice boundary conditions from observations. Our aim is to assess how much of the multidecadal variability in wintertime predictability comes from the observed SST variability alone and whether this is dependent on the type of model or the initialisation method.

2 | DATASETS AND METHODS

2.1 | Ensemble atmospheric model simulations

In this study we analyse data from six atmospheric model ensembles, which are listed in Table 1. The first ensemble which we will analyse is the ASF-20C simulations, which were initialised from the ERA-20C reanalysis on November 1 of each year from 1901 to 2009 (Weisheimer *et al.*, 2017). These atmospheric model simulations were performed with the ECMWF's Integrated Forecast System (IFS) model, using 51 ensemble members, and their performance has been analysed in several other studies (e.g., O'Reilly *et al.*, 2017; MacLeod *et al.*, 2018; Weisheimer *et al.*, 2018; Parker *et al.*, 2019). The second ensemble also consists of initialised seasonal atmospheric model simulations, using the same system as the ASF-20C simulations (though with 25 ensemble members). However, in this ensemble, the initial conditions for all ensemble members for a given year were taken from November 1 from the

TABLE 1 List of atmospheric model datasets analysed in this study, including the number of ensemble members in each dataset

Experiment/model	Initialisation type and simulation details	Ensemble members
ASF-20C (IFS)	Initialised on November 1 of each year from ERA-20C	51
ASF-RAND (IFS)	Initialised from November 1 of randomly selected years from ERA-20C	25
ERA-20CM (IFS)	Free-running AMIP-style simulation	10
CanESM5	Free-running AMIP-HIST experiment (from CMIP6 archive)	10
CNRM-CM6	Free-running AMIP-HIST experiment (from CMIP6 archive)	9
CFSv2	Free-running AMIP-style simulation (1957–2009)	101

Note: All the datasets cover the period 1901–2009 except for CFSv2 which covers only 1957–2009. The ASF-20C, ASF-RAND and ERA-20CM ensembles all use a version of ECMWF's IFS model.

ERA-20C reanalysis for a different year selected each time at random. These randomised years were selected without replacement such that each November 1 initial condition (for each year over the period 1900–2009) was used for one season of simulations. The SST boundary condition is identical to ASF-20C. We hereafter refer to this random initial condition experiment as "ASF-RAND". ASF-RAND will allow us to test how important the initial conditions used in the ASF-20C ensemble are for determining the multidecadal variability in skill.

Four other ensemble datasets are analysed, all of which are free-running atmospheric model simulations with prescribed SSTs (i.e., "AMIP-style" simulations). In these datasets, therefore, there is no information from any initial conditions for a given season, and any decadal variability in skill is not due to variation in initialisation quality. Instead, we can interpret the ensemble mean anomalies as being forced by the surface boundary conditions alone. The first of these free-running ensemble simulations is the ERA-20CM dataset. This was performed using the same model (ECMWF's IFS) as that used to produce the ERA-20C reanalysis but without any data assimilation, instead being forced only by prescribed SSTs and external forcing (Hersbach *et al.*, 2015). The next two ensemble datasets were taken from the Coupled Model Intercomparison Project 6 (CMIP6) database, specifically the "AMIP-HIST" experiments from the Global Monsoons Model Inter-comparison Project (GMMIP; Zhou *et al.*, 2016). These experiments include the natural and anthropogenic (external) historical forcings used in CMIP6 historical simulations, with prescribed SSTs and sea-ice boundary conditions from HadISST (Rayner *et al.*, 2003). Two models from the "AMIP-HIST" database were included that had the most ensemble members: CanESM5, with 10 ensemble members (Gillett *et al.*, 2019), and CNRM-CM6, with nine ensemble members (Voldoire *et al.*, 2019). Ensemble sizes of 10 are still quite small in the context of seasonal hindcast experiments, so these ensembles might be expected to have some difficulty in capturing some of the smaller predictable signals, particularly over the extratropical North Atlantic (e.g., Scaife *et al.*, 2014). At the time of the analysis, no other models that had uploaded AMIP-HIST experiments had generated more than five ensemble members and were thus not included in our analysis. The fourth free-running atmospheric model simulation we analyse is a 101-member ensemble performed using the CFSv2 model (Kumar and Chen, 2018). However, because the simulations were only performed from 1957 onwards, data from the CFSv2 ensemble is taken from the period 1957–2009, whereas for all the other datasets we use data from the period 1901–2009.

Whilst the prescribed surface boundary conditions are important for directly comparing the responses across these different ensembles, it is important to note that such a model set-up can be misleading in some instances. The reason for this is that, in some regions, SST anomalies that are generated by anomalous atmospheric circulation actually act to force the atmosphere above when prescribed as a boundary condition (e.g., Bretherton and Battisti, 2000). An important example of this is over the Indian Ocean, as shown by Copsey *et al.* (2006). Therefore, the results from the prescribed SST simulations should be considered with these caveats in mind. The potential impact of using prescribed surface boundary conditions on assessing predictability is demonstrated in several studies by relaxing the tropical atmosphere towards observations, thereby directly prescribing the correct atmospheric diabatic heating anomalies and circumventing the potentially erroneous air–sea coupling; this results in more accurate circulation anomalies in the extratropics (e.g., Greatbatch *et al.*, 2012; 2015; Hansen *et al.*, 2017). Despite the potential limitations, the use of the same prescribed surface boundary conditions in the ensembles used in the present study is beneficial as it allows for a comparison of the responses across the different models and different initialisation methods.

We use the ERA-20C reanalysis (Poli *et al.*, 2016) as a reference dataset; in addition to this we also use the

HadSLP2 dataset (for sea-level pressure [SLP]; Allan and Ansell, 2006). The HadSLP2 dataset is produced using only atmospheric pressure observations to reconstruct a gridded product using only statistical methods. This is quite different from the ERA-20C dataset, which is produced by assimilating surface pressure and wind observations into a model with prescribed surface boundary conditions from observed SSTs. A comparison of the results from these two datasets is used to assess the robustness of the results that follow. A preliminary analysis of the grid-point correlation between the wintertime (December–February; DJF) SLP data from the two datasets reveals high levels of agreement over much of the northern extratropics (Figure S1 in the Supporting Information) and the conclusions that follow are not qualitatively sensitive to the choice of reference dataset.

For the geopotential height calculations, the data were detrended prior to computing the indices and correlation skill statistics. However, we found that this actually made little difference to the results because the magnitude of the trends was very small compared to the interannual variability.

2.2 | Atmospheric circulation indices

In this study we analyse several different measures of the extratropical atmospheric circulation over the whole of the boreal winter season (defined as the DJF average). These measures are defined as follows.

- *North Atlantic Oscillation (NAO).* Here we define the NAO (e.g., Hurrell *et al.*, 2003) as the (normalised) principal component time series of the first empirical orthogonal function (EOF) of 500 hPa geopotential height anomalies (i.e., Z500) over the Euro-Atlantic sector (90°W–30°E, 30–90°N). The reference NAO index was calculated from the ERA-20C reanalysis. NAO indices were calculated from each ensemble member in each of the datasets by projecting the model Z500 anomalies onto the reference NAO pattern from the ERA-20C reanalysis.
- *Pacific/North American (PNA) index.* We use a grid-point definition of the PNA, following Wallace and Gutzler (1981), as $\mathrm{PNA} = \frac{1}{4}[Z'(20°\mathrm{N}, 160°\mathrm{W}) - Z'(45°\mathrm{N}, 165°\mathrm{W}) + Z'(55°\mathrm{N}, 115°\mathrm{W}) - Z'(30°\mathrm{N}, 85°\mathrm{W})]$, where Z' is the geopotential height anomaly at 500 hPa. The PNA indices are calculated for the ERA-20C reanalysis and for all ensemble members in each of the datasets.
- *Arctic Oscillation (AO).* Here we define the AO as the (normalised) principal component time series of the first EOF of mean SLP anomalies polewards of 20°N. The reference AO index was calculated from the HadSLP2 dataset, using the same region as defined in Thompson and Wallace (1998). AO indices were calculated from each ensemble member in each of the datasets by projecting the model SLP anomalies onto the reference AO pattern from the HadSLP2 dataset. All analysis for the AO was also repeated for the ERA-20C SLP, and comparison between HadSLP and ERA-20C AO is included in the Supporting Information (Figures S2, S3 and S5) and also discussed in the results section below.
- *Total fraction of explained variance (EV) in the extratropics.* In addition to the three common indices for characterising extratropical circulation variability, we also calculated the total fraction of extratropical Z500 (or SLP) variance explained in each of the ensembles. This was calculated as

$$\mathrm{EV} = \frac{\iint_A r^2 \overline{Z'^2} dA}{\iint_A \overline{Z'^2} dA},$$

where $\overline{Z'^2}$ is the grid-point geopotential height variance, r is the ensemble correlation skill (such that r^2 is the EV at each grid point, taken to be zero where r is negative) and A is the area over which this is integrated. To focus on the extratropical circulation anomalies we calculate the EV over the entire region polewards of 30°N. The EV gives a more general measure of the skill of the extratropical circulation than those that are focused on single-circulation indices, and is therefore complementary to the more common large-scale indices described above.

3 | RESULTS AND DISCUSSION

We begin by assessing the correlation between the observed NAO index and the ensemble mean NAO indices in the different datasets. This is shown for moving 30-year periods, as well as the full 1901–2009 period, in Figure 1a. The shading in the plots of Figure 1 indicates the sampling uncertainty of the 30-year correlation skill values for the ASF-20C ensemble. The sampling uncertainty is calculated by resampling years in the sample with replacement – the process is repeated 10,000 times to generate uncertainty estimates. The sampling uncertainty is used here because it is generally substantially higher than the ensemble uncertainty, which can be calculated by resampling over ensemble members with replacement (an example of this is shown in Figure S4 in the Supporting Information). The uncertainty is similar in the

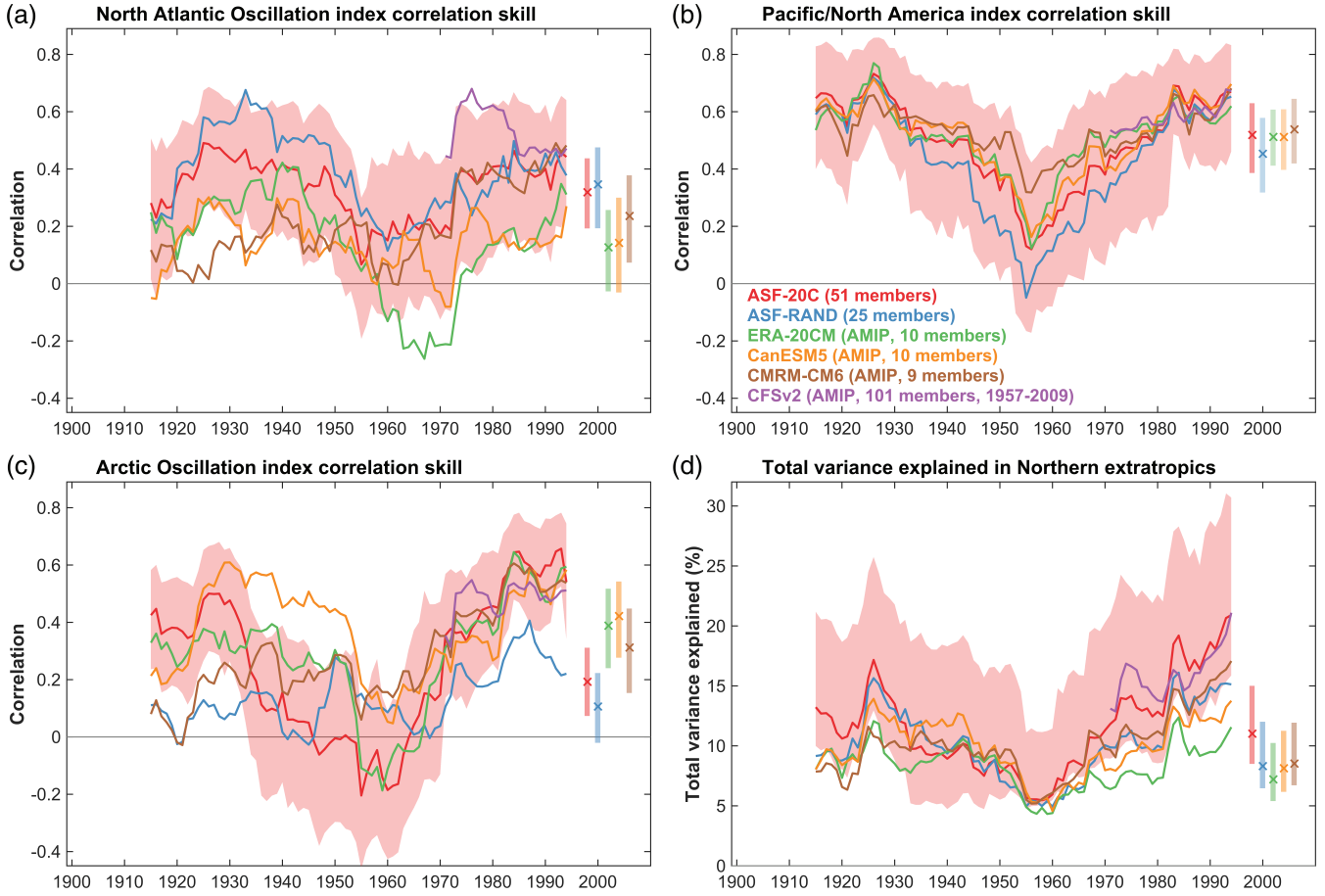

FIGURE 1 (a) The North Atlantic Oscillation (NAO) ensemble mean correlation skill over moving 30-year windows, shown for all model ensemble datasets. The red shading shows the 5–95% uncertainty range for the ASF-20C dataset based on a random bootstrap resampling over the years in each sample (performed 10,000 times). The crosses on the right-hand side of each plot show the overall correlation skill over the period 1901–2009. (b) As for panel (a), but for the Pacific/North America pattern (PNA) correlation skill. (c) As for panel (a), but for the Arctic Oscillation (AO) index correlation skill. (d) The total fraction of the 500 hPa geopotential height anomalies (Z500) explained by each ensemble dataset in moving 30-year windows (defined in Section 2.2)

other ensembles, but to aid clarity we only show the uncertainty for the ASF-20C ensemble in these plots. Towards the end of the analysis period the ASF-20C correlation skill is around $r \approx 0.4$, and the ASF-RAND, CNRM-CM6 and CFSv2 models all exhibit similar levels of skill. As shown in Weisheimer *et al.* (2017), the NAO correlation skill in the ASF-20C simulations drops in the mid-century period, centred around 1960 ($r < 0.2$). Although the difference between the late period and this mid-century period is not clearly statistically significant by itself (because the confidence intervals overlap) it is useful to compare it to the ensembles produced using different models and initialisation methods in an attempt to examine whether there is agreement across the ensembles. During this same mid-century period, none of the other ensemble datasets exhibit any significant skill; in fact, all are lower than in ASF-20C. Most of the datasets, regardless of initialisation type, exhibit higher levels of correlation skill in the later periods than in the mid-century period. Some of the datasets also display slightly higher levels of NAO correlation skill in the early-century periods – specifically ASF-20C, ASF-RAND, ERA-20CM and CNRM-CM6 – although in some of these cases the skill level is low and the differences are marginal. Nonetheless, it is clear that the mid-century drop in NAO skill seen in ASF-20C is not unique to this simulation and is a robust feature across the ensembles. Therefore, the drop in NAO skill during the mid-century period is not due to a change in the quality of the initialisation because other datasets – with either no initialisation or incorrect initial conditions – also exhibit higher skill in the later periods as compared to the mid-century periods.

In addition to the EOF-based NAO, we also repeated the same analysis for a grid-point-based NAO index, defined as the difference in the normalised SLP between Iceland and the Azores (following, e.g., Hurrell *et al.*, 2003); this is shown in Figure S3 in the Supporting Information for both the HadSLP2 and ERA-20C datasets. The

overall behaviour of the skill over the twentieth century is qualitatively similar to that seen in Figure 1a for the EOF-based index. The overall skill is generally lower than the EOF-based NAO across all the datasets. This is consistent with the results of Baker *et al.* (2018), who found that the skill of the grid-point-based NAO is typically lower than in NAO indices derived over larger regions.

We next analyse the variability in the PNA correlation skill. Previous analysis of the ASF-20C dataset showed that the PNA correlation skill exhibits particularly pronounced variability over the twentieth century (O'Reilly *et al.*, 2017). Figure 1b shows the PNA ensemble mean correlation skill for moving 30-year periods for all the ensemble datasets. All ensemble datasets exhibit similar multidecadal variability in the PNA correlation skill, with high levels of correlation skill in the early and late periods and a distinct drop in correlation skill during the mid-century period. The overall skill in the atmospheric ensemble simulations is very similar, despite the different models and initialisation types (or lack thereof). O'Reilly *et al.* (2017) showed that this drop in correlation skill closely follows the observed correlation between the PNA and the observed Niño-3 SST anomalies, which exhibited a weaker teleconnection to the extratropical North Pacific during the mid-twentieth century (Minobe and Mantua, 1999; Diaz *et al.*, 2001; O'Reilly, 2018). Comparing the PNA correlation skill with the NAO correlation skill in Figure 1a, it is clear that during the mid-century period there is a common period around the late 1950s/early 1960s when there is substantially less predictability coming from the SST boundary conditions over both the Euro-Atlantic and North Pacific sectors. Since we see similar variability in skill across all the ensembles, including those with an incorrect initial condition (i.e., ASF-RAND) and with no initialisation (i.e., ERA-20CM, CanESM5 and CNRM-CM6), we can conclude that the mid-century drop in skill is independent of the initialisation type.

The AO index is defined over a region covering the entire northern extratropics, which is larger than the NAO and PNA regions. The AO correlation skill over moving 30-year periods, shown in Figure 1c, exhibits multidecadal variability that is broadly similar to the PNA and the NAO across all the datasets. The most notable similarity is the drop in skill in the mid-century period, in which none of the different ensemble simulations exhibit any significant skill, specifically during periods centred on the late 1950s/early 1960s. The similarities with the behaviour seen in the NAO and PNA correlation skill is perhaps not surprising since the centres of action of the AO closely correspond to those that feature in both the NAO and PNA patterns (e.g., Thompson and Wallace, 1998).

The variability in AO index correlation is broadly similar when using the ERA-20C SLP data as the reference dataset (shown in Figure S5 in the Supporting Information). This is in spite of the fact that the reference AO patterns and indices exhibit some substantial differences (Figures S2 and S3 in the Supporting Information). For example, the correlation of the AO indices calculated from the HadSLP2 and ERA-20C SLP datasets is $r = 0.66$, despite grid-point correlations between the two datasets exceeding 0.6 over almost the entire Northern Hemisphere extratropical region (i.e., Figure S1 in the Supporting Information). The differences between the AO indices from the two datasets seem to originate from the different SLP patterns, which show clear differences in the relative magnitude and pattern, particularly over the Eurasian continent.

In a recent study, Kumar and Chen (2018) found no notable variability in the AO correlation skill in the CFSv2 ensemble, which we also examine here. Looking more closely at the CFSv2 ensemble (purple lines in Figure 1), it is clear that because the CFSv2 ensemble was only integrated from 1957 onwards it does not cover the mid-century period, during which there is a notable drop in correlation skill in the other ensembles. Interestingly, the CFSv2 ensemble does exhibit larger skill than some of the other ensembles for both the AO and NAO indices during periods in the latter part of the century, but for the PNA the skill is very similar. The higher levels of NAO and AO skill in the CFSv2 ensemble are not simply due to the larger ensemble size (Figure S6 in the Supporting Information). Nonetheless, the multidecadal variability seen in the five ensembles that cover the entire twentieth century is not contradicted by the results from the shorter CFSv2 ensemble.

The variability of the ensemble mean correlation skill seen in the NAO, PNA and AO indices suggests that the mid-century drop in correlation skill is not due to variability in initialisation quality and is robust across different atmospheric models. However, each of these indices targets a particular large-scale atmospheric circulation pattern. It is possible, for example, that during this mid-century period there were other dominant patterns of large-scale variability that exhibit larger correlation skill. To examine how universal the variability in skill is across the northern extratropics, we calculate the total fraction of EV in the extratropics (as described above in Section 2.2), which is plotted in Figure 1d (the equivalent plots for SLP data from the HadSLP2 and ERA-20C datasets are shown in Figure S7 in the Supporting Information). In all of the ensemble datasets, the largest fraction of observed variance is generally explained during the later periods of the simulations, peaking at just over 20% in the ASF-20C and CFSv2 ensembles. Substantial fractions of explained extratropical circulation variance are also seen during the early-century periods in all of the ensemble simulations. Perhaps most notable, however, is the fact that all the

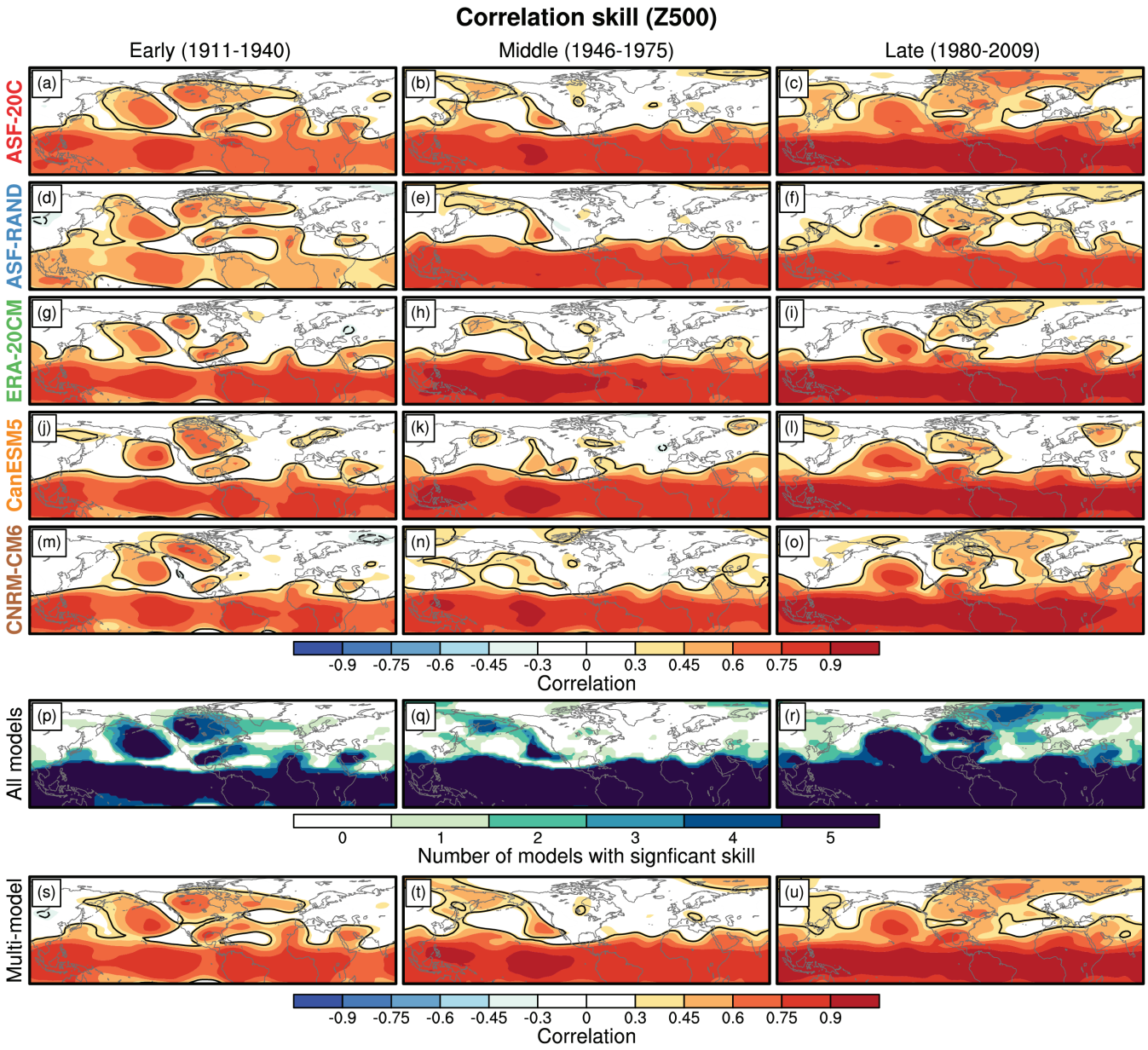

FIGURE 2 (a–o) Correlation skill of the ensemble mean 500 hPa geopotential height anomalies (Z500) in each of the datasets with the ERA-20C reanalysis, for the early period (1911–1940), middle period (1946–1975) and late period (1980–2009). The black contours show where the correlation skill is significant at the 95% confidence level according to a two-sided t-test. (p–r) The number of model ensembles (out of the five models above) that exhibit significant skill at the 95% confidence level. (s–u) As for panels (a–o), but for the correlation skill of the multi-model ensemble mean

ensembles exhibit similarly low levels of EV during periods centred around the late 1950s. Similar behaviour is also seen in the evolution of the total EV of SLP (Figure S7 in the Supporting Information). Overall, these results suggest that the drop in skill is a robust feature of the amount of seasonal predictability emanating from the prescribed SST boundary conditions in all of these ensembles, in spite of the differences in atmospheric models and initialisation methods.

The variability in skill over the twentieth century can be further demonstrated by analysing maps of Z500 correlation skill in 30-year periods during the early (1911–1940), middle (1946–1975) and late century (1980–2009), shown in Figure 2. The skill across the range of ensembles is summarised in the bottom two rows of Figure 2, which shows how many of the five ensembles exhibit significant correlation skill (at the 95% level) during each of the three periods, as well as the multi-model ensemble mean correlation skill. In the extratropics there are high levels of skill evident in the early period over the North Pacific and North America (left column), with some notable skill also over the North Atlantic in ASF-20C and

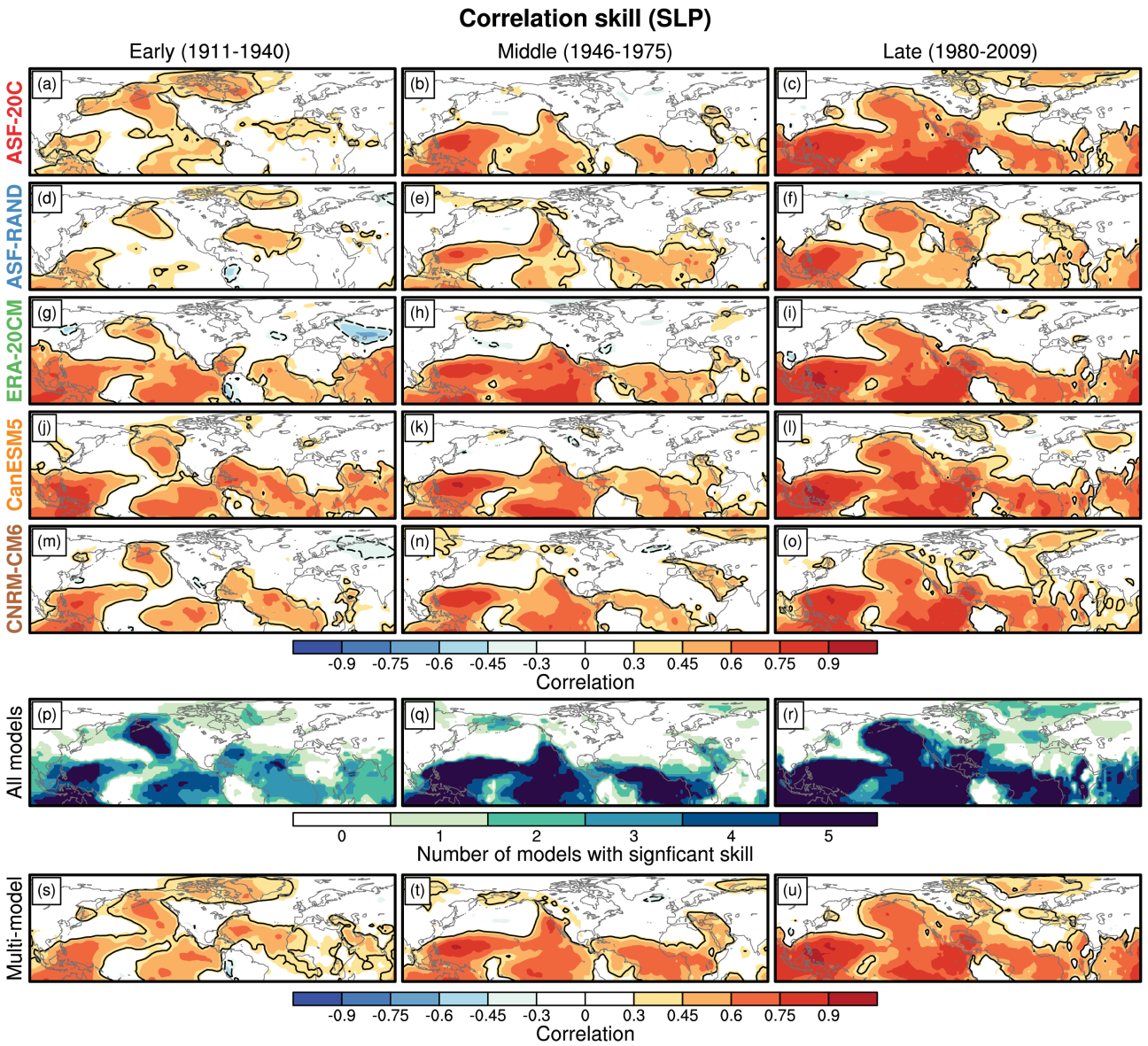

FIGURE 3 (a–o) Correlation skill of the ensemble mean sea-level pressure (SLP) anomalies in each of the datasets with the HadSLP2 dataset, for the early period (1911–1940), middle period (1946–1975) and late period (1980–2009). The black contours show where the correlation skill is significant at the 95% confidence level according to a two-sided t-test. (p–r) The number of model ensembles (out of the five models above) that exhibit significant skill at the 95% confidence level. (s–u) As for panels (a–o), but for the correlation skill of the multi-model ensemble mean. It is notable that the locations of largest skill in the Atlantic are not located directly over the centres chosen for the grid-point NAO (i.e., Iceland and the Azores), which may explain why the grid-point NAO correlations (i.e., Figure S5 in the Supporting Information) are typically lower than the empirical orthogonal function (EOF)-based NAO correlations (i.e., Figure 1a)

ASF-RAND, likely due to the larger number of members in these ensembles. There are also high levels of skill across all ensembles over the North Pacific and North America during the late period (right column); in addition, there are notable levels of skill over the North Atlantic in most of the ensembles during the late period. During the mid-century period, however, there are only small areas over the extratropics that exhibit skill in many of the ensembles. In particular, over North America and the North Atlantic almost none of the ensembles exhibit any significant skill.

Similar results are evident in the equivalent maps of SLP correlation skill, using the HadSLP2 dataset as the reference, which are shown in Figure 3 (the equivalent maps for ERA-20C SLP are very similar and are shown in Figure S8 in the Supporting Information). Again, there are significant levels of correlation skill across all ensembles over the North Pacific during the early and late periods, and also significant skill over the North Atlantic in several of the ensembles during the early and late periods. During the mid-century period, however, there are only very small

areas that exhibit significant correlation skill in any of the ensembles, similarly to the maps of Z500 correlation skill (i.e., Figure 2). The correlation skill completely disappears over the North Atlantic in all of the ensembles during the mid-century period.

The maps shown in Figures 2 and 3 are wholly consistent with the variability in skill of the different indices discussed above and shown in Figure 1. During the mid-century period there is less skill in capturing the observed circulation anomalies in the ensembles forced by surface boundary conditions. One conclusion from this is that the potentially predictable signal from SSTs in this mid-century period is lower than in the late-century period and, to a certain extent, lower than in the early-century period as well. However, it is possible that there are common errors in the SST dataset used in all the ensembles, which all used a version of the HadISST dataset (Rayner *et al.*, 2003). To examine this we calculated the signal-to-noise ratio of the extratropical Z500 anomalies, area-averaged over the entire region polewards of 30°N, in each of the ensembles. Specifically, the signal-to-noise ratio is defined as $\sigma_{em}/\langle\sigma_e\rangle$, where σ_{em} is the standard deviation of the model ensemble mean and $\langle\sigma_e\rangle$ is the average of the standard deviation in each of the individual ensemble members.

The signal-to-noise ratio in each of the ensembles is plotted in Figure 4a for moving 30-year periods. The signal-to-noise ratios were calculated for resampled ensemble sizes of 10 for ensembles with more than 10 members to ensure a sensible comparison of the ensemble size (see, e.g., Scaife and Smith, 2018). In all of the ensembles, there is a larger signal-to-noise ratio in the late-century than in the mid-century period. In the two larger ensembles covering the entire period, namely ASF-20C and ASF-RAND, the signal-to-noise ratio is also larger in the early-century period than during the mid-century period. The variability of the signal-to-noise ratio in the ensembles can, in one sense, be considered a measure of "potential" predictability, as it is not measured relative to observations. The signal-to-noise variability over the twentieth century in the ensembles clearly demonstrates that there is less forced predictable signal during the mid-century period, centred around the late 1950s, as compared to the later periods. This indicates that the drop in correlation skill seen in all ensembles in the mid-century period represents a drop in predictability driven by the observed evolution of SSTs.

The drop in predictable signal coming from the SST boundary conditions is likely linked to the variance of the tropical Pacific SSTs, shown in Figure 4b. During the late-century period there is a much greater variance in the Niño-3.4 SST index, which is known to be a major source of skill in seasonal forecasts (e.g., Smith *et al.*, 2012; Dunstone

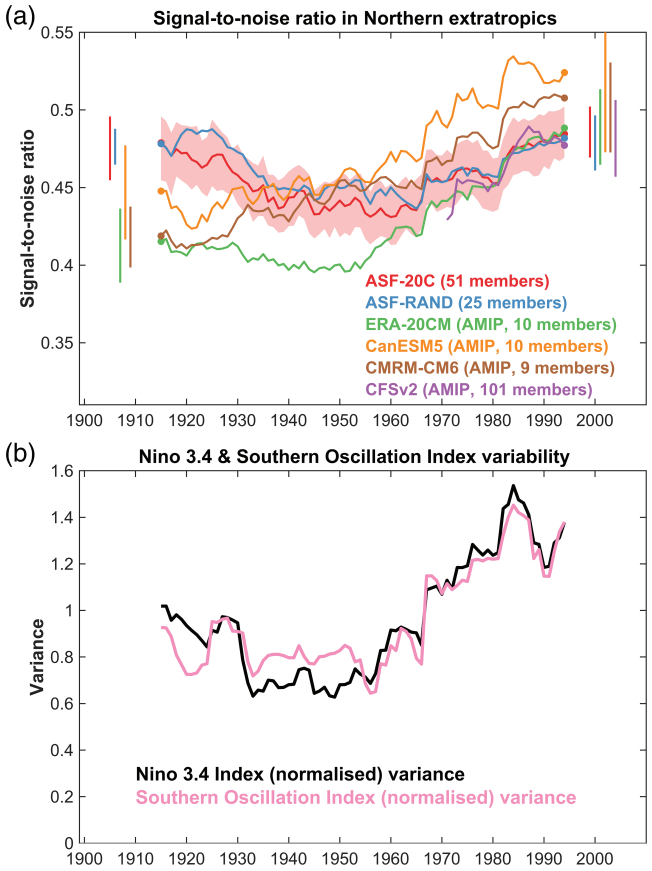

FIGURE 4 (a) The signal-to-noise ratio of the 500 hPa geopotential height anomalies in each ensemble dataset over moving 30-year periods (see the text for further details). To ensure a fair comparison, for the ensembles with more that 10 ensemble members, 10 members were sampled randomly without replacement and the signal-to-noise ratio was calculated for this subsample of the ensemble. This process was repeated 10,000 times and the median of the resulting distribution is shown. The red shading shows the 5–95% uncertainty range for the ASF-20C dataset based on the random bootstrap resampling. The vertical lines show the 5–95% uncertainty range for each of the ensemble datasets for the first and last 30-year periods (indicated by the filled circles). (b) Variance, in 30-year periods, of the observed Niño-3.4 SST index and the observed Southern Oscillation Index (defined as the Tahiti–Darwin SLP index, following Ropelewski and Jones, 1987, and downloaded from https://crudata.uea.ac.uk/cru/data/soi/). The time series were both normalised over the entire 1901–2009 period before the variance was calculated over each 30-year window

et al., 2016). Previously, Greatbatch and Jung (2007) found that there was a reduction in the predictable skill from the tropics towards the mid-twentieth century. O'Reilly (2018) showed that the observed link between the tropical Pacific SSTs and the atmospheric circulation in the extratropical Pacific was substantially weaker in the mid-century period, which is clearly replicated here in the ensemble simulations with prescribed SSTs. The link to the North Atlantic skill is less direct, but O'Reilly *et al.* (2017) showed

that in the ASF-20C ensemble during the mid-century period, winters with the largest forecast error over the North Atlantic were typically associated with large forecast errors over the extratropical North Pacific.

It is possible that the larger errors in SST observations evident around the mid-century period (e.g., Thompson *et al.*, 2008; Kennedy *et al.*, 2011; Davis *et al.*, 2019) contribute to the lower predictability in the ensembles during this period. To examine this hypothesis, we have analysed the variance in the Southern Oscillation Index (SOI; defined as the Tahiti–Darwin SLP difference), shown in Figure 4b. The SOI demonstrates remarkably similar changes in variance as the Niño-3.4 SST index. Since these two indices are produced using independent observations, this provides strong evidence that the drop in tropical Pacific variability in the mid-century period is a robust feature, which is likely responsible for the apparent drop in predictability during this period. Further examination of the variance of the two indices reveals that for periods centred on the 1930s and 1940s there is lower variance in the Niño-3.4 SST index than in the SOI, suggesting that the data issues may be limiting the SST variability and hence the predictable signals seen in the ensembles during this period. However, during the late 1950s and in later periods, there is a clear correspondence between the variance of the Niño-3.4 index and the SOI, indicating that there is no SST variance that is obviously missing during this period. Since these two indices are produced using independent observations, this provides strong evidence that the drop in tropical Pacific variability during the mid-century period around 1960 is a robust feature, which is likely responsible for the apparent drop in predictability during this period.

4 | CONCLUDING REMARKS

In this article we have examined how the multidecadal variability in seasonal forecast skill of Northern Hemisphere wintertime circulation depends on the type of initialisation and atmospheric model. We analysed a set of ensemble atmospheric model simulations, including atmospheric model simulations initialised with reanalysis data (performed with prescribed SST and sea-ice boundary conditions) and free-running atmospheric model simulations constrained by observed ocean conditions. Analysis of indices of large-scale circulation over the Northern Hemisphere, namely the NAO, PNA and AO, demonstrate significant correlation skill over the whole of the twentieth century. The ensembles generally exhibit higher correlation skill in the late-century periods than in mid-century periods, for all the circulation indices. Similar multidecadal skill variability is found in a more general measure of overall skill – the total explained Z500 variance – which encompasses the entire extratropical region. Maps of the correlation skill in the early-, mid- and late-century periods show that most of the differences in skill between the early and late periods and the less skilful mid-century period is due to a reduction in skill over the North Pacific and a disappearance in skill over North America and the North Atlantic.

The results are robust across different models and different initialisation types, indicating that the multidecadal variability in skill is a robust feature of twentieth-century climate variability for Northern Hemisphere winters. This is most likely related to a weakening of the influence of the ENSO on the predictable signal during the middle of the twentieth century (e.g., O'Reilly *et al.*, 2017), which is evident in the signal-to-noise ratio of the different ensembles, and in particular the larger ensembles analysed here, which exhibit lower predictable signals of extratropical large-scale circulation anomalies during the mid-century period. Whilst all these ensembles use prescribed SSTs, in a recent study Weisheimer *et al.* (2020) presented a coupled hindcast using a similar set-up to the ASF-20C but coupled to a version of the NEMO ocean model (and initialised from a coupled reanalysis). This coupled hindcast experiment also shows similar variations in NAO and PNA correlation skill over the twentieth century, with a distinct minimum during the mid-century period, demonstrating that this variability in skill is not merely a feature of atmospheric model ensembles with prescribed SST boundary conditions.

The results presented here have important implications for assessing the expected skill of seasonal forecasting systems. Using only the most recent periods to evaluate the skill of such systems, as is typically done, should be expected to give higher levels of skill. Estimates of skill from hindcast simulations covering a recent period of ≈30 years are subject to a large amount of uncertainty due to the relatively small sample size (e.g., Figure 1). Our results here suggest that if hindcasts of seasonal forecast models were performed over longer periods, extending further back in time, then the uncertainty in the skill estimates would be reduced and we would expect the overall estimates of skill to be lower than in hindcasts performed over only a recent period. However, the existence of substantial low-frequency variability in skill would suggest that hindcasts over the most recent periods might be more representative of current forecast skill.

ACKNOWLEDGEMENTS

We thank Dr. Arun Kumar and colleagues from the Climate Prediction Center (NCEP) for providing the CFSv2 ensemble dataset. We also thank the modelling centres that provided the AMIP-HIST datasets and made them available as part of the GMMIP.

We thank the editor and two anonymous reviewers who provided very insightful and helpful comments on previous versions of the manuscript.

ORCID

Christopher H. O'Reilly https://orcid.org/0000-0002-8630-1650
Antje Weisheimer https://orcid.org/0000-0002-7231-6974
Daniel J. Befort https://orcid.org/0000-0002-2851-0470
Tim Palmer https://orcid.org/0000-0002-7121-2196

REFERENCES

Allan, R. and Ansell, T. (2006) A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2): 1850–2004. *Journal of Climate*, 19, 5816–5842.

Baker, L., Shaffrey, L., Sutton, R., Weisheimer, A. and Scaife, A. (2018) An intercomparison of skill and overconfidence/underconfidence of the wintertime North Atlantic Oscillation in multimodel seasonal forecasts. *Geophysical Research Letters*, 45, 7808–7817.

Bretherton, C.S. and Battisti, D.S. (2000) An interpretation of the results from atmospheric general circulation models forced by the time history of the observed sea surface temperature distribution. *Geophysical Research Letters*, 27, 767–770.

Copsey, D., Sutton, R. and Knight, J.R. (2006) Recent trends in sea level pressure in the Indian Ocean region. *Geophysical Research Letters*, 33, L19712.

Davis, L.L., Thompson, D.W., Kennedy, J.J. and Kent, E.C. (2019) The importance of unresolved biases in twentieth-century sea surface temperature observations. *Bulletin of the American Meteorological Society*, 100, 621–629.

Diaz, H.F., Hoerling, M.P. and Eischeid, J.K. (2001) ENSO variability, teleconnections and climate change. *International Journal of Climatology*, 21, 1845–1862.

Dunstone, N., Smith, D., Scaife, 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.

Gillett, J.A., Arora, V., Christian, J.R., Hanna, S., Jiao, Y. and Lee, W.G. (2019) *The Canadian Earth System Model version 5 (CanESM5. 0.3)*. https://gmd.copernicus.org/articles/12/4823/2019/

Greatbatch, R.J., Gollan, G., Jung, T. and Kunz, T. (2012) Factors influencing Northern Hemisphere winter mean atmospheric circulation anomalies during the period 1960/61 to 2001/02. *Quarterly Journal of the Royal Meteorological Society*, 138, 1970–1982.

Greatbatch, R.J., Gollan, G., Jung, T. and Kunz, T. (2015) Tropical origin of the severe European winter of 1962/1963. *Quarterly Journal of the Royal Meteorological Society*, 141, 153–165.

Greatbatch, R.J. 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 North Atlantic Oscillation predictability via the stratosphere. *Quarterly Journal of the Royal Meteorological Society*, 143, 706–719.

Hersbach, H., Peubey, C., Simmons, A., Berrisford, P., Poli, P. and Dee, D. (2015) ERA-20CM: a twentieth-century atmospheric model ensemble. *Quarterly Journal of the Royal Meteorological Society*, 141, 2350–2375.

Hurrell, J.W., Kushnir, Y., Ottersen, G. and Visbeck, M. (2003). An overview of the North Atlantic Oscillation, In: Hurrell, J.W., Kushnir, Y., Ottersen, G., Visbeck, M. *The North Atlantic Oscillation: Climatic Significance and Environmental Impact*, pp. 1–36. Washington, DC: American Geophysical Union.

Kennedy, J., Rayner, N., Smith, R., Parker, D. and Saunby, M. (2011) Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *Journal of Geophysical Research: Atmospheres*, 116, D14104

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

MacLeod, D., O'Reilly, C., Palmer, T. and Weisheimer, A. (2018) Flow dependent ensemble spread in seasonal forecasts of the boreal winter extratropics. *Atmospheric Science Letters*, 19, e815

Minobe, S. and Mantua, N. (1999) Interdecadal modulation of interannual atmospheric and oceanic variability over the North Pacific. *Progress in Oceanography*, 43, 163–192.

Nie, Y., Scaife, A.A., Ren, H.-L., Comer, R.E., Andrews, M.B., Davis, P. and Martin, N. (2019) Stratospheric initial conditions provide seasonal predictability of the North Atlantic and Arctic Oscillations. *Environmental Research Letters*, 14, 034006

O'Reilly, C.H. (2018) Interdecadal variability of the ENSO teleconnection to the wintertime North Pacific. *Climate Dynamics*, 51, 3333–3350.

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.

O'Reilly, C.H., Weisheimer, A., Woollings, T., Gray, L.J. and MacLeod, D. (2019) 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*, 145, 131–146.

Parker, T., Woollings, T., Weisheimer, A., O'Reilly, C., Baker, L. and Shaffrey, L. (2019) Seasonal predictability of the winter North Atlantic Oscillation from a jet stream perspective. *Geophysical Research Letters*, 46, 10159–10167.

Poli, P., Hersbach, H., Dee, D.P., Berrisford, P., Simmons, A.J., Vitart, F., Laloyaux, P., Tan, D.G., Peubey, C., Thépaut, J.-N., Trémolet, Y., Hólm, E.V., Bonavita, M., Isasken, L. and Risher, M. (2016) ERA-20C: An atmospheric reanalysis of the twentieth century. *Journal of Climate*, 29, 4083–4097.

Rayner, N., Parker, D.E., Horton, E., Folland, C.K., Alexander, L.V., Rowell, D., Kent, E. and Kaplan, A. (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres*, 108, 4407

Riddle, E.E., Butler, A.H., Furtado, J.C., Cohen, J.L. and Kumar, A. (2013) CFSv2 ensemble prediction of the wintertime Arctic Oscillation. *Climate Dynamics*, 41, 1099–1116.

Ropelewski, C.F. and Jones, P.D. (1987) An extension of the Tahiti–Darwin Southern Oscillation Index. *Monthly Weather Review*, 115, 2161–2165.

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) Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, 41, 2514–2519.

Scaife, A.A. and Smith, D. (2018) A signal-to-noise paradox in climate science. *npj Climate and Atmospheric Science*, 1, 28

Smith, D.M., Scaife, A.A. and Kirtman, B.P. (2012) What is the current state of scientific knowledge with regard to seasonal and decadal forecasting?. *Environmental Research Letters*, 7, 015602

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.

Thompson, D.W., Kennedy, J.J., Wallace, J.M. and Jones, P.D. (2008) A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. *Nature*, 453, 646–649.

Thompson, D.W. and Wallace, J.M. (1998) The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical Research Letters*, 25, 1297–1300.

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., Colin, J., Guérémy, J.-F., Michou, M., Moine, M.-P., Nabat, P., Roehring, R., Salas y Mélia, D., Séférian, R., Valcke, S., Beau, I., Belamari, S., Berthet, S., Cassou, C., Cattiaux, J., Deshayes, J., Douville, H., Ethé, C., Franchistéguy, L., Geoffroy, O., Lévy, C., Madec, G., Meurdesoif, Y., Msadek, R., Ribes, A., Sanchez-Gomez, E., Terray, L. and Waldman, R. (2019) Evaluation of CMIP6 DECK experiments with CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, 11, 2177–2213.

Wallace, J.M. and Gutzler, D.S. (1981) Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Monthly Weather Review*, 109, 784–812.

Weisheimer, A., Befort, D.J., MacLeod, D., Palmer, T., O'Reilly, C. and Strömmen, K. (2020) Seasonal forecasts of the 20th century. *Bulletin of the American Meteorological Society*. https://doi.org/10.1175/BAMS-D-19-0019.1

Weisheimer, A., Decremer, D., MacLeod, D., O'Reilly, C., Stockdale, T.N., Johnson, S. and Palmer, T.N. (2018) How confident are predictability estimates of the winter North Atlantic Oscillation?. *Quarterly Journal of the Royal Meteorological Society*, 145, 145–159.

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

Zhou, T., Turner, A.G., Kinter, J.L., Wang, B., Qian, Y., Chen, X., Wu, B., Liu, B., Zou, L. and Bian, H. (2016) GMMIP (v1.0) contribution to CMIP6: Global Monsoons Model Inter-comparison Project. *Geoscientific Model Development*, 9, 3589–3604.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: O'Reilly CH, Weisheimer A, MacLeod D, Befort DJ, Palmer T. Assessing the robustness of multidecadal variability in Northern Hemisphere wintertime seasonal forecast skill. *Q J R Meteorol Soc*. 2020;146:4055–4066. https://doi.org/10.1002/qj.3890