



The importance of stratospheric initial conditions for winter North Atlantic Oscillation predictability and implications for the signal-to-noise paradox

Christopher H. O'Reilly¹  | Antje Weisheimer^{2,3}  | Tim Woollings¹ | Lesley J. Gray² | Dave MacLeod¹

¹Atmospheric, Oceanic and Planetary Physics, University of Oxford, UK

²National Centre for Atmospheric Science, Department of Physics, University of Oxford, UK

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

Correspondence

Christopher H. O'Reilly, Atmospheric, Oceanic and Planetary Physics, Department of Physics, Parks Road, Oxford OX1 3PU, UK.

Email: Christopher.O'Reilly@physics.ox.ac.uk

Funding information

NERC grant NE/M005887/1, the NCAS and the NERC Atlantic Climate System Integrated Study,

This study investigates the influence of atmospheric initial conditions on winter seasonal forecasts of the North Atlantic Oscillation (NAO). Hindcast (or reforecast) experiments – which differ only in their initial conditions – are performed over the period 1960–2009, using prescribed sea surface temperature (SST) and sea-ice boundary conditions. The first experiment (“ERA-40/Int IC”) is initialized using the ERA-40 and ERA-Interim reanalysis datasets, which assimilate upper-air, satellite and surface observations; the second experiment (“ERA-20C IC”) is initialized using the ERA-20C reanalysis dataset, which assimilates only surface observations. The ensemble mean NAO skill is largest in ERA-40/Int IC ($r = 0.54$), which is initialized with the superior reanalysis data. Moreover, ERA-20C IC did not exhibit significantly more NAO hindcast skill ($r = 0.38$) than in a third experiment, which was initialized with incorrect (shuffled) initial conditions. The ERA-40/Interim and ERA-20C initial conditions differ substantially in the tropical stratosphere, where the quasi-biennial oscillation (QBO) of zonal winds is not present in ERA-20C. The QBO hindcasts are highly skilful in ERA-40/Int IC – albeit with a somewhat weaker equatorial zonal wind amplitude in the lower stratosphere – but are incorrect in ERA-20C IC, indicating that the QBO is responsible for the additional NAO hindcast skill; this is despite the model exhibiting a relatively weak teleconnection between the QBO and NAO. The influence of the QBO is further demonstrated by regressing out the QBO influence from each of the hindcast experiments, after which the difference in NAO hindcast skill between the experiments is negligible. Whilst ERA-40/Int IC demonstrates a more skilful NAO hindcast, it appears to have a relatively weak predictable signal; this is the so-called “signal-to-noise paradox” identified in previous studies. Diagnostically amplifying the (weak) QBO–NAO teleconnection increases the ensemble-mean NAO signal with negligible impact on the NAO hindcast skill, after which the signal-to-noise problem seemingly disappears.

KEYWORDS

North Atlantic Oscillation (NAO), polar vortex, initial conditions, Quasi-Biennial Oscillation (QBO), seasonal predictability, signal-to-noise paradox

1 | INTRODUCTION

Wintertime climate anomalies over Europe and Asia are strongly dependent on the North Atlantic Oscillation (NAO),

which is the dominant mode of large-scale atmospheric circulation variability over the Euro-Atlantic sector (e.g. Hurrell *et al.*, 2001). Historically, the NAO has proven difficult to predict on seasonal time-scales (Palmer *et al.*, 2004; Müller

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.

et al., 2005). However, recent developments in seasonal forecasting systems have shown that the wintertime NAO may be more predictable than previously thought, with winter forecasts (for December–January–February, initialized at the beginning of November) exhibiting appreciable skill. Scaife *et al.* (2014a) showed an ensemble mean correlation skill for the NAO of $r \approx 0.6$ over a 20-year hindcast period on the Met Office's GloSEA5 seasonal forecasting system. Similar levels of NAO hindcast skill were also found in a longer 35-year hindcast using the Met Office DePreSys3 model by Dunstone *et al.* (2016). Other seasonal forecasting systems have demonstrated wintertime seasonal forecast skill in hindcasts of the Arctic Oscillation (AO), the dominant mode of extratropical variability over the Northern Hemisphere (Riddle *et al.*, 2013; Stockdale *et al.*, 2015).

A notable feature of the recent seasonal forecast systems which have exhibited increased NAO hindcast skill, is that the ensemble mean signal in the NAO (or AO) seems to be somewhat weaker than one might expect for the level of demonstrated correlation skill. This was shown by Eade *et al.* (2014), who examined the “ratio-of-predictable-components” in the Met Office seasonal and decadal forecasting systems (recent review by Scaife and Smith, 2018). The ensemble mean signal was also found to be too weak in the hindcasts of the wintertime AO in the study by Stockdale *et al.* (2015) using the ECMWF System-4. As a result of the weak signal in the seasonal forecasting systems, large numbers of ensemble members are required to achieve ensemble mean hindcast skill. Another practical limitation of the weak signal is that the ensemble spread exhibits little variability between more or less accurate hindcasts. This means that forecasts made by the models with a signal-to-noise problem do not provide useful ensemble spreads that can be interpreted as estimates of their own forecast error, unlike in medium-range ensemble forecasts (e.g. Molteni *et al.*, 1996). MacLeod *et al.* (2018) demonstrated that seasonal forecast models can demonstrate good relationships between forecast spread and error in the Extratropics. The hindcast analysed in the study by MacLeod *et al.* (2018) also does not demonstrate any signal-to-noise problem (A. Weisheimer personal communication, 2018), however this hindcast is also appreciably less skilful than those in which a signal-to-noise problem has been found (e.g. Scaife *et al.*, 2014a; Stockdale *et al.*, 2015; Dunstone *et al.*, 2016).

Wintertime NAO predictability has been attributed to numerous sources. Statistical models demonstrate that both extratropical and tropical sea-surface temperature (SST), sea-ice anomalies and anomalous circulation in the stratosphere provide predictable precursors for the wintertime NAO (Folland *et al.*, 2012; Wang *et al.*, 2017). The skilful prediction of these factors was cited as the source of the NAO hindcast skill in the study by Dunstone *et al.* (2016). Tropical SSTs, most notably associated with the El Niño/Southern Oscillation (ENSO), can generate predictable tropical precipitation (and associated atmospheric heating) anomalies, which

trigger stationary Rossby waves in the Extratropics and are the dominant source of seasonal predictability in the Extratropics (Smith *et al.*, 2012; Scaife *et al.*, 2017). Extratropical SST anomalies in the North Atlantic have also been linked to predictability in the wintertime NAO (e.g. Rodwell *et al.*, 1999; Czaja and Frankignoul, 2002; Dunstone *et al.*, 2016), as have autumn sea-ice anomalies – particularly in the Barents and Kara Seas (e.g. Alexander *et al.*, 2004; Nakamura *et al.*, 2015; Yang *et al.*, 2016).

In addition, the influence of the stratosphere on seasonal circulation anomalies has been demonstrated in many studies. For example, precursors at the beginning of the winter have been linked to seasonal anomalies in the polar vortex and the coupling to the lower-tropospheric circulation anomalies (e.g. Baldwin *et al.*, 2001; Thompson *et al.*, 2002; Kidston *et al.*, 2015). An example of the stratospheric influence on extratropical tropospheric circulation is the quasi-biennial oscillation (QBO) of zonal winds in the tropical stratosphere, which has been shown to influence the wintertime NAO in observational and modelling studies (e.g. Holton and Tan, 1980; 1982; Anstey and Shepherd, 2014). Boer and Hamilton (2008) argued that the QBO teleconnection to the Extratropics may act as an important source of skill in seasonal forecasts, using a statistical approach. In a more recent study, Scaife *et al.* (2014b) showed that the long time-scales of the QBO are well captured in initialized seasonal forecast models at 1-month lead times, although the link to the winter NAO in seasonal forecasts is either relatively weak or absent in these models. Over a short eight-year hindcast period, Stockdale *et al.* (2015) shifted the oceanic initial conditions by 1 year and demonstrated that the skilful forecasts of the winter AO in the control hindcast were strongly dependent on atmospheric initial conditions, most likely through the influence of the stratosphere. In a recent paper, Hansen *et al.* (2017) demonstrated that the downward influence from the stratosphere is potentially important for seasonal forecasting skill by imposing stratospheric relaxation in seasonal hindcast experiments, which dramatically increased the hindcast NAO skill. However, whilst it seems that the stratosphere may play an important role in seasonal forecasts, the relative importance of stratospheric initial conditions on winter NAO predictability is not clear.

In this study, we investigate the influence of atmospheric initial conditions – and in particular those in the stratosphere – by analysing a series of seasonal hindcast experiments. Specifically, hindcast experiments initialized with two different reanalysis products are analysed; these products are similar in the troposphere but differ substantially in the stratosphere. These experiments are described in more detail in the next section. In section 3, we analyse the skill of the NAO and the role of the stratospheric initial conditions in determining the hindcast skill. In particular, the QBO teleconnection to the NAO stands out as an important source of skill in the hindcast experiments. However, the QBO–NAO teleconnection appears to be somewhat weaker than in the

observations, which results in a signal-to-noise issue. In section 4 we summarize our results and discuss the potential implications for the signal-to-noise paradox in operational seasonal forecasting systems.

2 | DATASETS, EXPERIMENTS AND METHODS

2.1 | Observational datasets

In this study we use several reanalysis datasets. We use ERA-Interim reanalysis (Dee *et al.*, 2011) over the period 1979–2010. ERA-Interim assimilates a vast number of surface and upper-air observations, including from satellite products. Since ERA-Interim is only available from 1979, we use data from the ERA-40 reanalysis (Uppala *et al.*, 2005) over the period 1960–1978, which allows us to extend our hindcast period further back in the twentieth century and increase the sample size. ERA-40 is an older version of the ECMWF reanalysis using an earlier model version, but still assimilates a large number of surface and upper-air observations. However, over the period 1960–1978 in the ERA-40 dataset there are fewer observations from satellites. We combine data from the ERA-Interim (from January 1979 onwards) and ERA-40 (prior to 1979) reanalyses to produce a dataset, hereafter referred to as ERA-40/Interim, which spans the period 1960–2010. In addition, we also use data from the ERA-20C reanalysis (Poli *et al.*, 2016). ERA-20C was produced by assimilating only surface pressure and wind observations, over the period 1900–2010. The most significant difference between ERA-20C and the more comprehensive ERA-40/Interim dataset is that ERA-20C includes no upper-air or satellite observations. We will here focus on the 1960–2009 period which is covered by both ERA-20C and ERA-40/Interim. In addition to the reanalysis data, we will also use SSTs and sea-ice concentrations from the HadISST2.1.0.0 dataset (Rayner *et al.*, 2003) to provide the surface boundary conditions for the hindcast experiments.

2.2 | Hindcast experiments

We analyse the results from three ensemble hindcast experiments. All of these experiments were performed using the ECMWF Integrated Forecasting System (cycle 41r1), which is an older version of the model than is used for the current operational ECMWF seasonal forecasting system, SEAS5. The experiments were performed at T255 horizontal resolution (≈ 80 km), with 91 levels in the vertical up to 0.01 hPa in the top level. The experiments were initialized using reanalysis data (interpolated to the model grid) on every 01 November (at 0000 UTC) between 1960 and 2009 and analysed for the DJF winter season. The experiments were performed with explicit stochastic physics, which represents unresolved subgrid-scale atmospheric processes and creates

the spread between the ensemble members (e.g. Weisheimer *et al.*, 2014). Prescribed SST and sea-ice surface boundary conditions are used in the experiments and these are taken from the HadISST dataset. As a result of this idealized approach, these hindcast experiments should be considered as an indication of potential predictability in which SSTs are perfectly predicted. An advantage of this idealized approach here is that we are able to isolate the roles of SST boundary conditions and atmospheric initial conditions, as we outline below. Details of the hindcast experiments are also shown in Table 1. The first of the hindcast experiments takes (atmospheric) initial conditions from ERA-40/Interim with observed SST and sea-ice boundary conditions. This experiment is hereafter referred to as “ERA-40/Int IC”, where IC indicates the initial conditions used in the ensemble. The second experiment is identical to the first experiment, except that initial conditions are taken from the ERA-20C dataset; this experiment is hereafter referred to as “ERA-20C IC”. The ERA-20C IC experiment is the hindcast that was performed by Weisheimer *et al.* (2017) over the entire twentieth century. They found that the skill of the NAO (as well as the Pacific/North American pattern) exhibited a minimum in the mid-twentieth century, when there was relatively weak forcing from tropical Pacific SSTs (O'Reilly *et al.*, 2017; O'Reilly, 2018). The hindcast period analysed here primarily covers the later period where the ERA-20C IC experiment exhibits higher levels of skill. Both ERA-40/Int IC and ERA-20C IC were performed with 51 ensemble members for each winter season over the 50-year hindcast period. A comparison of these two hindcast experiments will highlight sensitivity to atmospheric initial conditions.

The third experiment uses initial conditions from ERA-40/Interim (as in the ERA-40/Int IC experiment), but for each initial condition year the prescribed SST and sea-ice boundary conditions are taken from each of the other 49 years over the period (i.e. 49 ensemble members per start date); this simulation is hereafter referred to as the “Shuffled” experiment. A comparison of this experiment with the ERA-40/Int IC experiment will highlight sensitivities to surface boundary conditions. This hindcast can be viewed in two ways (see schematic in Figure 1). Firstly, the ensemble members in the Shuffled experiment can be averaged over all 49 surface boundary conditions for a given initial condition year, hereafter the “Correct IC-only” ensemble; secondly, the ensemble members can be averaged over all the different atmospheric initial conditions for the SST boundary conditions for a given year, hereafter the “Correct SST-only” ensemble, as shown in the schematic (Figure 1). Therefore, the Shuffled experiment allows us to examine the relative importance of the atmospheric initial conditions and the SST boundary conditions by computing the Correct IC-only and Correct SST-only ensemble means. These ensemble means from the Shuffled experiment will also be compared with the ERA-40/Int IC experiment, which contains the correct initial conditions and SST boundary conditions, though the

TABLE 1 Information about the hindcast experiments analysed in this study. Values of the ensemble mean correlation skill for the polar vortex and NAO indices are also shown. Values that are significant at the 5% level are shown in bold (based on a two-sided *t*-test)

Experiment name	Surface forcing	Atmospheric initial conditions	Ensemble size/length	Polar vortex skill	NAO hindcast skill	
					Full hindcast	QBO regressed out
ERA-40/Int IC	Correct SSTs/sea-ice	Accurate atmospheric initial conditions (from ERA-40/Interim reanalysis)	51 members, 50 years	0.41	0.54	0.40
ERA-20C IC	Correct SSTs sea-ice	Inaccurate atmospheric initial conditions (from ERA-20C reanalysis)	51 members, 50 years	0.27	0.38	0.36
Shuffled:						
(i) Correct SST-only	Correct SST/sea-ice	Shuffled atmospheric initial conditions (from ERA-40/Interim reanalysis)	49 members, 50 years	0.19	0.36	0.36
(ii) Correct IC-only	Shuffled SST/sea-ice	Correct atmospheric initial conditions (from ERA-40/Interim reanalysis)	49 members, 50 years	0.28	0.16	0.10

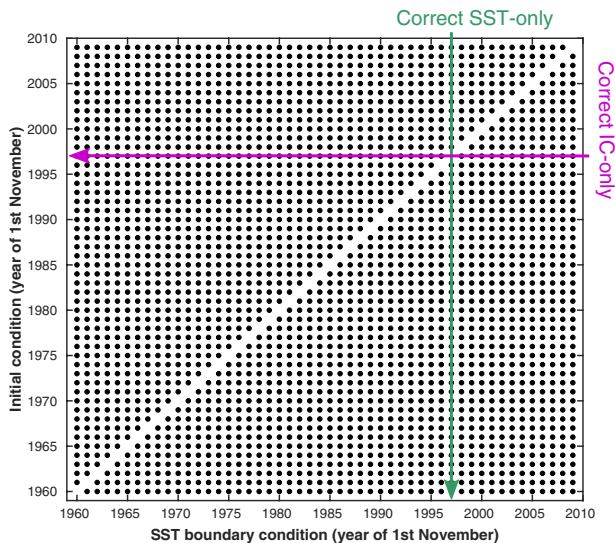


FIGURE 1 The points in the schematic indicate individual ensemble integrations that were performed as part of the Shuffled hindcast experiment. The Correct SST-only and Correct IC-only statistics are calculated by averaging across all ensemble members with the same prescribed SST boundary condition or same initial condition [Colour figure can be viewed at wileyonlinelibrary.com]

Shuffled experiment has greater initial condition variance or SST variance for each start date.

2.3 | NAO indices

In this study we analyse the predictability of the NAO by comparing the NAO from the hindcast experiments, with NAO indices from observations over the 50-year period (1960–2009). We define the observed NAO (e.g. Hurrell *et al.*, 2003) as the (normalized) principle component time series of the first empirical orthogonal function (EOF) of the wintertime (DJF) 500 hPa geopotential height (Z500) anomalies over the Atlantic sector (90°W–30°E, 30–90°N), in the ERA-40/Interim dataset. The NAO indices from each hindcast ensemble member are calculated by projecting the modelled Z500 anomalies onto the reference NAO pattern calculated from the ERA-40/Interim dataset (note that the

NAO index calculated from the ERA-20C reanalysis dataset is almost identical to that from the ERA-40/Interim dataset, with a correlation of $r = 0.99$).

2.4 | QBO indices

To assess the ability of the hindcast experiments to skilfully reproduce the seasonal evolution of the QBO, we analyse indices of the zonal-mean zonal winds – averaged along the Equator – at 30 hPa (following e.g. Scaife *et al.*, 2014b). The 30 hPa level was selected because, when compared with other levels (e.g. 10, 20, 50 or 70 hPa), it exhibits the largest correlation with wintertime (DJF) NAO index in both the reanalysis and the ERA-40/Int IC experiment, which is consistent with other studies of the QBO teleconnection to the northern extratropical tropospheric circulation (e.g. Gray *et al.*, 2018). To produce winter composites based on the different QBO phases, we define westerly QBO (QBO-W) as winters in which the QBO index from reanalysis (defined at 30 hPa) exceeds 2.5 m/s and define easterly QBO (QBO-E) as winters in which the QBO index is less than –5 m/s.

2.5 | Monte Carlo significance testing

To test the significance of the difference in hindcast NAO correlation skill between any two of the experiments (e.g. the ERA-40/Int IC and ERA-20C IC experiments in Figure 3a), we performed a Monte Carlo test by combining the ensemble members from the two experiments being compared and then randomly splitting them into two 51-member samples. These two random ensembles were then averaged to give two ensemble mean time series, which were correlated with the observed NAO time series. The difference between the correlation values was saved and the process was repeated to generate 10,000 random correlation differences. The observed difference in hindcast NAO correlation skill was then compared with the two-tailed random distribution to give a *p*-value which reflects the probability that a correlation difference of that magnitude could have occurred through random sampling.

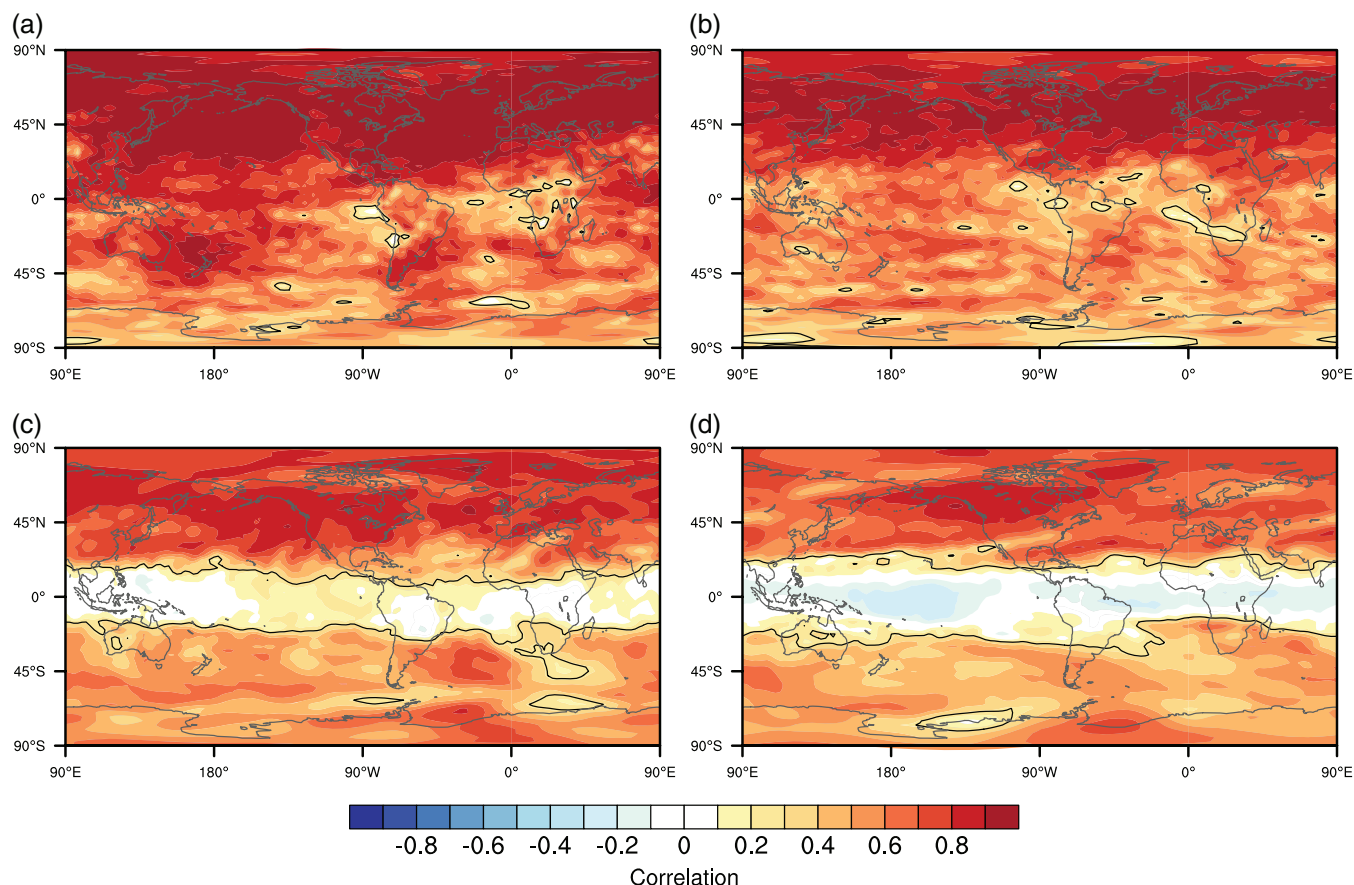


FIGURE 2 Correlation between the zonal wind fields in the 01 November initial conditions (i.e. $n = 50$ years) taken from ERA-40/Interim and ERA-20C (in the ERA-40/Int IC and ERA-20C IC hindcast experiments, respectively) at (a) 850 hPa, (b) 300 hPa, (c) 50 hPa and (d) 10 hPa. The black contours show regions where the correlation is not significant at the 5% level (based on a two-sided t -test) [Colour figure can be viewed at wileyonlinelibrary.com]

3 | RESULTS

We will now assess the influence of the stratospheric initial conditions, and in particular the QBO, on the hindcast skill of the NAO. We do so by analysing the representation of the QBO in the different hindcast experiments and the teleconnection of the QBO to the NAO in the Extratropics. We begin, though, by comparing the the initial conditions used in the hindcast experiments.

3.1 | Initial condition comparison

To demonstrate the initial condition differences between the ERA-40/Int IC and ERA-20C IC experiments, maps of the grid-point correlations between the initial zonal winds in each year (i.e. 0000 UTC 01 November snapshots) at various levels are shown in Figure 2. In the Northern Hemisphere, particularly in the Extratropics, there is a good agreement between the ERA-20C and ERA-40/Interim initial conditions in the lower and upper troposphere (i.e. Figure 2a,b). In the northern Extratropics, there are also significant correlations between the zonal wind anomalies in the stratosphere (i.e. Figure 2c,d). This is likely due to the fact that in early winter the strength of the polar vortex is largely determined by radiative processes that are well modelled and also the upward propagation

of planetary waves – and their subsequent influence on the stratospheric circulation (e.g. Andrews and McIntyre, 1976) – being reasonably well constrained through the assimilation of surface observations. In the Tropics and the Southern Hemisphere, the zonal winds are less well correlated. This is likely due to the absence of upper-air observations in the Tropics in ERA-20C, and the relative paucity of surface pressure and wind observations in the Southern Hemisphere compared to the Northern Hemisphere (Poli *et al.*, 2016); in addition, November is a month of large dynamical variability of the Southern Hemisphere stratospheric polar vortex, when it undergoes its final warming. However, the most striking differences are in the tropical stratosphere, where the zonal wind initial conditions are essentially uncorrelated (or even slightly negatively correlated at 10 hPa; Figure 2d). Therefore, one of the main differences between the ERA-40/Interim and ERA-20C initial conditions is the representation of the correct phase (and amplitude) of the equatorial zonal winds associated with the QBO in the tropical stratosphere, which we will discuss in more detail shortly.

3.2 | NAO skill in the hindcast experiments

The time series of the ensemble mean NAO hindcasts in the ERA-40/Int IC and ERA-20C IC experiments are shown

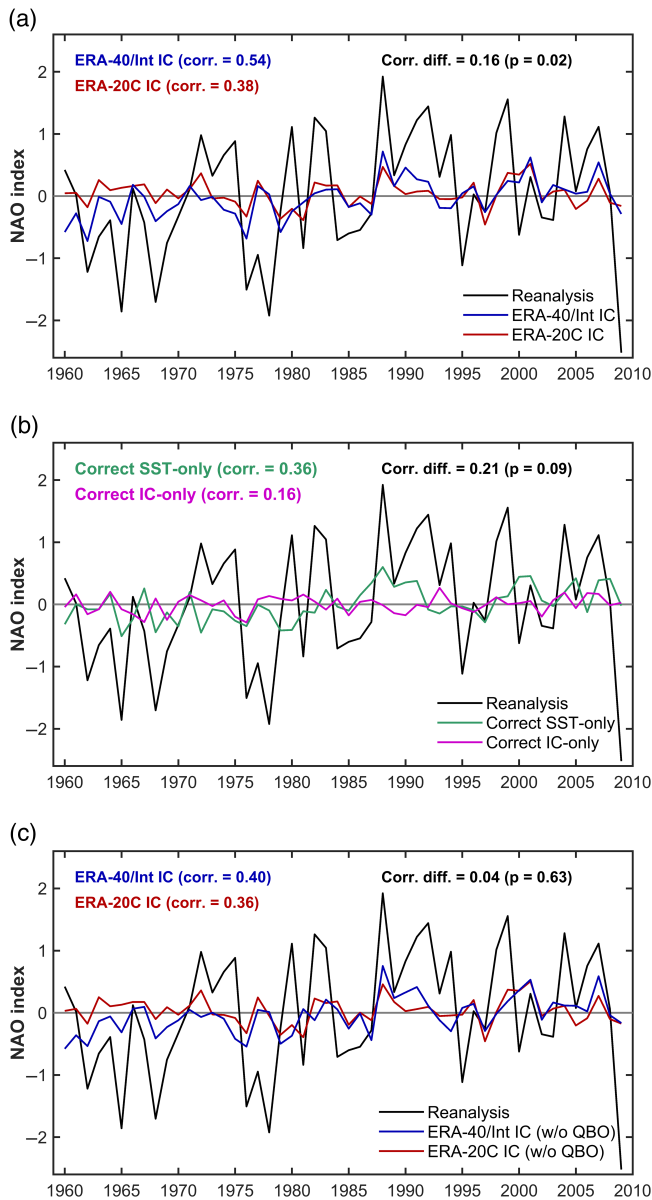


FIGURE 3 (a) The wintertime (DJF) NAO index from the ERA-40/Interim reanalysis data plotted against the ensemble mean NAO indices for the ERA-40/Int IC and ERA-20C IC hindcast experiments. (b) The wintertime (DJF) NAO index from reanalysis data plotted against the ensemble mean NAO indices for the Correct SST-only IC and Correct IC-only ensembles from the Shuffled hindcast experiment. (c) is as (a), but the contribution to the NAO from the QBO teleconnection in each ensemble member has been removed prior to recalculating the ensemble mean indices. Also shown are the correlations between the the ensemble mean NAO indices and the reanalysis NAO index [Colour figure can be viewed at wileyonlinelibrary.com]

in Figure 3a. Also shown are the correlation skill scores for the hindcast experiments compared to the observed NAO index, both of which are statistically significant (also Table 1). The correlation skill in the ERA-40/Int IC hindcast, $r = 0.54$ ($p < 0.001$), is notably higher than in the ERA-20C IC hindcast, $r = 0.38$ ($p < 0.001$). The Monte Carlo significance test indicates the difference in correlation skill between the two experiments, $\Delta r = 0.16$, is highly unlikely to have occurred by chance ($p = 0.02$). Recent studies using the ERA-20C IC experiment have highlighted that the hindcast

NAO skill in the recent period is more predictable than in the mid-twentieth century (Weisheimer *et al.*, 2017; O'Reilly *et al.*, 2017), therefore, it is interesting to examine this in the ERA-40/Int IC experiment as well. To test the stationarity of the hindcast NAO skill, we calculated the ensemble mean correlation skill over the first and second halves of the hindcast period. In the first 25 years of the hindcast period, 1960–1984, the hindcast NAO skill in ERA-40/Int IC is $r = 0.34$ and in ERA-20C IC is $r = 0.29$. In the second 25 years of the hindcast period, 1985–2009, the hindcast NAO skill in ERA-40/Int IC is $r = 0.60$ and in ERA-20C IC is $r = 0.47$. Therefore the ERA-40/Int IC experiment also exhibits more hindcast NAO skill in the latter part of the hindcast period, in agreement with the ERA-20C IC experiment.

Given that the ERA-40/Int IC and ERA-20C IC experiments differ only in initial condition, the difference in correlation skill between the two experiments indicates that the superior initial conditions from ERA-40/Interim are providing an appreciable amount of additional NAO correlation skill in the hindcasts. Also shown, in Figure 3b, are the Correct SST-only and Correct IC-only ensemble mean NAO indices from the Shuffled experiment. The correlation skill score of the Correct SST-only ensemble mean NAO is $r = 0.36$. However, the correlation skill score of the Correct IC-only ensemble mean NAO is only $r = 0.16$ and the difference in correlation skill compared to the the Correct SST-only ensemble, $\Delta r = 0.21$, is statistically significant at the 10% level ($p = 0.09$) according to the Monte Carlo significance test. Interestingly, the Correct SST-only ensemble exhibits almost the same hindcast NAO correlation skill as the ERA-20C IC experiment, which suggests that the surface boundary conditions alone are responsible for the skill in each of these ensembles. In other words, the atmospheric initial conditions from the ERA-20C hindcast (i.e. from the ERA-20C reanalysis dataset) do not contribute to the hindcast NAO skill. Furthermore, the similarity in NAO hindcast skill in ERA-20C IC and Correct SST-only implies that the major sources of atmospheric initial condition skill may originate from the region where the initial conditions in the ERA-40/Int IC experiment differ from those in the ERA-20C IC experiment. Analysis of the zonal wind initial conditions (i.e. Figure 2) demonstrates that the tropical stratosphere is the region of most substantial difference between the two sets of initial conditions, which is explored further in the next section.

3.3 | Representation of the QBO in the hindcast experiments

The dominant mode of interannual variability in the tropical stratosphere is the quasi-biennial oscillation (QBO). The QBO is characterized by alternating, descending bands of easterly and westerly equatorial zonal winds, as shown in the ERA-40/Interim reanalysis dataset in Figure 4a. The QBO is well represented in the ERA-40/Interim reanalysis datasets (e.g. Pascoe *et al.*, 2005) due to the assimilation

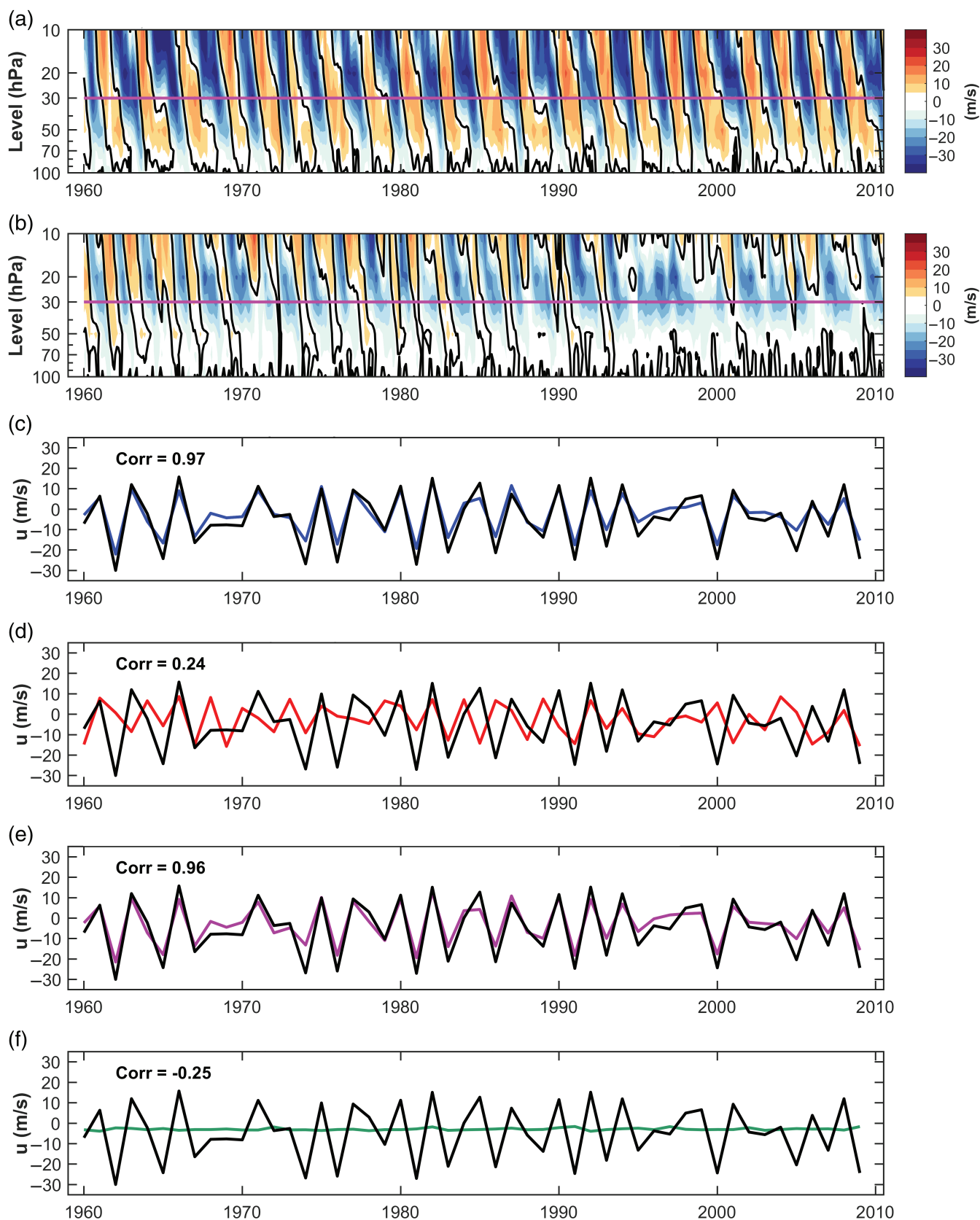


FIGURE 4 Hovmöller plots of the equatorial zonal-mean zonal winds in the stratosphere in (a) ERA-40/Interim reanalysis and (b) ERA-20C reanalysis. The ensemble mean wintertime (DJF) QBO indices, defined as equatorial zonal-mean zonal wind at 30 hPa, are shown by the coloured lines for (c) ERA-40/Int IC, (d) ERA-20C IC, (e) Correct IC-only and (f) Correct SST-only hindcast experiments. The black lines show the wintertime QBO index from the ERA-40/Interim reanalysis. The correlation between the ensemble mean and reanalysis QBO indices are also shown [Colour figure can be viewed at wileyonlinelibrary.com]

of regular radiosonde observations in the reanalysis. Interestingly, the ERA-20C exhibits QBO-like variability in the mid-upper stratosphere but this does not extend deep enough into the equatorial lower stratosphere (also Hersbach *et al.*, 2015; Fujiwara *et al.*, 2017). The model version used for the ERA-20C reanalysis does produce an internal QBO in free-running simulations, albeit with a period which is too short (2 years) and an amplitude which is too weak (Hersbach *et al.*, 2015), therefore it is not surprising that the ERA-20C reanalysis exhibits QBO-like variability. However, since the ERA-20C dataset does not assimilate upper-atmospheric observations, the stratospheric equatorial zonal wind variability is much reduced (Figure 4b) and does not exhibit the correct periodic alternating, descending zonal winds that are characteristic of the QBO. The winter mean (DJF) QBO indices from the hindcast experiments along with the ERA-40/Interim reanalysis are shown in Figure 4c–f.

In the ERA-40-Int IC experiment, the QBO initial condition from the reanalysis ensures a high correlation skill score for the hindcast QBO index compared with the QBO index from the ERA-40/Interim reanalysis (Figure 4c). This high level of seasonal QBO hindcast skill ($r = 0.97$) has been demonstrated in operational seasonal forecasting systems (Scaife *et al.*, 2014b) and is perhaps not surprising given the 2–4 month lead time is substantially less than the ≈ 28 month period of the QBO (cf. Figure 4a). The radiative time-scale in the equatorial lower stratosphere is of the order of months-to-years, thus providing an atmospheric “memory” in this region. In the ERA-20C IC experiment (Figure 4d), the lack of accurate initialization in the tropical stratosphere from the ERA-20C reanalysis, perhaps unsurprisingly, results in a very low seasonal QBO hindcast correlation skill ($r = 0.24$). The QBO hindcast skill is as high in the Correct IC-only ensemble mean (Figure 4e; $r = 0.96$) as in the ERA-40/Int IC experiment, indicating that SSTs are not important in determining the QBO on seasonal time-scales. However, in the Correct SST-only ensemble, the QBO exhibits a wide spread across the ensemble (Figure 4f) and exhibits no significant skill ($r = -0.25$). Therefore, the zonal winds in the tropical stratosphere are reasonably well represented in the hindcast ensembles that are initialized with more accurate stratospheric initial conditions.

We now further assess the representation of the QBO in the hindcast experiments. To do so, we analyse composites of the zonal-mean zonal wind in the westerly (QBO-W) and easterly (QBO-E) phases of the QBO, shown in Figure 5. In both QBO-W and QBO-E, the zonal winds systematically weaken along the Equator as the hindcast progresses from December to February in the hindcast experiments (Figure 5c–f). This can be seen very clearly in the scatter plot (Figure 5g), with the relative amplitude of the equatorial zonal winds becoming systematically weaker than the reanalysis as the hindcast progresses through the season. In addition, the zero wind contours in Figure 5a,b demonstrate that the extent of the equatorial winds associated with the

QBO narrows slightly compared to the reanalysis; the narrowing of QBO winds in the equatorial lower stratosphere appears to be a feature of many models (Schenzinger *et al.*, 2017). Overall, this analysis indicates that, despite exhibiting significant QBO skill on seasonal time-scales, there are still some significant shortcomings in the representation of the QBO in the hindcast experiments.

3.4 | QBO teleconnection to the NAO in the hindcast experiments

To assess how the representation of the QBO in the hindcasts is influencing the circulation in the Extratropics, we have regressed the (DJF averaged) zonal-mean zonal wind anomalies onto the normalized QBO indices for the observations and the hindcast experiments (Supporting Information Figure S1). These regression maps were produced by regressing across all ensemble members, such that there are substantially more data points in the experiments than in the ERA-40/Interim reanalysis. In the reanalysis, the westerly phase of the QBO is associated with a strengthening of the stratospheric polar vortex, peaking at around 65°N (Figure S1a, as shown in previous studies (Holton and Tan, 1980; 1982; Baldwin *et al.*, 2001; Anstey and Shepherd, 2014; Gray *et al.*, 2018). The stratospheric polar vortex is strengthened in the ERA-40/Int IC experiment during the westerly phase of the QBO (Figure S1b), though the strengthening of the polar vortex is less than in the reanalysis (Figure S1d). The ensemble members in the Shuffled experiment (i.e. $n = 49 \times 50 = 2,450$ members) also exhibit a strengthening of the stratospheric polar vortex during the westerly phase of the QBO (Figure S1c), though this is weaker than in the ERA-40/Int IC experiment. The QBO is substantially weaker in the ERA-20C IC experiment and there is only a weak association with the stratospheric polar vortex (not shown).

In order to more quantitatively evaluate and compare the relationships between the QBO and the stratospheric polar vortex, we define a stratospheric polar vortex index as the zonal-mean 10 hPa zonal wind anomaly at 65°N . The standard deviation of the winter QBO index, the regression of the polar vortex anomaly onto the normalized QBO index and the regression of the NAO onto the normalized QBO index are shown in Figure 6 for the reanalysis and hindcast experiments. The regressions are displayed as probability density functions (pdfs) to provide a measure of sampling uncertainty; the pdfs were calculated using a bootstrap with replacement method, performed 10,000 times. A striking feature of Figure 6a is that the standard deviation of the seasonal mean QBO in the ERA-40/Int IC and Shuffled experiments is around 30% lower than in the reanalysis, in spite of the extremely high hindcast correlation skill (c.f. Figure 4). Since the QBO is initialized on 01 November in the hindcasts, the lower standard deviation shows that the amplitude of the QBO is damped over the four-month hindcast period, consistent with the results shown in Figure 5. The IFS model used for the hindcast

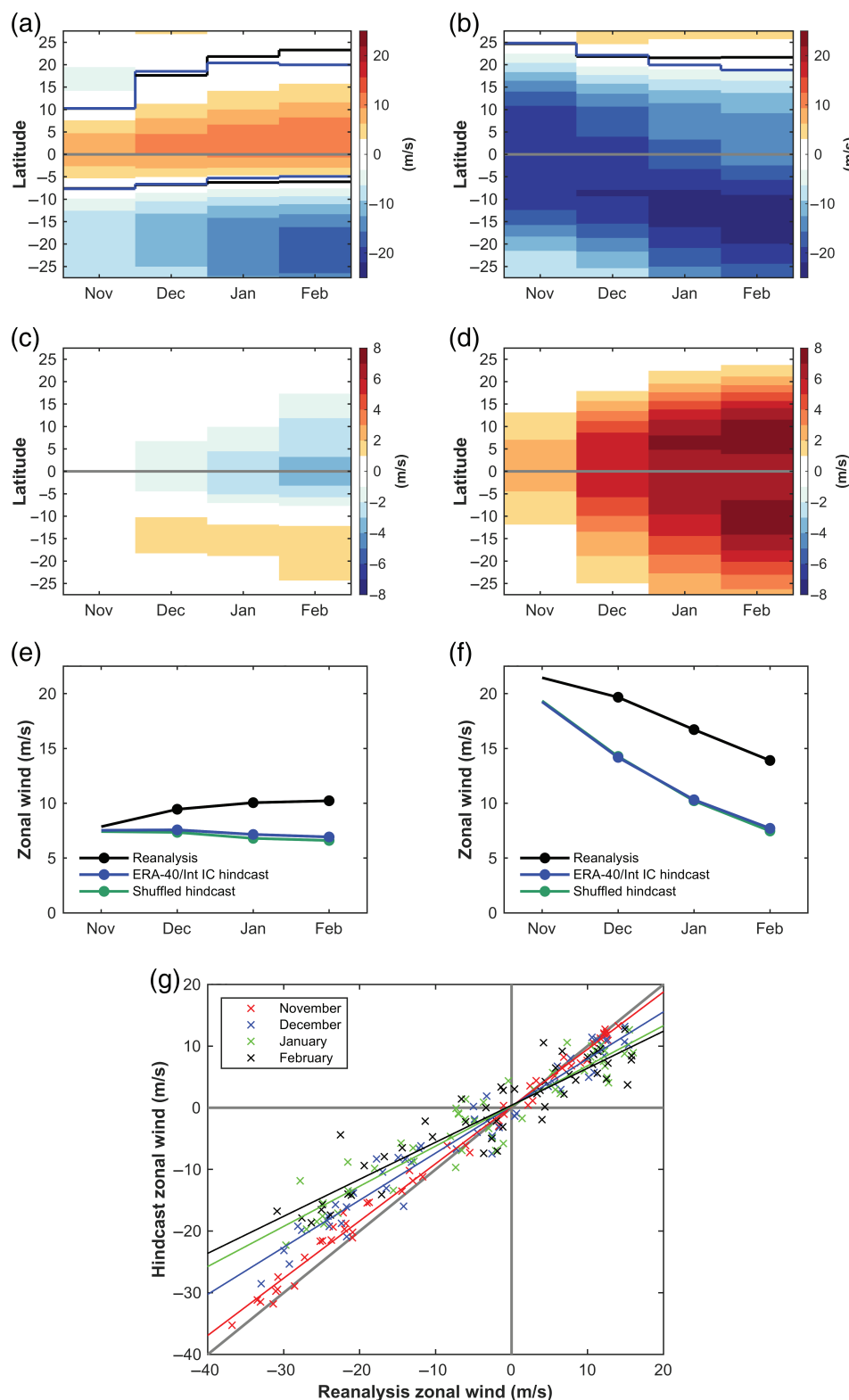


FIGURE 5 Hovmöller plots of zonal-mean zonal wind composites at 30 hPa for (a) westerly QBO (QBO-W) and (b) easterly QBO phases in the ERA-40/Interim reanalysis. QBO-W and QBO-E winters were selected based on the (DJF) QBO index at 30 hPa being greater than 2.5 m/s ($n = 19$ winters) or less than 5 m/s ($n = 24$ winters) respectively. The lines in (a) and (b) show where the zonal wind composite is zero for the reanalysis (black) and ERA-40/Int IC experiment (blue). (c) and (d) show the corresponding differences between the zonal wind composites from the ERA-40/Int IC experiment and the composites from the reanalysis (defined as ERA-40/Int IC minus reanalysis). (e) and (f) are composite plots of the equatorial zonal-mean zonal wind amplitude at 30 hPa for the QBO phases in the reanalysis, ERA-40/Int IC experiment and Shuffled experiment (note the data from the Shuffled experiment is very similar to the ERA-40/Int IC and is hard to distinguish). The filled circles in (e) and (f) show where the zonal wind composite is significantly different between the reanalysis and the experiments (at the 5% level, according to a two-sided t -test). (g) shows the scatter plots of monthly values of the equatorial zonal-mean zonal wind at 30 hPa in the reanalysis and the ERA-40/Int IC ensemble mean, with lines of best fit. The grey line in (g) indicates where the hindcast ensemble mean is equal to the reanalysis [Colour figure can be viewed at wileyonlinelibrary.com]

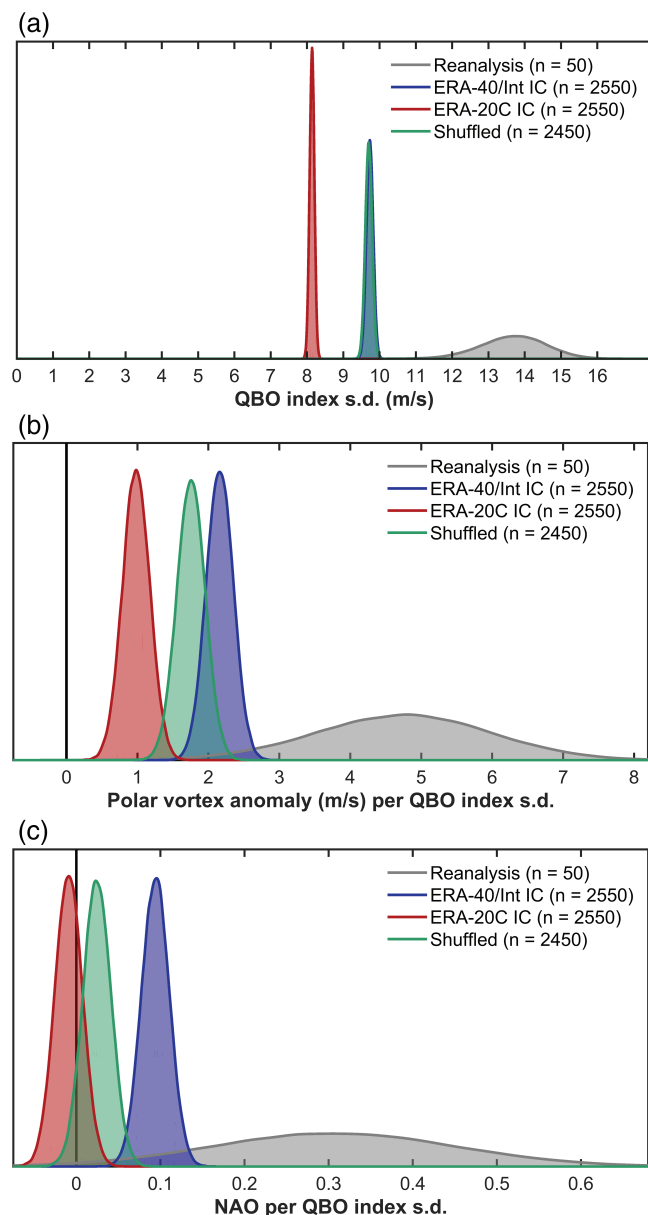


FIGURE 6 Probability density functions (pdfs) showing the sampling uncertainty in (a) the standard deviation of the winter QBO indices (as plotted in Figure 4), (b) the polar vortex anomaly regressed onto the normalized QBO index and (c) the NAO anomaly regressed onto the normalized QBO index. The distributions were calculated using a bootstrap with replacement method, performed 10,000 times to provide a measure of the sampling uncertainty, which is reflected in the width of the distributions. The sample sizes of the datasets are indicated in the legend. For presentation purposes, the pdfs were calculated from the 10,000 resamples using the kernel method of Silverman (1981). Note that in (a), because the pdfs for the ERA-40/Int IC and Shuffled experiments are almost identical, they are difficult to distinguish in the figure [Colour figure can be viewed at wileyonlinelibrary.com]

experiments is able to reproduce QBO-like behaviour in the tropical stratosphere in free-running simulations at comparable resolution (albeit in an earlier model cycle), though the amplitude of the QBO is much reduced (Orr *et al.*, 2010; Christiansen *et al.*, 2016). The damped QBO magnitude seen in the ERA-40/Int IC and Shuffled experiments is consistent with the inability of the non-initialized IFS to produce a QBO

of the magnitude seen in observations. The standard deviation of the 30 hPa QBO index in the ERA-20C IC experiment has an even smaller magnitude than in the ERA-40/Int IC and Shuffled experiments. (We note that, as seen in Figure 4, the ERA-20C reanalysis does not strictly have a QBO but nevertheless displays substantial variability at 30 hPa.)

The polar vortex anomaly associated with the QBO is weaker in all the hindcast experiments than in the reanalysis, though there is considerable sampling uncertainty in the estimate from reanalysis (Figure 6b). The polar vortex anomaly associated with the QBO in the ERA-40/Int IC experiment is larger than in the Shuffled experiment and both experiments exhibit a larger polar vortex anomaly than in the ERA-20C IC experiment. The NAO anomaly associated with the QBO in the ERA-40/Int IC is positive, as seen in the reanalysis (Figure 6c). However, the NAO anomaly associated with the QBO is approximately three times larger in the reanalysis than in the ERA-40/Int IC experiment although, again, there is substantial sampling uncertainty in the reanalysis relationship. The NAO anomaly associated with the QBO in the Shuffled experiment ensemble members is also positive but is much smaller than in the ERA-40/Int IC experiment. In the ERA-20C IC experiment, there is essentially no relationship between the NAO and the QBO. However, the differences in the relationship between the QBO and both the polar vortex and NAO in the ERA-40/Int IC and Shuffled experiments is more striking because these two experiments have the same initial conditions and identical QBO magnitudes (Figure 6a). In the Shuffled experiment, there is a wide spread in the SST boundary conditions associated with each initial condition (i.e. the Correct IC-only ensemble), whereas in the ERA-40-Int IC experiment each initial condition is associated with the same SST boundary condition. Therefore, there is likely to be more non-QBO related variability in the Shuffled IC experiment, which may contribute to the weaker relation between the QBO and both the polar vortex and NAO. We now turn to analyse why these relationships are weaker in the Shuffled IC experiment.

Many previous observational and modelling studies have demonstrated the important influence of the SSTs, particularly in the tropical Pacific, on the stratospheric polar vortex (e.g. Brönnimann *et al.*, 2004; Garcia-Herrera *et al.*, 2006; Manzini *et al.*, 2006; Free and Seidel, 2009). The influence of the tropical Pacific SST anomalies associated with the ENSO is understood to occur primarily through its teleconnection to the Extratropics or, more specifically, the Pacific/North America (PNA) pattern (Manzini *et al.*, 2006). Garfinkel and Hartmann (2008) showed that the positive phase of the PNA pattern is associated with an increase in the convergence of upward propagating planetary waves into the polar stratosphere, which acts to warm the polar stratosphere and weaken the polar vortex. In the Correct IC-only ensemble, for each QBO state that is very similar to the ERA-40/Int IC, there is a wide variety of tropical Pacific SST patterns. Figure 7 shows the DJF ensemble spread (defined as the average of the

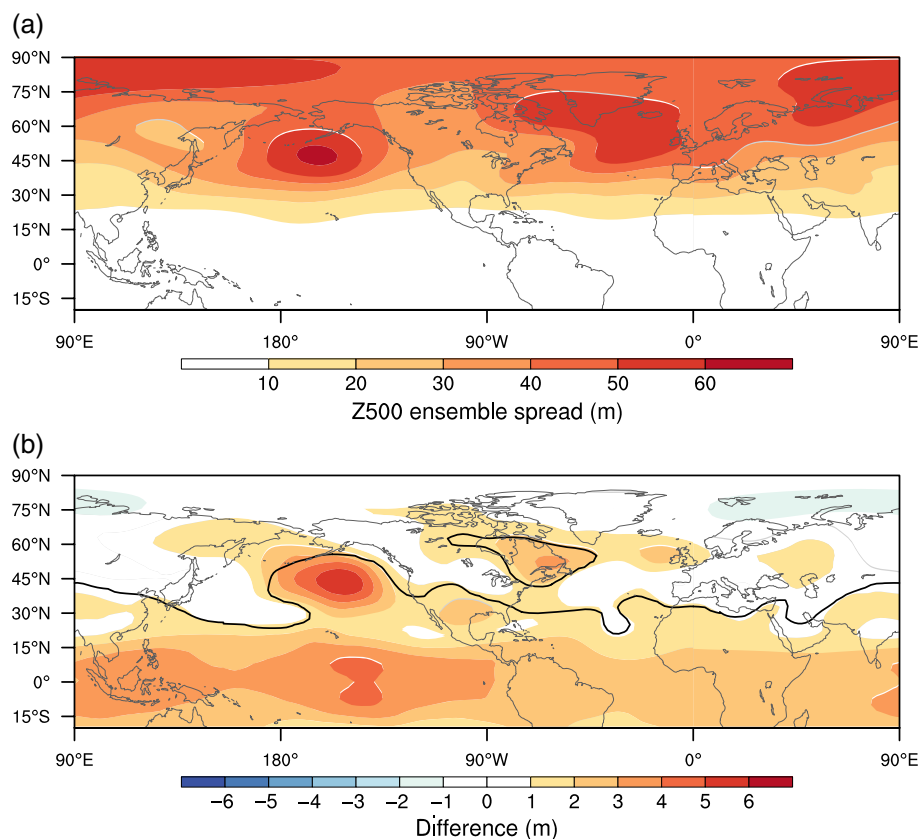


FIGURE 7 (a) Average DJF Z500 ensemble spread in the ERA-40/Int IC experiment. (b) Difference in Z500 ensemble spread between the Correct IC-only ensemble (from the Shuffled experiment) and the ERA-40/Int IC experiment. In (b) positive values show regions of larger spread in the Correct IC-only hindcast. The black contour in (b) outlines regions where the difference in the ensemble spread is significantly different at the 5% level, according to a two-sided *t*-test [Colour figure can be viewed at wileyonlinelibrary.com]

standard deviation across ensemble members for each season) in the ERA-40/Int IC experiment and the difference in the correct IC-only ensemble. The spread in SSTs in the Correct IC-only ensemble leads to a wider spread in the 500 hPa geopotential height field (i.e. Z500) in the Tropics and, even more clearly, in the extratropical North Pacific in the Correct IC-only ensemble. It is also interesting to note that the spread in SST and sea-ice has little influence on the Z500 spread over the Arctic. The $\approx 10\%$ increase in spread over the PNA region in the Correct IC-only ensemble is likely due to the spread in the ENSO SST anomalies and therefore the teleconnection to the Extratropics. There is also an increase in the spread of the stratospheric polar vortex strength (defined using zonal wind at 10 hPa, as before) in the Correct IC-only ensemble of over 5% compared to the ERA-40/Int IC experiment.¹ The increased spread in the PNA region could plausibly be acting to increase the polar vortex spread across the ensemble, which would reduce the relative influence of the QBO on the polar vortex (Garfinkel and Hartmann, 2007; Hansen *et al.*, 2016).

The relationship between the QBO and the NAO in the Correct IC-only experiment is even weaker than the

QBO/polar vortex relationship, in comparison with the ERA-40/Int IC experiment (Figure 6c). Our analysis indicates that the differing SST boundary conditions in the Correct IC-only ensemble increases the ensemble spread of the modelled NAO. The relatively weak relationship between the QBO and the stratospheric polar vortex in the Shuffled IC experiment would be expected to weaken the subsequent downward influence of the stratospheric circulation anomalies on the NAO (Baldwin and Dunkerton, 2001; Song and Robinson, 2004; Scaife *et al.*, 2005; Hitchcock and Simpson, 2014). The variance in the SST boundary condition also seems to have contributed to additional ensemble spread in the troposphere over the North Atlantic sector (Figure 7), which would also effectively weaken any NAO signal related to the QBO.

3.5 | Contribution of the QBO to the hindcast NAO skill

To this point, our analysis indicates that the initial conditions associated with the QBO in the tropical stratosphere lead to increased hindcast skill compared to hindcast experiments which lack accurate stratospheric initial conditions (i.e. ERA-20C IC) or have incorrect atmospheric initial conditions (i.e. the Correct SST-only ensemble). To further evaluate the contribution of the QBO to the NAO hindcast skill, we now statistically remove the QBO signal from the hindcast NAO indices using a linear regression technique. For each of the

¹The average ensemble spread of the polar vortex strength in the Correct IC-only ensemble and ERA-40/Int IC were 9.74 and 9.23 m/s, respectively. The difference between these two ensemble spreads is significant at the 1% level, according to a two-sided *t*-test.

experiments, we took the value of the NAO regression per unit QBO anomaly (i.e. from Figure 6c) and removed the QBO influence from each ensemble member individually, based on the value of the DJF QBO index in each particular ensemble member. The adjusted ensemble mean NAO indices, after regressing out the QBO influence, are shown in Figure 3c. After regressing out the influence of the QBO, the NAO hindcast correlation skill in the ERA-40/Int IC experiment reduces from $r = 0.54$ (Figure 3a) to $r = 0.40$, which is still statistically significant (Table 1). However, in the ERA-20C IC experiment, regressing out the influence of the QBO makes little difference to the NAO hindcast skill, confirming that almost none of the NAO skill was coming from the (incorrect) QBO initial condition. Interestingly, after regressing out the influence of the QBO, the NAO hindcast skill in the ERA-40/Int IC and ERA-20C IC experiments are essentially indistinguishable ($\Delta r = 0.04, p = 0.63$), indicating that much of the increased skill in the ERA-40/Int IC experiment can be directly attributed to the initialization of the QBO in the tropical stratosphere. In the Correct SST-only ensemble, there is no change in the NAO hindcast correlation skill when the influence of the (incorrect) QBO is regressed out (Table 1). The hindcast skill of the NAO in the Correct SST-only ensemble is approximately the same as both the ERA-40/Int IC and ERA-20C IC experiments once the influence of the QBO is regressed out. The NAO hindcast correlation skill of the Correct IC-only ensemble is reduced slightly when the QBO is regressed out, though the skill is very low both with and without the QBO influence, suggesting that any skill originating from the QBO initial conditions is swamped by the wide variety of incorrect SST forcing in the Correct-IC only hindcast.

3.6 | The QBO teleconnection and the signal-to-noise paradox

Whilst it is clear that the QBO initial condition is contributing to the increased NAO hindcast correlation skill in the ERA-40/Int IC experiment, it is also interesting to analyse the influence of the QBO on the *signal* in the ensemble NAO hindcasts. Recently, the Met Office's GloSEA5 and DePreSys3 wintertime seasonal forecasts have demonstrated impressive NAO hindcast skill ($r \approx 0.6$), however the ensemble-mean NAO signal in these systems is seemingly too weak (Scaife *et al.*, 2014a; Dunstone *et al.*, 2016); this so-called “signal-to-noise paradox” has also been found in skilful seasonal hindcasts of summertime European precipitation (Dunstone *et al.*, 2018). To calculate whether the signal-to-noise ratio in the model can be considered appropriate, Eade *et al.* (2014) defined the *Ratio of Predictable Components (RPC)* as the ratio of the predictable component in the real world to the predictable component (or signal-to-noise) in the model:

$$RPC = \frac{r}{\sqrt{\sigma_{\text{ensmean}}^2 / \sigma_{\text{total}}^2}}, \quad (1)$$

where r is the ensemble mean hindcast correlation skill, $\sigma_{\text{ensmean}}^2$ is the ensemble mean variance and σ_{total}^2 is the average variance of individual ensemble members. For an ideal ensemble forecast, *RPC* would be equal to one. However, if the ensemble mean signal is too weak, as is the case in the Met Office seasonal hindcasts of the winter NAO, then *RPC* is greater than one and the real world is seemingly more predictable than the model alone would indicate. The ratio calculated in Equation 1 should be considered a lower bound of the actual *RPC* of the system, since a larger ensemble size would be expected to increase the ensemble mean correlation skill (Kumar, 2009; Eade *et al.*, 2014).

The ensemble NAO hindcast in the ERA-40/Int IC experiment has an *RPC* of 1.51, shown in Figure 8a (blue). To calculate the confidence intervals on the *RPC* value, we performed a bootstrapped resampling (with replacement) 10,000 times over the years of the hindcast; these confidence intervals of the *RPC* are also shown in Figure 8a. In the bootstrapped resampling, only 3% of the random samples have an *RPC* < 1 (i.e. $p = 0.06$ for a two-sided test), suggesting that there is a signal-to-noise issue in the NAO hindcasts in the ERA-40/Int IC experiment. The *RPC* is too high in the ensemble NAO hindcast because the signal-to-noise ratio is too low; more specifically, it is the signal that is too low in the NAO hindcasts, as also found in the hindcasts of Scaife *et al.* (2014a), rather than the noise being too high in the individual ensemble members. The *RPC* values of the NAO hindcasts in the other experiments – the ERA-20C IC experiment and the Correct SST-only and Correct IC-only ensembles – are not significantly different from one (Figure 8a).

We also calculated the *RPC* for the ERA-40/Int IC NAO hindcasts with the QBO influence regressed out (as in Figure 3c), which is shown in Figure 8a (i.e. where the magnitude of QBO amplification is zero). Interestingly, when the influence of the QBO is regressed out of the NAO hindcasts, the *RPC* is just 1.12, which is a reduction compared to the full NAO hindcast in the ERA-40/Int IC experiment (i.e. *RPC* = 1.51). The reduction of the *RPC* once the influence of the QBO has been regressed out indicates that, as well as being a source of increased NAO hindcast skill, the QBO is also a likely source of the signal-to-noise problem in the hindcast. In addition to removing the QBO influence from the ensemble NAO hindcast through linear regression, we can also amplify the QBO influence in the NAO hindcast using a similar linear regression approach. The *RPC*, hindcast NAO correlation skill, and signal-to-noise ratio are shown in Figure 8 for varying magnitudes of QBO amplification from zero (i.e. QBO regressed out entirely) up to six times the QBO signal in the ensemble NAO hindcast. As the QBO magnitude is increased from zero to one, the *RPC* increases along with the NAO correlation skill. However, as the QBO influence on the NAO is amplified further, the *RPC* reduces and is approximately equal to 1 when the amplification of the QBO signal in the ensemble NAO is about four times the original signal. However, despite the reduction in the *RPC*, the hindcast NAO

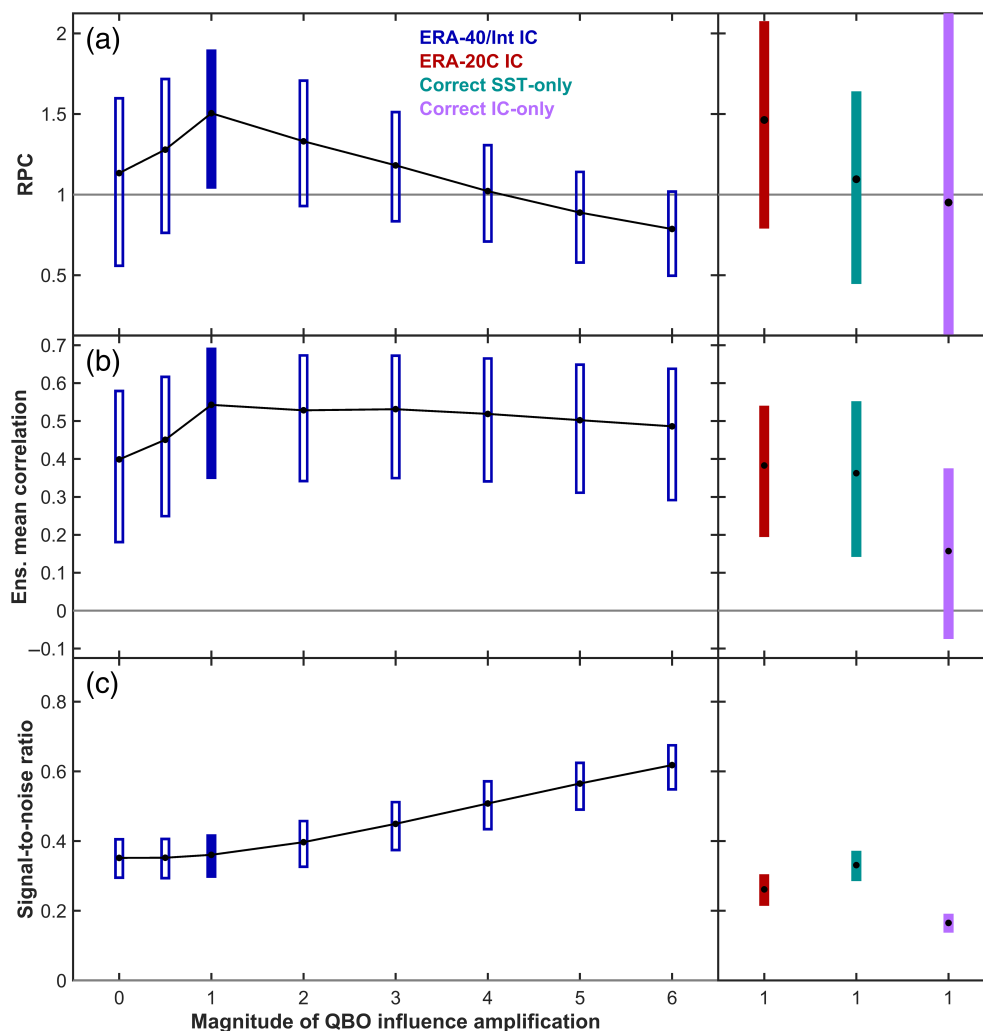


FIGURE 8 (a) The ratio-of-predictable-components (*RPC*) for the NAO in the ERA-40/Int IC, ERA-20C IC, Correct SST-only and Correct IC-only hindcast experiments. Also shown, outlined in blue, are *RPC* values for the post-processed ERA-40/Int IC hindcast experiment, where the QBO influence on the NAO in each ensemble member has been amplified (or reduced) by various factors. (b) and (c) show the same plot for the two components of the *RPC*, namely (b) the ensemble mean NAO correlation (i.e. r) and (c) signal-to-noise ratio (i.e. $\sqrt{\sigma_{\text{ensmean}}^2 / \sigma_{\text{total}}^2}$). The bars indicate the 5–95% confidence intervals estimated using a bootstrap resampling (with replacement) [Colour figure can be viewed at wileyonlinelibrary.com]

correlation skill of $r = 0.52$ is essentially indistinguishable with a $\times 4$ QBO amplification, compared with $r = 0.54$ in the original hindcast (Figure 8b). The *RPC* reduction at these amplified QBO levels is due to the increased ensemble mean NAO signal in the hindcast, as seen in Figure 8c.

That the QBO influence on the NAO resolves the signal-to-noise problem when amplified by four times, whilst retaining approximately the same ensemble mean hindcast skill, suggests that the QBO influence on the NAO is too weak in the ERA-40/Int IC experiment. Our analysis of the QBO influence on the NAO in the hindcast ensemble members, in Figure 6c, also indicated that the influence of the QBO on the NAO was too weak in the ERA-40/Int IC experiment. In the hindcast ensemble members, the regression coefficient is $+0.09$ (NAO per unit QBO anomaly), whereas in the reanalysis the regression coefficient is $+0.31$ (NAO per unit QBO anomaly), although there is considerably more uncertainty in the latter. Nonetheless, the QBO influence on the NAO in the hindcast ensemble is approximately 3.4

times smaller than in the reanalysis, which is very similar to the amplification of 4 needed to reduce the *RPC* to 1 in the scaled QBO hindcasts (i.e. Figure 8a). Therefore, in the ERA-40/Int IC hindcast experiment, it appears that the signal-to-noise problem in the ensemble NAO hindcast can largely be attributed to the weak QBO teleconnection to the NAO.

4 | SUMMARY AND FURTHER DISCUSSION

In this study we have investigated the influence of atmospheric initial conditions on winter seasonal hindcasts of the NAO. The ensemble mean NAO skill is largest in the ERA-40/Int IC experiment (Figure 3), which is initialized with superior initial conditions in the tropical stratosphere (Figure 2). Moreover, the hindcast initialized with reanalysis that only assimilates surface observations, the ERA-20C IC experiment, did not exhibit significantly more NAO hindcast skill than an ensemble in which the atmospheric initial conditions were taken from all the other 49 years (i.e. the

Correct SST-only ensemble). Seasonal hindcasts of the QBO were highly skilful in the ERA-40/Int IC experiment (albeit with a somewhat weaker equatorial zonal wind amplitude in the lower stratosphere) but exhibited no significant skill in the ERA-20C IC experiment, suggesting that the QBO may explain the additional NAO skill in the ERA-40/Int IC experiment. The influence of the QBO was further demonstrated by regressing out the QBO influence from each of the hindcast experiments, after which the difference in NAO hindcast skill between the experiments was negligible.

Whilst the QBO initial condition is identified as the key source of additional NAO skill, the model exhibits some substantial deficiencies in the representation of the QBO and its influence on other aspects of the atmospheric circulation. In particular, the amplitude of the QBO is too low in the model and reduces markedly as the winter progresses. Additionally, the influence of the QBO on both the stratospheric polar vortex (the so-called Holton–Tan relationship, e.g. Holton and Tan (1980; 1982) and the NAO is found to be too weak (Figure 6). The weak Holton–Tan relationship is a plausible explanation for the weak impact on the NAO, since the polar vortex has been shown to impact the NAO (e.g. Baldwin and Dunkerton, 2001; Thompson *et al.*, 2002). However, some modelling studies have argued that the equatorial lower stratosphere (and hence the QBO) can influence the NAO through changes in the subtropical jet and the subsequent impact on the growth of baroclinic waves in the troposphere (e.g. Simpson *et al.*, 2009; Garfinkel and Hartmann, 2011). To test the importance of the polar vortex influence route on the QBO–NAO teleconnection in the skill in our hindcast experiments, we followed the methodology of Gray *et al.* (2018; their figure 1) and performed an additional regression analysis. We linearly regressed the polar vortex influence on the NAO from each ensemble member, as we did for the QBO in Figure 3c. After regressing out the polar vortex influence, the NAO hindcast skill in the ERA-40/Int IC experiment is reduced (from $r = 0.54$) to $r = 0.41$, suggesting that the QBO influence on the NAO in the hindcast occurs via the polar vortex.

Additional evidence for the role of the polar vortex in the hindcast QBO–NAO teleconnection is seen in the ensemble mean polar vortex indices and their correlation skill with reanalysis (Supporting Information Figure S2; summarized in Table 1). The ensemble mean polar vortex skill in the ERA-40/Int IC hindcast ($r = 0.41$) is substantially higher than the ERA-20C IC ($r = 0.27$, $p_{\Delta r} = 0.07$). This further demonstrates that the QBO initial condition is exhibiting an important influence on the seasonal polar vortex strength. The Correct IC-only ensemble also exhibits significant ensemble mean polar vortex skill ($r = 0.28$), however this is somewhat lower than in the ERA-40/Int IC experiment, suggesting that some of the polar vortex hindcast skill in the latter is coming from the coherent tropospheric response to the SST boundary conditions.

The weak QBO teleconnection to the NAO in the model results in a skilful NAO hindcast that nevertheless has too little signal in the ERA-40/Int IC experiment; this is the so-called “signal-to-noise paradox” that has been found in previous studies of wintertime NAO predictability (Eade *et al.*, 2014). Diagnostically amplifying the influence of the QBO on the NAO substantially increases the ensemble mean NAO signal with negligible impact on the ensemble mean NAO hindcast skill. After applying this artificial amplification, the signal-to-noise problem seemingly disappears.

That the QBO is a source of skill in seasonal hindcasts is perhaps to be expected. However, that the QBO is the source of the signal-to-noise problem in these hindcasts is more surprising. Based on our study, we would suggest that it is likely that a weak teleconnection between the QBO and the NAO is contributing to the signal-to-noise problems seen in seasonal hindcasts of the NAO using fully coupled systems (Scaife *et al.*, 2014a; Eade *et al.*, 2014; Stockdale *et al.*, 2015). Given that the QBO has a period of ≈ 28 months and can be skilfully forecast well over a year in advance by some models (e.g. Scaife *et al.*, 2014a), we suggest that the weak QBO teleconnection to the NAO could possibly contribute to the signal-to-noise problem found by Dunstone *et al.* (2016) in their skilful hindcasts of the second winter NAO in the DePreSys3 model; this would particularly be the case if the QBO amplitude is reduced, as we find in the hindcast experiments described in this study. However, it should be noted that, due to sampling uncertainty, the observed QBO–NAO teleconnection is not well constrained. It is possible that some of the additional ensemble mean NAO skill from the QBO teleconnection occurs because in the observational period the NAO exhibits a relatively strong correlation with the QBO largely by chance. Nonetheless, this study highlights the importance of evaluating the representation of the QBO – and its connection to the NAO – in operational seasonal forecasting systems. A similar signal-to-noise issue has been found in skilful seasonal forecasts of summertime European precipitation (Dunstone *et al.*, 2018) and the QBO is unlikely to be responsible for the signal-to-noise issue in these summer hindcasts; however, it could be that another relevant teleconnection mechanism, such as from the tropical Pacific (O'Reilly *et al.*, 2018), could be too weakly represented in these summer hindcasts. Nonetheless, the results of the present study suggest that seasonal forecasting models which capture predictable relationships too weakly may be able to exhibit substantial ensemble mean skill with a large number of ensemble members, but will have a signal that is somewhat too weak.

ACKNOWLEDGEMENTS

We are grateful to ECMWF for making the IFS and necessary computing time available through a member state

special project. COR, TW and AW were funded by Natural Environment Research Council (NERC) grant number NE/M005887/1. LTG acknowledges NERC funding through the National Centre for Atmospheric Science (NCAS) and the NERC ACSIS project (Atlantic Climate System Integrated Study). The insightful comments by the anonymous reviews greatly helped in improving the paper.

ORCID

Christopher H. O'Reilly  <https://orcid.org/0000-0002-8630-1650>

Antje Weisheimer  <https://orcid.org/0000-0002-7231-6974>

Dave MacLeod

REFERENCES

- Alexander, M.A., Bhatt, U.S., Walsh, J.E., Timlin, M.S., Miller, J.S. and Scott, J.D. (2004) The atmospheric response to realistic Arctic sea ice anomalies in an AGCM during winter. *Journal of Climate*, 17, 890–905.
- Andrews, D. and McIntyre, M.E. (1976) Planetary waves in horizontal and vertical shear: the generalized Eliassen–Palm relation and the mean zonal acceleration. *Journal of the Atmospheric Sciences*, 33, 2031–2048.
- Anstey, J.A. and Shepherd, T.G. (2014) High-latitude influence of the quasi-biennial oscillation. *Quarterly Journal of the Royal Meteorological Society*, 140, 1–21.
- Baldwin, M.P. and Dunkerton, T.J. (2001) Stratospheric harbingers of anomalous weather regimes. *Science*, 294, 581–584.
- Baldwin, M.P., Gray, L.J., Dunkerton, T.J., Hamilton, K., Haynes, P.H., Randel, W.J., Holton, J.R., Alexander, M.J., Hirota, I., Horinouchi, T., Jones, D.B.A., Kinnerson, J.S., Marquardt, C., Sato, K. and Takahashi, M. (2001) The quasi-biennial oscillation. *Reviews of Geophysics*, 39, 179–229.
- Boer, G. and Hamilton, K. (2008) QBO influence on extratropical predictive skill. *Climate Dynamics*, 31, 987–1000.
- Brönnimann, S., Luterbacher, J., Staehelin, J., Svendby, T., Hansen, G. and Svenøe, T. (2004) Extreme climate of the global troposphere and stratosphere in 1940–42 related to El Niño. *Nature*, 431, 971–974.
- Christiansen, B., Yang, S. and Madsen, M.S. (2016) Do strong warm ENSO events control the phase of the stratospheric QBO? *Geophysical Research Letters*, 43, 10489–10495.
- Czaja, A. and Frankignoul, C. (2002) Observed impact of Atlantic SST anomalies on the North Atlantic Oscillation. *Journal of Climate*, 15, 606–623.
- Dee, D.P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, I., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F. (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137, 553–597.
- 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.
- Dunstone, N., Smith, D., Scaife, A.A., Hermanson, L., Fereday, D., O'Reilly, C.H., Stirling, A., Eade, R., Gordon, M., MacLachlan, C., Woollings, T., Sheen, K. and Belcher, S.E. (2018) Skilful seasonal predictions of Summer European rainfall. *Geophysical Research Letters*, 45, 3246–3254.
- 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.
- Folland, C.K., Scaife, A.A., Lindesay, J. and Stephenson, D. (2012) How potentially predictable is Northern European winter climate a season ahead? *International Journal of Climatology*, 32, 801–818.
- Free, M. and Seidel, D.J. (2009) Observed El Niño–Southern Oscillation temperature signal in the stratosphere. *Journal of Geophysical Research: Atmospheres*, 114, D23. <https://doi.org/10.1029/2009JD012420>.
- Fujiwara, M., Wright, J.S., Manney, G.L., Gray, L.J., Anstey, J., Birner, T., Davis, S., Gerber, E.P., Harvey, V.L., Hegglin, M.I., Homeyer, C.R., Knox, J.A., Krüger, K., Lambert, A., Long, C.S., Martineau, P., Molod, A., Monge-Sanz, B.M., Santee, M.L., Tegtmeier, S., Chabrilat, S., Tan, D.G.H., Jackson, D.R., Polavarapu, S., Compo, G.P., Dragani, R., Ebisuzaki, W., Harada, Y., Kobayashi, C., McCarty, W., Onogi, K., Pawson, S., Simmons, A., Wargan, K., Whitaker, J.S. and Zou, C.-Z. (2017) Introduction to the SPARC Reanalysis Intercomparison Project (S-RIP) and overview of the reanalysis systems. *Atmospheric Chemistry and Physics*, 17, 1417–1452. <https://doi.org/10.5194/acp-17-1417-2017>.
- García-Herrera, R., Calvo, N., García, R. and Giorgetta, M. (2006) Propagation of ENSO temperature signals into the middle atmosphere: a comparison of two general circulation models and ERA-40 reanalysis data. *Journal of Geophysical Research: Atmospheres*, 111(D6). <https://doi.org/10.1029/2005JD006061>.
- Garfinkel, C. and Hartmann, D. (2007) Effects of the El Niño–Southern Oscillation and the Quasi-Biennial Oscillation on polar temperatures in the stratosphere. *Journal of Geophysical Research: Atmospheres*, 112(D19). <https://doi.org/10.1029/2007JD008481>.
- Garfinkel, C. and Hartmann, D. (2008) Different ENSO teleconnections and their effects on the stratospheric polar vortex. *Journal of Geophysical Research: Atmospheres*, 113(D18). <https://doi.org/10.1029/2008JD009920>.
- Garfinkel, C.I. and Hartmann, D.L. (2011) The influence of the quasi-biennial oscillation on the troposphere in winter in a hierarchy of models. Part I: simplified dry GCMs. *Journal of the Atmospheric Sciences*, 68, 1273–1289.
- Gray, L.J., Anstey, J.A., Kawatani, Y., Lu, H., Osprey, S. and Schenzinger, V. (2018) Surface impacts of the Quasi-Biennial Oscillation. *Atmospheric Chemistry and Physics*, 18, 8227–8247. <https://doi.org/10.5194/acp-18-8227-2018>.
- Hansen, F., Matthes, K. and Wahl, S. (2016) Tropospheric QBO–ENSO interactions and differences between the Atlantic and Pacific. *Journal of Climate*, 29, 1353–1368.
- 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.P. (2015) ERA-20CM: a twentieth-century atmospheric model ensemble. *Quarterly Journal of the Royal Meteorological Society*, 141, 2350–2375.
- Hitchcock, P. and Simpson, I.R. (2014) The downward influence of stratospheric sudden warmings. *Journal of the Atmospheric Sciences*, 71, 3856–3876.
- Holton, J.R. and Tan, H.-C. (1980) The influence of the equatorial quasi-biennial oscillation on the global circulation at 50 mb. *Journal of the Atmospheric Sciences*, 37, 2200–2208.
- Holton, J.R. and Tan, H.-C. (1982) The quasi-biennial oscillation in the Northern Hemisphere lower stratosphere. *Journal of the Meteorological Society of Japan, Series II*, 60, 140–148.
- Hurrell, J.W., Kushnir, Y. and Visbeck, M. (2001) The North Atlantic Oscillation. *Science*, 291, 603–605.
- 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., and Visbeck, M. [Eds.] *The North Atlantic Oscillation: Climatic Significance and Environmental Impact* (Geophysical Monograph 134. Washington, DC: American Geophysical Union, pp. 1–35.
- Kidston, J., Scaife, A.A., Hardiman, S.C., Mitchell, D.M., Butchart, N., Baldwin, M.P. and Gray, L.J. (2015) Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. *Nature Geoscience*, 8, 433.
- Kumar, A. (2009) Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Monthly Weather Review*, 137, 2622–2631.
- MacLeod, D., O'Reilly, C.H., Palmer, T.N. and Weisheimer, A. (2018) Flow-dependent ensemble spread in seasonal forecasts of the boreal winter extratropics. *Atmospheric Science Letters*, 19. <https://doi.org/10.1002/asl.815>.
- Manzini, E., Giorgetta, M., Esch, M., Kornblüeh, L. and Roeckner, E. (2006) The influence of sea surface temperatures on the northern winter stratosphere: ensemble simulations with the MAECHAM5 model. *Journal of Climate*, 19, 3863–3881.
- Molteni, F., Buizza, R., Palmer, T.N. and Petroliagis, T. (1996) The ECMWF ensemble prediction system: methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122, 73–119.

