An intercomparison of skill and overconfidence/underconfidence of the wintertime North Atlantic Oscillation in multimodel seasonal forecasts

AGU100 ADVANCING EARTH AND SPACE SCIENCE

Geophysical Research Letters

RESEARCH LETTER
10.1029/2018GL078838

Key Points:
- Five seasonal forecast systems are shown to skillfully forecast the wintertime North Atlantic Oscillation at 2- to 4-month lead times
- Four of these five systems are underconfident at forecasting the North Atlantic Oscillation
- Winters when the North Atlantic Oscillation is successfully forecast tend to be common to different seasonal forecast systems

Supporting Information:
- Supporting Information S1

Correspondence to:
L. H. Baker,
l.h.baker@reading.ac.uk

Citation:
Baker, L. H., Shaffrey, L. C., Sutton, R. T., Weisheimer, A., & Scaife, A. 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. https://doi.org/10.1029/2018GL078838

Received 18 MAY 2018
Accepted 26 JUN 2018
Accepted article online 5 JUL 2018
Published online 8 AUG 2018

An Intercomparison of Skill and Overconfidence/Underconfidence of the Wintertime North Atlantic Oscillation in Multimodel Seasonal Forecasts

L. H. Baker[1], L. C. Shaffrey[1], R. T. Sutton[1], A. Weisheimer[2,3], and A. A. Scaife[4,5]

[1]National Centre for Atmospheric Science, Department of Meteorology, University of Reading, Reading, UK, [2]NCAS, Department of Physics, University of Oxford, Oxford, UK, [3]European Centre for Medium-Range Weather Forecasts, Reading, UK, [4]Met Office Hadley Centre, Exeter, UK, [5]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

Abstract Recent studies of individual seasonal forecast systems have shown that the wintertime North Atlantic Oscillation (NAO) can be skillfully forecast. However, it has also been suggested that these skillful forecasts tend to be underconfident, meaning that there is too high a proportion of unpredictable noise in the forecasts. We assess the skill and overconfidence/underconfidence of the seasonal forecast systems contributing to the EUROpean Seasonal to Interannual Prediction (EUROSIP) multimodel ensemble system. Five of the seven systems studied have significant skill for forecasting the wintertime NAO at 2- to 4-month lead times. Four of these skillful systems are underconfident for forecasting the NAO. A multimodel ensemble (ensemble size 126 members) is both skillful and clearly underconfident. Underconfidence becomes more pronounced as the ensemble size increases. Certain years in the hindcast period are well forecast by all or most models. This implies that common teleconnections and drivers of the NAO are being captured by the EUROSIP seasonal forecasts.

Plain Language Summary In this paper we provide an intercomparison of seven seasonal forecast systems, with particular focus on the wintertime North Atlantic Oscillation (NAO). The wintertime NAO is the main driver of winter weather variability in the United Kingdom and Europe, and being able to forecast the NAO for the season ahead has potential benefits for many different sectors such as agriculture, energy, health, transport, and water resource management. We show that five of the seven systems studied can skillfully forecast the NAO, and a multimodel ensemble has even higher skill. Four of these skillful systems are found to be underconfident, which means that there is too high a proportion of unpredictable noise in the model. Being underconfident makes it more difficult to fully utilize the skill of a forecast. However, one system is skillful but not underconfident. We also find that there are common years in which the NAO is well forecast by all the skillful systems. This is an important result because it implies that common drivers of NAO predictability are being captured by these systems. These results are an important contribution to our understanding of seasonal forecasts systems and the predictability of the NAO.

1. Introduction

The wintertime North Atlantic Oscillation (NAO) is the main driver of variability in winter sea level pressure, temperature and precipitation over much of Europe and the North Atlantic (e.g., Hanna & Cropper, 2017; Hurrell, 1995; 1996; Hurrell et al., 2003; Trigo et al., 2002). The NAO is a measure of the difference in surface pressure between the Icelandic Low and the Azores High. Positive values of the NAO indicate that this pressure difference is larger than normal, resulting in a stronger North Atlantic jet, often associated with mild, wet winters across northern Europe. Negative NAO values indicate a weaker meridional pressure gradient across the North Atlantic, leading to a weaker jet and more frequent occurrence of atmospheric blocking, often associated with cold, dry winters in northern Europe.

Recent advances in seasonal forecasting models mean that some systems have skill in the midlatitudes for modes of variability such as the NAO. Scaife et al. (2014) showed that the Met Office's seasonal forecast system,

GloSea5, has significant skill (correlation 0.62) for the wintertime NAO at 2- to 4-month lead times for a 20-year period (1992–2011). The GloSea5 forecasts have since been used for applications such as seasonal forecasts of regional UK precipitation (Baker et al., 2017), weather-related impacts on UK transport (Palin et al., 2016), Baltic Sea ice (Karpechko et al., 2015), and UK energy demand and supply (Clark et al., 2017). Athanasiadis et al. (2017) showed that even higher skill (correlations of 0.85) can be obtained for the NAO using a multimodel ensemble (MME) of three seasonal forecast systems (GloSea5, NCEP-CFSv2, and CMCC) but found that the exact level of skill was sensitive to both the hindcast period and the NAO definition used. Dobrynin et al. (2018) showed that similarly high levels of skill (correlations of 0.86) can be obtained for the NAO using a combination of a single seasonal forecast system and a statistical model. Similar levels of skill have also been found for the wintertime Arctic Oscillation in some individual and multimodel seasonal forecast systems (e.g., Athanasiadis et al., 2017; Kang et al., 2014; Stockdale et al., 2015).

As mentioned above, previous studies have shown some sensitivity to the definition of the NAO used when evaluating seasonal forecasts. Observational records at Stykkisholmur, Iceland, and Ponta Delgada, Azores, mean that the NAO is often defined as the difference in surface pressure between these two points (Hurrell et al., 2003) or other similar pairs of points (Li & Wang, 2003). For a more up-to-date, extended Azores-Iceland NAO index, see Cropper et al. (2015). Alternatively, the NAO is defined as the principal component of the first empirical orthogonal function (EOF) in the North Atlantic region (e.g., Barnston & Livezey, 1987; Rogers, 1990; Weisheimer et al., 2017). Both these methods have some disadvantages. The point-based methods may not give a good representation of the full spatial pattern of the NAO (Li & Wang, 2003) and may also suffer from homogeneity issues of the long-running climate stations from which they are constructed (Hanna et al., 2015). Furthermore, they have the disadvantage that even if a model successfully represents the broader scale NAO, errors in the position of the northern and southern centers of action mean that the model skill may appear lower than it should (Stephenson et al., 2006). Because of this, EOF-based NAO indices will generally be forecast with more skill than point-based indices on seasonal time scales, although we note that Domeisen et al. (2018) found that point-based indices have better predictability on time scales of days to weeks, due to missing smaller-scale variability in EOF-based indices. However, using the EOF method can lead to problems with the physical interpretation of results (Li & Wang, 2003; Osborn et al., 1999) and due to differences in the division of modes of variability in models compared to observations (Stephenson et al., 2006).

Another issue identified in previous studies is that some seasonal forecasting systems, while skillful, are underconfident. This means that there is too high a proportion of unpredictable noise in the model, resulting in a smaller ensemble mean variance than would be expected given the correlation between the ensemble mean and observations (Kumar et al., 2014; Scaife et al., 2014). Eade et al. (2014) showed that the GloSea5 system is underconfident for the NAO and in the wider North Atlantic region. The same is true for interannual predictions (Dunstone et al., 2016). Athanasiadis et al. (2017) found that the three models in their study and the MME were also underconfident. A similar underconfidence was found for the European Centre for Medium-Range Weather Forecasts (ECMWF) system Arctic Oscillation forecasts by Stockdale et al. (2015). The cause of the underconfidence is not fully understood and is an ongoing area of research. However, it is not due to incorrect internal variability, since the interannual variance of individual ensemble members matches the observed interannual variance.

It is clear from the above discussion that there are major knowledge gaps regarding the level of skill in seasonal forecasts of the wintertime NAO, the extent that skill depends on the definition of the NAO and the degree to which seasonal forecasts of the wintertime NAO are overconfident or underconfident. To address these gaps, we assess the skill and overconfidence/underconfidence of NAO forecasts from seasonal forecast systems that participate in the EUROSIP (EUROpean Seasonal to Interannual Prediction) MME (Stockdale, 2012). The EUROSIP ensemble contains seasonal hindcasts from each of the contributing modeling systems, and these are used in this study.

In section 2 the model data and NAO definitions used in this study are described. In section 3 the skill and overconfidence/underconfidence of the forecast systems in the North Atlantic are evaluated, and in section 4 the seasonal forecasts of the wintertime NAO are evaluated. Finally, in section 5 these results are discussed.

2. Data and NAO Definition

2.1. Data

Hindcasts are analyzed for the winter season (December, January, and February [DJF] averages) in seven seasonal forecasting systems from EUROSIP for the 20-year period 1992/1993–2011/2012. This period is the longest hindcast period common to all systems. For those models for which data are available, results for a 30-year period (1982/1983–2011/2012) are shown in the supporting information. In general, the results are similar for the two different periods.

The systems studied are Met Office GloSea5 (MacLachlan et al., 2015), ECMWF System 4 (Molteni et al., 2011; hereafter ECMWF), NCEP-CFSv2 (Saha et al., 2014; hereafter NCEP), Meteo-France Systems 3 and 4 (Voldoire et al., 2013; hereafter MFr-S3 and MFr-S4), and the Japanese Meteorological Agency (JMA) system 2 (Takaya et al., 2017). Two versions of the Met Office's GloSea5 system, which we refer to as MO-GA3 and MO-GA6, were used instead of the Met Office data in the EUROSIP archive, which originates from more than one model version due to operational upgrades to the Met Office system. Furthermore, NCEP hindcasts from http://nomads.ncdc.noaa.gov/data.php were used instead of the NCEP data in the EUROSIP archive, since members initialized after the start of November did not have hindcast data for December archived. Key details of these forecast systems are given in Table S1. The ensemble size varies between systems, ranging from 10 to 51 members. Although the hindcasts are all initialized around the start of November, the exact start dates of the hindcasts differ between models: Ensemble members in the Meteo-France, ECMWF, and JMA systems all start on 1 November; Met Office ensemble members start on 25 October, 1 November, and 9 November (eight members each); and NCEP ensemble members start on 23 October, 28 October, 2 November, and 7 November (four members on each). The winter forecasts are therefore forecasts of approximately 2- to 4-month lead times. The start dates after 1 November were included for the Met Office and NCEP systems in order to increase their ensemble size, and the range of dates was chosen to be centered around 1 November. We note that ensemble members initialized after 1 November will in general have slightly better skill than those initialized earlier (although this is not the case in MO-GA6).

The mean sea level pressure (MSLP) from ERA-Interim (Dee et al., 2011) is used to characterize the atmospheric circulation and validate the hindcasts. All model and reanalysis data are regridded to a regular $5^\circ \times 5^\circ$ grid.

2.2. Definition of the NAO

Two definitions of the NAO are used. The first, defined by Stephenson et al. (2006), is the difference in MSLP averaged over a southern box (90°W–60°E, 20°N–55°N) and a northern box (90°W–60°E, 55°N–90°N). We refer to this as the box-based NAO index. The second, which we refer to as the point-based NAO index, is the difference in MSLP between the Azores and Iceland (as used by, e.g., Scaife et al., 2014). The boxes and the Azores and Iceland points are shown in Figure S1. Results for three intermediate box sizes are shown in the supporting information (Figure S2).

3. Evaluation of Wintertime Seasonal Forecasts of MSLP in the North Atlantic

In this section the skill and overconfidence/underconfidence of each seasonal forecast system is evaluated, for the ensemble size available for each system. The anomaly correlation coefficient (ACC) and the ratio of predictable components (RPC; Eade et al., 2014) are used to measure skill and overconfidence or underconfidence, respectively; these are described in the supporting information. RPC values greater than 1 indicate underconfidence. RPC values less than 1 may indicate overconfidence but may also occur due to small ensemble size. Note that we do not include the correction term suggested by Siegert et al. (2016) to correct for limited ensemble size; see the supporting information for more details. To examine the behavior of a very large ensemble (126 ensemble members), a MME is constructed by combining all ensemble members with equal weight from all systems excluding MFr-S4 and JMA, since these systems are shown in section 4 not to have significant skill for forecasting the wintertime NAO.

Maps of ACC skill for DJF MSLP are shown in Figure 1. All the systems have high skill (ACC > 0.7) in the tropics, with the highest values found in the tropical Pacific and around the Maritime Continent. High skill in this region is expected due to the predictability of El Niño–Southern Oscillation, which can be forecast skillfully several months ahead (e.g., Jin et al., 2008; Weisheimer et al., 2009). In the North Atlantic the level of skill differs between systems. The MO-GA3 system has significant skill everywhere except for the Gulf Stream region (Figure 1a). The skill around both NAO centers of action is greater than 0.5, and there is also significant skill

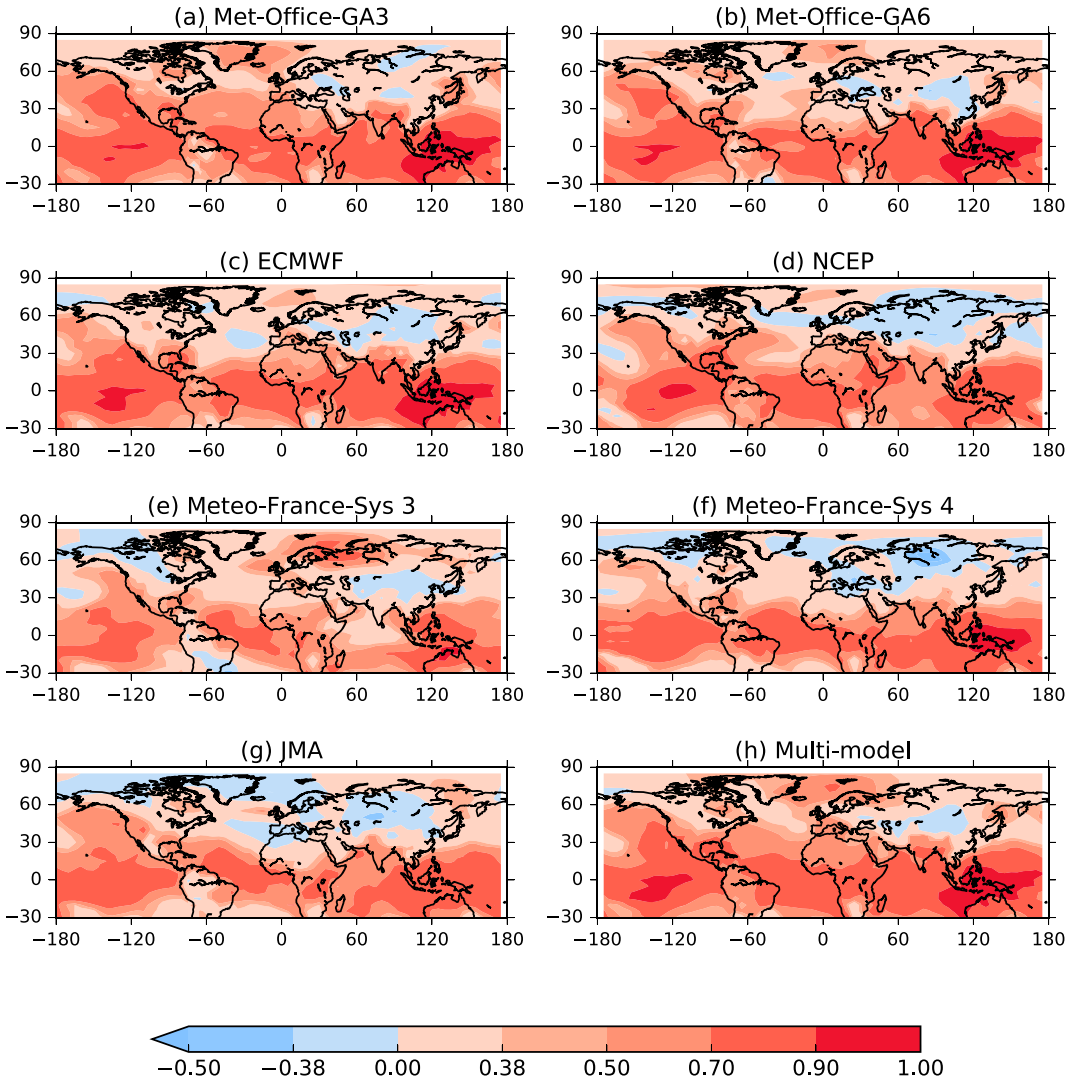

Figure 1. Anomaly correlation coefficient (ACC) for the ensemble mean in each system for winters 1992–1993 to 2011–2012. (a) Met Office GloSea5-GA3, (b) Met Office GloSea5-GA6, (c) European Centre for Medium-Range Weather Forecasts, (d) National Centers for Environmental Prediction, (e) Meteo-France-System 3, (f) Meteo-France-System 4, (g) Japanese Meteorological Agency, and (h) the multimodel ensemble. Values larger than 0.38 are significantly greater than 0 at the 95% confidence level.

extending over the United Kingdom. The MO-GA6 system has similar but lower skill, with significant skill only east of Greenland and around the Azores (Figure 1b). ECMWF has lower but significant skill to the east of Greenland, over central Europe, and in the subtropical Atlantic (Figure 1c). The NCEP system has significant skill in the western North Atlantic and to the east of Greenland (Figure 1d). MFr-S3 has significant skill extending from west of the United Kingdom down into the Mediterranean and over Scandinavia and the Barents Sea. The remaining two systems, MFr-S4 and JMA (Figures 1f and 1g), exhibit less skill in the North Atlantic, although JMA does show a small region of significant skill to the west of the United Kingdom. The ACC for the MME is significant over most of the North Atlantic and over Greenland and northern Europe (Figure 1h).

Maps of RPC for DJF MSLP forecasts are shown in Figure 2. All systems have RPC values close to 1 in the tropics, indicating that the forecasts are neither overconfident nor underconfident. The two Met Office systems have high RPC values in the North Atlantic (Figures 2a and 2b), forming a tripole structure, as shown by Eade et al. (2014). In other systems, while the RPC is generally below 1, there are some areas in and around the North Atlantic with RPC greater than 1. The ECMWF and NCEP systems have RPC > 1 over southern Europe and to the northeast of Iceland (Figures 2c and 2d). MFr-S3 (Figure 2e) has RPC values greater than 1 extending from the west of the United Kingdom across Scandinavia and the Barents Sea, corresponding to the high ACC

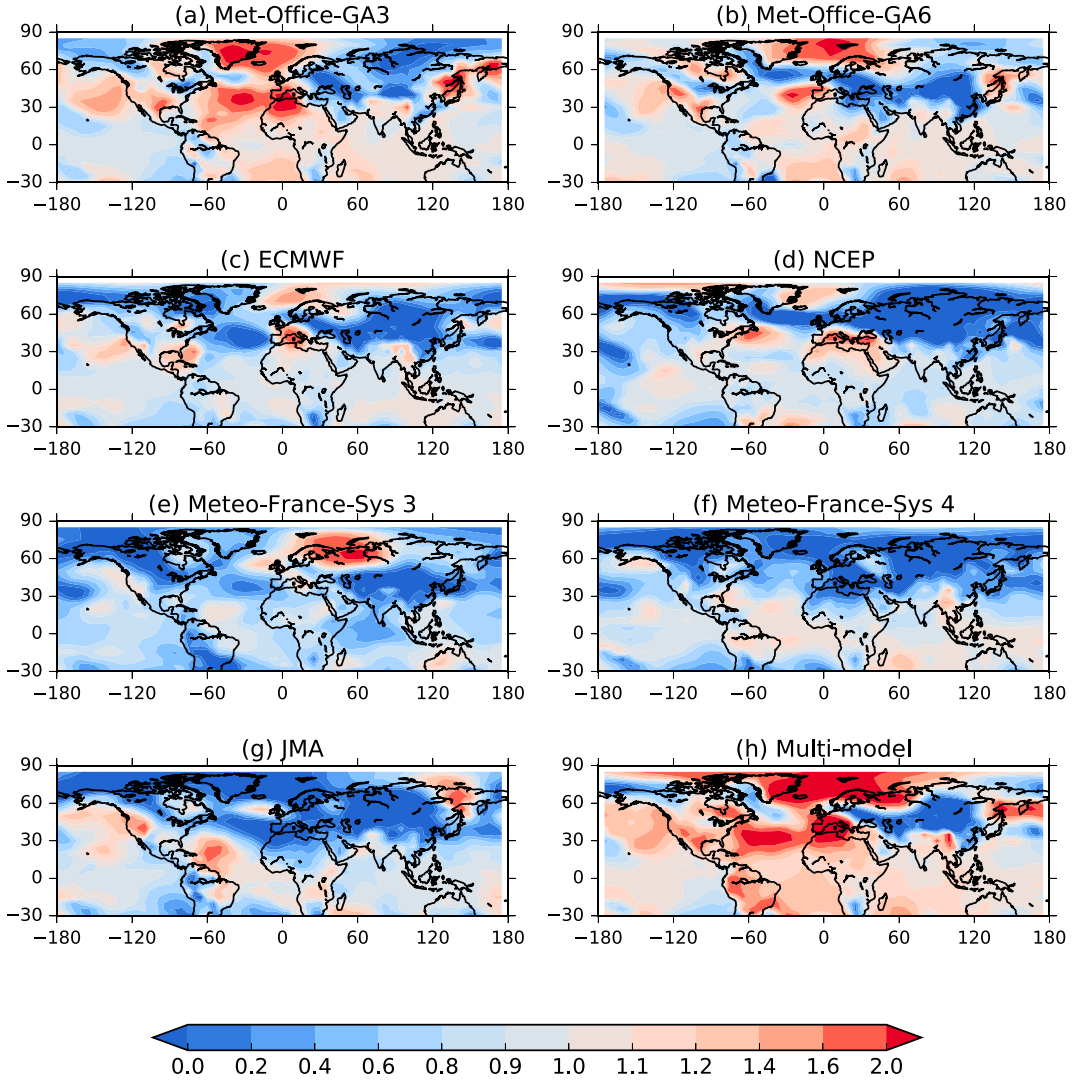

Figure 2. Ratio of predictable components (RPCs) in each system for winters 1992–1993 to 2011–2012. (a) Met Office GloSea5-GA3, (b) Met Office GloSea5-GA6, (c) European Centre for Medium-Range Weather Forecasts, (d) National Centers for Environmental Prediction, (e) Meteo-France-System 3, (f) Meteo-France-System 4, (g) Japanese Meteorological Agency, and (h) the multimodel ensemble.

values in the same regions (see Figure 1e). MFr-S4 and JMA (Figures 2f and 2g) have RPC < 1 over the whole North Atlantic, except for a small region to the west of the United Kingdom in JMA. These results suggest that the two Met Office systems are underconfident over much of the North Atlantic and in particular over the NAO centers of action; other systems show less evidence of underconfidence in the North Atlantic, except in limited regions. It should be noted, however, that these RPC values may be underestimates due to the limited ensemble size, and therefore, with larger ensemble size these systems may also show underconfidence in these areas. The MME (Figure 2h) shows RPC > 2 over most of the North Atlantic and over Greenland and northern Europe, indicating that it is underconfident.

4. Evaluation of Seasonal Hindcasts of the Wintertime NAO

4.1. Wintertime NAO Skill and RPC

The DJF ACC skill for each system for the box-based and point-based NAO indices defined in section 2 is shown in Figure 3a. Consistent with the results from section 3, the MO-GA3 system has the highest skill for the point-based NAO index. Other systems have lower, but still positive, skill, except for MFr-S4 and JMA, which show very little skill. In contrast, five of the systems (all but MFr-S4 and JMA) have significant skill for the box-based NAO index, although this skill is lower and only marginally significant in the ECMWF system. The

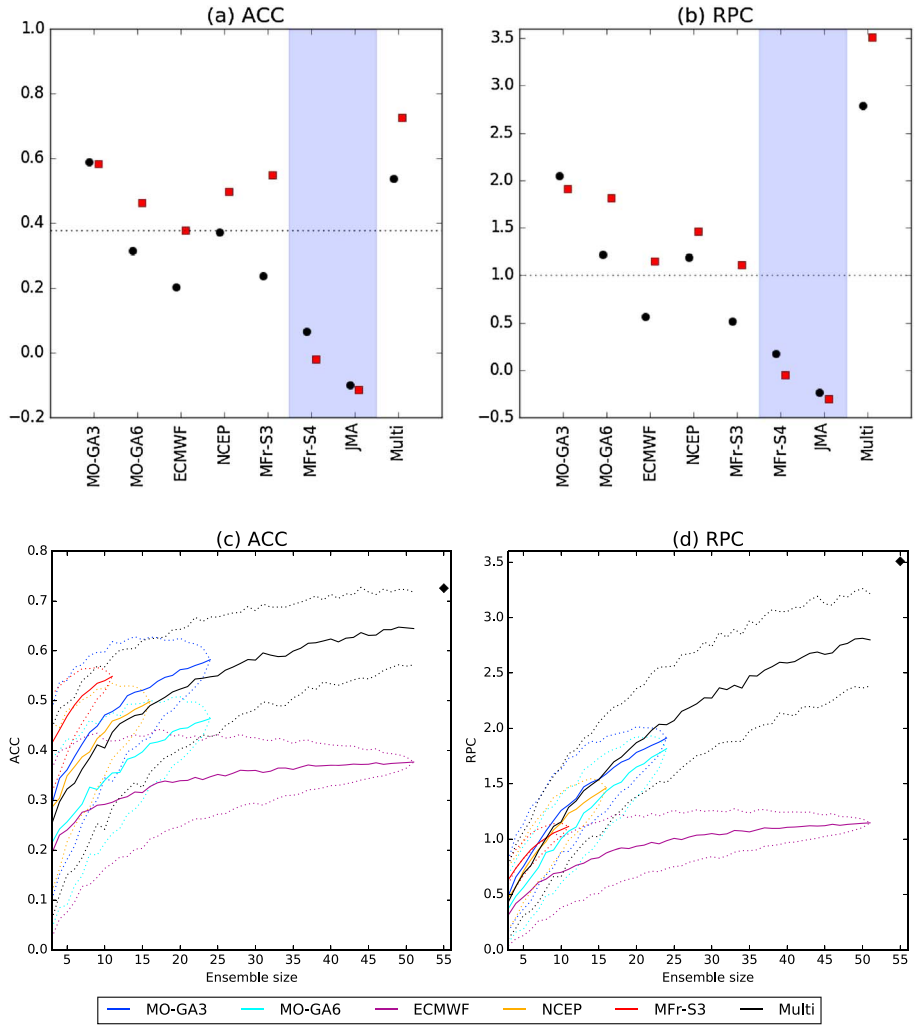

Figure 3. (a) Anomaly correlation coefficient (ACC) and (b) ratio of predictable components (RPC) for the point-based (black circles) and box-based (red squares) winter North Atlantic Oscillation (NAO) indices for all ensemble members in each system, for the hindcast period 1992 −2011. The dotted line in (a) shows the 95% significance level for nonzero skill; the dotted line in (b) shows the divide between overconfidence and underconfidence. (c) ACC and (d) RPC against ensemble size for each system for the box-based NAO index (omitting MFr-S4 and JMA for clarity). Solid and dotted lines show the mean and mean ± one standard deviation, respectively, over 500 random selections of each number of ensemble members. Black diamonds show the values for the multimodel ensemble with 126 members. MO-GA3 = Met Office GloSea5-GA3; MO-GA6 = Met Office GloSea5-GA6; ECMWF = European Centre for Medium-Range Weather Forecasts; NCEP = National Centers for Environmental Prediction; MFr-S3 = Meteo-France-System 3; MFr-S4 = Meteo-France-System 4; JMA = Japanese Meteorological Agency.

MO-GA3 system has almost identical skill for the two NAO indices. In the remaining systems, the significant skill for the box-based NAO suggests that these systems capture the broad-scale pressure pattern well. Reducing the sizes of the boxes used to define the NAO index results in a reduction in skill (Figure S2). MFr-S4 and JMA show no skill for either NAO index, consistent with the lack of skill seen over the North Atlantic region for these systems in Figure 1.

For the box-based NAO index the MME (comprising the five skillful systems) has a correlation skill of 0.73, which is higher than for any individual system; for the point-based NAO index the MME correlation skill is 0.54, which is lower than that obtained by the MO-GA3 system.

The RPC for the NAO forecasts for each system is shown in Figure 3b. Only the MO-GA3 system has RPC values higher than 2 for the point-based NAO index; the MO-GA6 and NCEP systems have RPC just over 1 for this index, while ECMWF and MFr-S3 have RPC values around 0.5, which may mean that they are overconfident

or may be a result of their limited ensemble size. MFr-S4 and JMA have RPC values close to 0, which is due to their very low NAO skill. For the box-based NAO index, the MO-GA3, MO-GA6, and NCEP systems have RPC values greater than 1.4, while ECMWF and MFr-S3 have RPC values around 1. These results suggest that, for the box-based NAO index, the Met Office and NCEP systems are skillful but underconfident, while the ECMWF and MFr-S3 systems are skillful but not underconfident. We note again, however, the underestimation of RPC for small ensemble size. For the MME, the RPC is 2.79 for the point-based NAO index and 3.51 for the box-based NAO index. This suggests that the MME, with its large ensemble size, is clearly underconfident.

For comparison, the ACC and RPC were calculated for the NAO index defined as the leading EOF over the North Atlantic. Results are very similar to those for the box-based NAO index and are shown in the supporting information (Figure S3).

To assess the dependence on ensemble size, Figure 3c shows the ACC for the box-based NAO index, for different numbers of subsampled ensemble members in each system. MFr-S4 and JMA are omitted for clarity, because they were shown to have no skill for the NAO. For all systems, the ACC increases with ensemble size, in agreement with Kumar (2009), Scaife et al. (2014), and Athanasiadis et al. (2017). With the exception of ECMWF, all systems and the MME have a similar slope of the ACC curve, and much of the variation in ACC between systems can be explained by their different ensemble sizes. It is reasonable to hypothesize, based on the similarity in the curves for the available ensemble sizes, that all these systems might achieve a correlation skill above 0.7 if they had as large an ensemble size as the MME. In contrast, the ECMWF system appears to be converging to a somewhat smaller ACC, despite its relatively large ensemble size (51 members). The correlations are generally higher than those shown in Figure 4 of Butler et al. (2016), which is due to the different NAO index used and may also be due to different hindcast periods used.

Like the ACC, the RPC also increases with ensemble size (Figure 3d). The RPC curves for the MO-GA3, MO-GA6, and NCEP systems follow roughly the same curve as the MME, tending to values significantly greater than 1 with increasing ensemble size. The MFr-S3 system curve is also similar to that of the MME, and extrapolation by eye suggests that with a larger ensemble size the RPC would increase further above 1 (i.e., the system is underconfident) although it is difficult to conclude this for definite with such a small ensemble size. In contrast, the ECMWF system has RPC values close to 1 and does not show a tendency toward underconfidence even for its relatively large ensemble size. The ECMWF system therefore has the correct proportion of unpredictable noise, given its skill.

4.2. Temporal Evolution of Wintertime NAO Skill

In this section the temporal evolution of the NAO forecasts is investigated. A time series of anomalies of the box-based winter NAO index for the period 1992–2011 is shown in Figure 4a, for ERA-Interim and for the ensemble mean from each seasonal forecast system (excluding MFr-S4 and JMA for clarity due to their lack of skill) and the MME. The interannual variability of the ensemble mean seasonal forecasts is clearly smaller than the variability in ERA-Interim (interannual standard deviations ranging from 1.02 to 1.57 hPa in the forecasting systems, compared with 5.59 hPa in ERA-Interim). This means that, although the forecasts are skillful in terms of correlation scores, a seasonal NAO forecast from these systems will generally consist of a small shift in the forecast probability density function (pdf) from the climatological pdf. The interannual standard deviation of individual ensemble members (not shown) roughly matches that of ERA-Interim.

To aid visual comparison between the forecast systems and ERA-Interim, the normalized NAO indices are shown in Figure 4b. There are some years when all or nearly all systems do well in capturing the sign and magnitude of the observed (normalized) index. Winters 2009–2010 and 2010–2011 are those with the most strongly negative observed NAO (with 2009–2010 having an exceptionally strong negative NAO; Rivière & Drouard, 2015) and are generally well forecast. Similarly, winters 1992–1993, 1994–1995, 1999–2000, 2007–2008, and 2011–2012 are those with the most strongly positive observed NAO and are also generally well forecast, with four or five systems forecasting the correct sign of NAO anomaly in each year. In years with weakly positive or weakly negative NAO, the forecast systems tend to do less well, consistent with the results of Athanasiadis et al. (2017) who showed low skill for middle-tercile MSLP forecasts and Weisheimer et al. (2017), who found low skill for weakly negative NAO years but much higher skill for strongly negative NAO winters in a long atmosphere-only seasonal hindcast data set. The correlation skill is robust to removing individual winters with particularly high or low skill, which gives correlations within 0.1 of the values obtained from the full time series for all models and the MME. The MME mean forecasts the correct sign of NAO anomaly

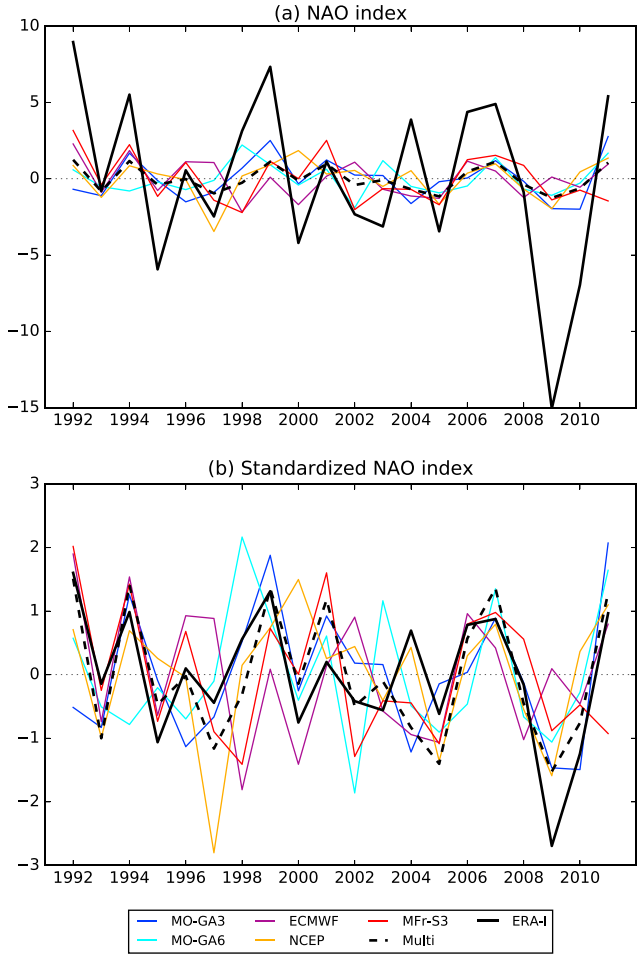

Figure 4. Time series of the box-based (a) NAO anomalies (relative to the 20-year mean), in hectopascals, and (b) NAO anomalies normalized by the standard deviation over the 20-year period, for ERA-Interim and for the ensemble mean in each seasonal forecast system and the multimodel ensemble. MFr-S4 and JMA systems are omitted for clarity. The years on the x axis indicate the start of the relevant winter (e.g., 1992 represents winter 1992 −1993). NAO = North Atlantic Oscillation; MO-GA3 = Met Office GloSea5-GA3; MO-GA6 = Met Office GloSea5-GA6; ECMWF = European Centre for Medium-Range Weather Forecasts; NCEP = National Centers for Environmental Prediction; MFr-S3 = Meteo-France-System 3; MFr-S4 = Meteo-France-System 4; JMA = Japanese Meteorological Agency.

in 18 of the 20 years. The fact that certain years are well forecast by most systems suggests that there are processes or teleconnections leading to these years being more predictable. Scaife et al. (2016) showed that the NAO seasonal forecast skill in GloSea5-GA3 is conditional on the presence of sudden stratospheric warmings (Sigmond et al., 2013) but originates from tropical rainfall (Scaife et al., 2017). Stockdale et al. (2015) and Scaife et al. (2017) argue that factors such as the El Niño–Southern Oscillation play a role, and Butler et al. (2016) showed that models that resolve the stratospheric response to tropical drivers give more skillful NAO forecasts. This is in agreement with Domeisen et al. (2015) who showed that predictability over Europe is only increased by El Niño events when sudden stratospheric warming events also occur. Due to these complexities, there is no clear pattern in our data showing better forecasts specifically in El Niño or La Niña years. Other recent studies (Caian et al., 2018; Dobrynin et al., 2018; Folland et al., 2012; Hall et al., 2017; Wang et al., 2017) have shown that factors such as Atlantic sea surface temperatures and Arctic sea ice extent also have an influence on the NAO, while Lin et al. (2015) and Garfinkel et al. (2014) suggested an influence of certain phases of the Madden-Julian Oscillation. Disentangling the sources of predictability in individual years is beyond the scope of this study and is an important ongoing area of investigation.

5. Conclusions

In this study, the skill and underconfidence/overconfidence of forecasts of the wintertime NAO from seven EUROSIP seasonal forecast systems were examined. Five systems were shown to have significant skill for forecasting the wintertime NAO at lead times of 2–4 months, over the hindcast period 1992–2011, when the NAO was defined as the pressure difference in two large areas; only one system skillfully forecasts the Azores-Iceland pressure difference. The MME had significant skill for both indices. A clear increase in skill with increasing ensemble size was found, consistent with other studies (Athanasiadis et al., 2017; Kumar, 2009; Scaife et al., 2014); therefore, higher correlation skill is expected if a larger ensemble size were used. Three systems (MO-GA3, MO-GA6, and NCEP) and the MME were found to have RPC values much greater than 1 for the box-based NAO forecasts, implying underconfidence, with a tendency toward larger RPC values with increasing ensemble size. The MFr-S3 system has RPC only just greater than 1, but this is likely due to its small ensemble size (11 members) and extrapolation by eye suggests that for a larger ensemble size this system would also show RPC much greater than 1 and thus be underconfident. In contrast, the ECMWF system has RPC values of around 1, even for a relatively large ensemble size (51 members) indicating that these forecasts are not underconfident, despite being skillful, and this system therefore has the correct proportion of unpredictable noise given its skill. This suggests that the results of Scaife et al. (2014) and Eade et al. (2014), that the skillful Met Office NAO forecasts are underconfident, apply to most, but not all, seasonal forecasting systems that can skillfully forecast the NAO. It should be noted that the interannual variance of individual ensemble members is close to that of ERA-Interim, so the underconfidence is not simply due to the incorrect amount of variability within the model. Finally, there are several years in the period studied in which the NAO is well forecast by all or nearly all of the forecasting systems. These tend to be years with strongly positive or strongly negative observed NAO. The fact that certain years are well forecast by all or most systems means that the processes that are key for forecasting the NAO in these cases are represented to at least some degree by several forecast systems. Future research efforts should aim to identify these processes and work to further improve their representation in forecasting systems, while also trying to identify the sources of underconfidence demonstrated by the more skillful systems.

Acknowledgments

The authors thank ECMWF for making the EUROSIP data available. L. H. B. was supported by the NERC project IMPETUS (ref. NE/L010488/1). A. A. S. was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). The authors thank Tim Palmer, Magdalena Balmaseda, Tim Stockdale, David Stephenson, and Stefan Siegert for useful discussion. EUROSIP data can be obtained from https://www.ecmwf.int/en/forecasts/datasets.

References

Athanasiadis, P. J., Bellucci, A., Scaife, A. A., Hermanson, L., Materia, S., Sanna, A., et al. (2017). A multi-system view of wintertime NAO seasonal predictions. *Journal of Climate*, *30*, 1461–1475. https://doi.org/10.1175/JCLI-D-16-0153.1

Baker, L. H., Shaffrey, L. C., & Scaife, A. A (2017). Improved seasonal prediction of UK regional precipitation using atmospheric circulation. *International Journal of Climatology*, *38*(S1), e437–e453. https://doi.org/10.1002/joc.5382

Barnston, A. G., & Livezey, R. E. (1987). Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review*, *115*, 1083–1126.

Butler, A. H, Arribas, A., Athanassiadou, M., Baehr, J., Calvo, N., Charlton-Perez, A., et al. (2016). The climate-system historical forecast project: Do stratosphere-resolving models make better seasonal climate predictions in boreal winter? *Quarterly Journal of the Royal Meteorological Society*, *142*(696), 1413–1427. https://doi.org/10.1002/qj.2743

Caian, M., Koenigk, T., Döscher, R., & Devasthale, A. (2018). An interannual link between Arctic sea-ice cover and the North Atlantic Oscillation. *Climate Dynamics*, *50*(1-2), 423–441.

Clark, R. T., Bett, P. E., Thornton, H. E., & Scaife, A. A. (2017). Skilful seasonal predictions for the European energy industry. *Environmental Research Letters*, *12*(2), 024002.

Cropper, T., Hanna, E., Valente, M. A., & Jónsson, T. (2015). A daily Azores–Iceland North Atlantic Oscillation index back to 1850. *Geoscience Data Journal*, *2*(1), 12–24. https://doi.org/10.1002/gdj3.23

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Dobrynin, M., Domeisen, D. I., Müller, W. A., Bell, L., Brune, S., Bunzel, F., et al. (2018). Improved teleconnection-based dynamical seasonal predictions of boreal winter. *Geophysical Research Letters*, *45*, 3605–3614. https://doi.org/10.1002/2018GL077209

Domeisen, D. I., Badin, G., & Koszalka, I. M. (2018). How predictable are the Arctic and North Atlantic Oscillations? Exploring the variability and predictability of the Northern Hemisphere. *Journal of Climate*, *31*(3), 997–1014.

Domeisen, D. I., Butler, A. H., Fröhlich, K., Bittner, M., Müller, W. A., & Baehr, J. (2015). Seasonal predictability over Europe arising from El Niño and stratospheric variability in the MPI-ESM seasonal prediction system. *Journal of Climate*, *28*(1), 256–271.

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

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

Folland, C., Scaife, A., Lindesay, J., & Stephenson, D. (2012). How potentially predictable is northern European winter climate a season ahead? *International Journal of Climatology*, *32*(6), 801–818.

Garfinkel, C. I., Benedict, J. J., & Maloney, E. D. (2014). Impact of the MJO on the boreal winter extratropical circulation. *Geophysical Research Letters*, *41*, 6055–6062. https://doi.org/10.1002/2014GL061094

Hall, R. J., Scaife, A. A., Hanna, E., Jones, J. M., & Erdélyi, R. (2017). Simple statistical probabilistic forecasts of the winter NAO. *Weather and Forecasting*, *32*(4), 1585–1601. https://doi.org/10.1175/WAF-D-16-0124.1

Hanna, E., & Cropper, T. E. (2017). North Atlantic Oscillation.

Hanna, E., Cropper, T. E., Jones, P. D., Scaife, A. A., & Allan, R. (2015). Recent seasonal asymmetric changes in the NAO (a marked summer decline and increased winter variability) and associated changes in the AO and Greenland Blocking Index. *International Journal of Climatology*, *35*(9), 2540–2554. https://doi.org/10.1002/joc.4157

Hurrell, J. W. (1995). Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation. *Science*, *269*, 676–679. https://doi.org/10.1126/science.269.5224.676

Hurrell, J. W. (1996). Influence of variations in extratropical wintertime teleconnections on Northern Hemisphere temperature. *Geophysical Research Letters*, *23*(6), 665–668. https://doi.org/10.1029/96GL00459

Hurrell, J., Kushner, Y., Ottersen, G., & Visbeck, M. (2003). The North Atlantic Oscillation: Climatic significance and environmental impact. In J. Hurrell, Y. Kushner, G. Ottersen, & M. Visbeck (Eds.), *An overview of the North Atlantic Oscillation* (pp. 1–35). Washington, DC: American Geophysical Union.

Jin, E. K., Kinter, J. L., Wang, B., Park, C.-K., Kang, I.-S., Kirtman, B. P., et al. (2008). Current status of ENSO prediction skill in coupled ocean–atmosphere models. *Climate Dynamics*, *31*(6), 647–664. https://doi.org/10.1007/s00382-008-0397-3

Kang, D., Lee, M.-I., Im, J., Kim, D., Kim, H.-M., Kang, H.-S., et al. (2014). Prediction of the Arctic Oscillation in boreal winter by dynamical seasonal forecasting systems. *Geophysical Research Letters*, *41*, 3577–3585. https://doi.org/10.1002/2014GL060011

Karpechko, A. Y., Peterson, K. A., Scaife, A. A., Vainio, J., & Gregow, H. (2015). Skilful seasonal predictions of Baltic sea ice cover. *Environmental Research Letters*, *10*, 044007.

Kumar, A. (2009). Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Monthly Weather Review*, *137*(8), 2622–2631.

Kumar, A., Peng, P., & Chen, M. (2014). Is there a relationship between potential and actual skill? *Monthly Weather Review*, *142*(6), 2220–2227. https://doi.org/10.1175/MWR-D-13-00287.1

Li, J., & Wang, J. X. (2003). A new North Atlantic Oscillation index and its variability. *Advances in Atmospheric Sciences*, *20*(5), 661–676. https://doi.org/10.1007/BF02915394

Lin, H., Brunet, G., & Yu, B. (2015). Interannual variability of the Madden-Julian Oscillation and its impact on the North Atlantic Oscillation in the boreal winter. *Geophysical Research Letters*, *42*, 5571–5576. https://doi.org/10.1002/2015GL064547

MacLachlan, C., Arribas, A., Peterson, K., Maidens, A., Fereday, D., Scaife, A., et al. (2015). Global seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, *141*, 1072–1084. https://doi.org/10.1002/qj.2396

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., et al. (2011). The new ECMWF seasonal forecast system (System 4) (*Technical Report*): ECMWF. www.ecmwf.int

Osborn, T. J., Briffa, K. R., Tett, S. F., Jones, P. D., & Trigo, R. M. (1999). Evaluation of the North Atlantic Oscillation as simulated by a coupled climate model. *Climate Dynamics*, *15*(9), 685–702.

Palin, E. J., Scaife, A. A., Wallace, E., Pope, E. C., Arribas, A., & Brookshaw, A. (2016). Skillful seasonal forecasts of winter disruption to the UK transport system. *Journal of Applied Meteorology and Climatology*, *55*(2), 325–344. https://doi.org/10.1175/JAMC-D-15-0102.1

Rivière, G., & Drouard, M. (2015). Understanding the contrasting North Atlantic Oscillation anomalies of the winters of 2010 and 2014. *Geophysical Research Letters*, *42*, 6868–6875. https://doi.org/10.1002/2015GL065493

Rogers, J. C (1990). Patterns of low-frequency monthly sea level pressure variability (1899–1986) and associated wave cyclone frequencies. *Journal of Climate*, *3*(12), 1364–1379. https://doi.org/10.1175/1520-0442(1990)003<1364:POLFMS>2.0.CO;2

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., et al. (2014). The NCEP climate forecast system version 2. *Journal of Climate*, *27*(6), 2185–2208. https://doi.org/10.1175/JCLI-D-12-00823.1

Scaife, A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R., Dunstone, N., et al. (2014). Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, *41*, 2514–2519. https://doi.org/10.1002/2014GL059637

Scaife, A. A., Comer, R., Dunstone, N., Fereday, D., Folland, C., Good, E., et al. (2017). Predictability of European winter 2015/2016. *Atmospheric Science Letters*, *18*(2), 38–44. https://doi.org/10.1002/asl.721

Scaife, A., Karpechko, A. Y., Baldwin, M., Brookshaw, A., Butler, A., Eade, R., et al. (2016). Seasonal winter forecasts and the stratosphere. *Atmospheric Science Letters*, *17*(1), 51–56. https://doi.org/10.1002/asl.598

Siegert, S., Stephenson, D. B., & Sansom, P. G. (2016). A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability? *Journal of Climate*, *19*(3), 995–1012. https://doi.org/10.1175/jcli-d-15-0196.1

Sigmond, M., Scinocca, J., Kharin, V., & Shepherd, T. (2013). Enhanced seasonal forecast skill following stratospheric sudden warmings. *Nature Geoscience*, *6*(2), 98–102.

Stephenson, D., Pavan, V., Collins, M., Junge, M., Quadrelli, R., et al. (2006). North Atlantic Oscillation response to transient greenhouse gas forcing and the impact on European winter climate: A CMIP2 multi-model assessment. *Climate Dynamics*, *27*(4), 401–420.

Stockdale, T. (2012). The EUROSIP system—A multi-model approach. https://www.ecmwf.int/sites/default/files/elibrary/2013/12429-eurosip-system-multi-model-approach.pdf

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

Takaya, Y., Yasuda, T., Fujii, Y., Matsumoto, S., Soga, T., Mori, H., et al. (2017). Japan Meteorological Agency/Meteorological Research Institute-Coupled Prediction System version 1 (JMA/MRI-CPS1) for operational seasonal forecasting. *Climate Dynamics*, *48*(1), 313–333. https://doi.org/10.1007/s00382-016-3076-9

Trigo, R. M., Osborn, T. J., Corte-Real, J. M., et al. (2002). The North Atlantic Oscillation influence on Europe: Climate impacts and associated physical mechanisms. *Climate Research*, *20*(1), 9–17.

Voldoire, A., Sanchez-Gomez, E., Salas, D., Mélia, B., Cassou, C., Sénési, S., et al. (2013). Decharme the CNRM-CM5.1 global climate model: Description and basic evaluation. *Climate Dynamics*, *40*(9), 2091–2121. https://doi.org/10.1007/s00382-011-1259-y

Wang, L., Ting, M., & Kushner, P. (2017). A robust empirical seasonal prediction of winter NAO and surface climate. *Scientific Reports*, *7*(1), 279. https://doi.org/10.1038/s41598-017-00353-y

Weisheimer, A., Doblas-Reyes, F. J., Palmer, T. N., Alessandri, A., Arribas, A., Déqué, M., et al. (2009). ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions - skill and progress beyond DEMETER in forecasting Tropical Pacific SSTs. *Geophysical Research Letters*, *36*, L21711. https://doi.org/10.1029/2009GL040896

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D. A., & 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*(703), 917–926. https://doi.org/10.1002/qj.2976