The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation

F. J. Doblas-Reyes, ECMWF, Reading, UK
V. Pavan, ARPA-SMR, Bologna, Italy
D. B. Stephenson, Department of Meteorology, University of Reading, Reading, UK

Submitted to Climate Dynamics, March 2002

Revised March 2003

Accepted June 2003

Abstract

The skill assessment of a set of wintertime North Atlantic Oscillation (NAO) seasonal predictions in a multi-model ensemble framework has been carried out. The multi-model approach consists in merging the ensemble hindcasts of four atmospheric general circulation models forced with observed sea surface temperatures to create a multi-model ensemble. Deterministic (ensemble-mean based) and probabilistic (categorical) NAO hindcasts have been considered. Two different sets of NAO indices have been used to create the hindcasts. A first set is defined as the projection of model anomalies onto the NAO spatial pattern obtained from atmospheric analyses. The second set obtains the NAO indices by standardizing the leading principal component of each single-model ensemble. Positive skill is found with both sets of indices, especially in the case of the multi-model ensemble. In addition, the NAO definition based upon the single-model leading principal component shows a higher skill than the hindcasts obtained using the projection method. Using the former definition, the multi-model ensemble shows statistically significant (at 5% level) positive skill in a variety of probabilistic scoring measures. This is interpreted as a consequence of the projection method being less suitable because of the presence of errors in the spatial NAO patterns of the models. The positive skill of the seasonal NAO found here seems to be due not to the persistence of the long-term (decadal) variability specified in the initial conditions, but rather to a good simulation of the year-to-year variability. Nevertheless, most of the NAO seasonal predictability seems to be due to the correct prediction of particular cases such as the winter of 1989. This behaviour has been explained on the basis of a more reliable description of large-scale tropospheric wave features by the multi-model ensemble, illustrating the potential of multi-model experiments to better identify mechanisms that explain seasonal variability in the atmosphere.

1. Introduction

Early last century, meteorologists noticed that year-to-year fluctuations in wintertime air temperatures in Western Greenland and Northern Europe were often out of phase with one another (Walker, 1924; Loewe, 1937; van Loon and Rogers, 1978; Stephenson et al. 2003). When temperatures are below normal over Greenland, they tend to be above normal in Scandinavia, and vice versa. This climate phenomenon inspired the concept of what was later called the North Atlantic Oscillation (NAO). The NAO is associated with significant changes in the intensity of the westerlies across the North Atlantic onto Europe, and so with a meridional oscillation in atmospheric mass with centers of action near the Icelandic low and the Azores high (e.g., van Loon and Rogers, 1978). During the positive phase the mean westerly flow over the North Atlantic and Western Europe is stronger than usual. The Icelandic Low and the Azores High, also known as the Atlantic dipole (Hastenrath, 2002) are then more intense than normal and tend to be located slightly further north and east (Glowienka-Hensa, 1985; Serreze et al., 1997). This anomalous flow increases the advection of warm and humid air over Northwest Europe. The negative phase of the NAO presents a weakened Atlantic dipole, with weakened westerly flow and increased advection of warm air over Greenland. The NAO is a mode that is robustly present in every month of the year (Barnston and Livezey, 1987). It accounts in a month-by-month basis for the largest amount of interannual variability in monthly North Atlantic sea level pressure in all but four months of the year (Rogers, 1990). However, the NAO is most pronounced in amplitude and areal coverage during winter (December to February) when it accounts for more than one third of the total variance in sea-level pressure (Wallace and Gutzler, 1981; Barnston and Livezey, 1987; Cayan, 1992; Stephenson and Pavan, 2003).

The NAO is linked to a wide range of climatic impacts. The changes in the mean circulation patterns over the North Atlantic are accompanied by pronounced shifts in the storm track (Rogers, 1990; Hurrell, 1995b) and associated synoptic eddy activity. This affects the transport of atmospheric temperature and moisture and produces changes in regional precipitation (Lamb and Peppler, 1987; Cayan, 1992; Hurrell, 1995a). Hurrell (1995a) has shown that drier conditions occur over much of Eastern and Southern Europe and the Mediterranean during high NAO index winters, while wetter-than-normal conditions occur from Iceland through Scandinavia. Winter snow depth and snow coverage duration over the Alps in the early 1990s, when the NAO was persistently positive, have been among the lowest recorded this century (Beniston and Rebetez, 1996), causing economic hardship on those industries dependent on winter snowfall. Other phenomena associated with the NAO include significant wave height (Bacon and Carter, 1993), changes in the Labrador Current transport (Marsh, 2000), in the Arctic sea-ice extent (Fang and Wallace, 1994), in the Davis Strait ice volume (Deser and Blackmon, 1993), in the total ozone column variability over the Northern Hemisphere (Orsolini and Doblas-Reyes, 2003) and in dust transport from Africa across the Mediterranean and the subtropical Atlantic (Moulin et al., 1997).

Atmospheric general circulation models (GCM), forced with both observed temporally varying (Rodwell et al., 1999) and constant (Barnett, 1985; Glowienka-Hensa, 1985; Marshall and Molteni, 1993) sea surface temperature (SST), are able to display NAO-like fluctuations. From those simulations, it would seem that the fundamental mechanisms in the interannual time scale of the NAO arise mostly from atmospheric processes (Hurrell, 1995a). In contrast, its decadal variations might be

slightly influenced by the local ocean (Marshall et al., 2001). Bretherton and Battisti (2000) have pointed out the consequences of forcing atmospheric models with prescribed SSTs in order to study the NAO predictability. Notably, they found an out-of-phase relationship between local surface fluxes and ocean heat content (measured through SST anomalies) at decadal time scales over the North Atlantic that would damp the SST anomalies. Using a coupled atmosphere-ocean simplified model, they detect a robust correlation of 0.4 for the seasonal average NAO, so that in this ideal experiment the seasonal predictability limit is set to less than six months.

A different process wherein atmospheric processes alone might produce strong interannual and perhaps longer-term variations in the intensity of the NAO relies upon the connection between the strength of the stratospheric cyclonic winter vortex and the tropospheric circulation over the North Atlantic (Perlwitz and Graf, 1995; Kodera et al., 1996; Ambaum and Hoskins, 2002). The strong link between the North Pacific basin and the North Atlantic through the Aleutian-Icelandic lows seesaw (Martineu et al., 1999; Honda et al., 2001) might be another source of potential NAO variability. The NAO on interannual time scales appears to be a preferred mode of the atmosphere that can be excited in a number of different ways. For instance, the NAO interannual variability seems to be linked to the large-scale atmospheric circulation (Shabbar et al., 2001) and, to some extent, to tropical (Czaja and Frankignoul, 1999; Hoerling et al., 2001) and extratropical SST (Drévillon et al, 2001) through the modulation of the storm track activity (Peng and Whitaker, 1999; Drevillon et al., 2002).

The nature of climate variability implies that, even if the global SST could be exactly prescribed, the associated NAO evolution would not be uniquely determined in a model given the diversity of strong interactions taking place. The chaotic nature of atmospheric dynamics would amplify any initial uncertainty blurring the NAO predictability. If links between SST anomalies and NAO variability exist, previous studies indicate that they are likely to be weak. Hence, the overall change of the atmospheric state with regard to climatology over the North Atlantic region associated with specified SST anomalies may not be large. Therefore, the amount of predictable signal associated with the boundary conditions will be small compared with the climatological variance (Palmer and Sun, 1985). In practice, an estimate of the atmospheric probability density function (PDF) can be determined from a set of integrations of a model (Epstein, 1969a; Leith, 1974; Brankovic and Palmer, 1997). This led to the concept of ensemble forecasting, whose basic idea is to run not just one deterministic model but to run a model many times with slightly different initial conditions. The set of initial conditions is obtained by introducing perturbations that sample the system uncertainty in the phase space. To tackle uncertainty in the generation of an initial state, multi-analysis forecasts have also been considered in medium-range weather forecasting (Richardson, 2001). Examples of seasonal ensemble integrations have been discussed in Brankovic et al. (1994), Palmer and Anderson (1994), Stern and Miyakoda (1995), Barnett (1995) and Palmer and Shukla (2000).

Initial conditions are not the only source of uncertainty in seasonal forecasting. There are many contributions to the error in a numerical forecast: truncation error, simplification in the model parameterizations, etc. A basic way of dealing with this kind of uncertainty is to use a multi-model approach (Tracton and Kalnay, 1993; Vislocky and Fritsch, 1995; Fritsch et al., 2000; Palmer et al., 2000; Krishnamurti et al., 1999, 2000; Karin and Zwiers, 2002; Stefanova and Krishnamurti, 2002). The

multi-model approach consists of merging forecasts from different models, either with the same initial conditions or not, to develop a wider range of possible outcomes that will allow for a better estimate of the atmospheric PDF. Model combination has already been applied in the development of better standards of reference (Murphy, 1992) or the forecast improvement by independent forecast combination (Thompson, 1977). Using several models in an ensemble is a way of taking into account our uncertainty about the atmospheric laws, since different models make different assumptions showing different performance in their variability simulation. Each model can as well produce an ensemble of simulations. This may be considered as yet another form of ensemble and is referred to as multi-model ensemble (Harrison et al., 1995; Atger, 1999; Doblas-Reyes et al., 2000; Evans et al., 2000, Pavan and Doblas-Reyes, 2000). By merging several single model ensembles into a unique multi-model ensemble, the effect of perturbations in both initial state and model formulation can be included, sampling in this way part of both sources of error. Long-range forecasting is probably one area of fruitful application for model merging, as forecasts from many different models are already available.

Given that large-scale climate features are more predictable than smaller scale anomalies, this study investigates the skill of seasonal NAO hindcasts, as a proxy to deliver seasonal forecasts over the Euro-Atlantic region, in a multi-model framework. In Section 2 we introduce the experiment. Section 3 describes the hindcast accuracy over the European region. The verification of different sets of NAO hindcasts is made in Section 4, and, finally, a discussion of the most important results is drawn along with the main conclusions in Section 5.

2. Experimental design

a. Data

500-hPa geopotential height (Z500) analyses were obtained from the 53-year (1948-2000) NCEP-NCAR reanalyses (Kalnay et al., 1996) as 2.5… horizontal resolution and twice per day data. The calculations were also repeated with the 1979-93 European Centre for Medium-Range Weather Forecasts (ECMWF) reanalyses (Gibson et al., 1997) to check the validity of the results.

b. Model experiment

The multi-model seasonal ensemble hindcasts which have been analysed in this paper, were run as part of the European project PROVOST (PRediction Of climate Variations On Seasonal to interannual Timescales) by four different institutions: the European Centre for Medium-Range Weather Forecasts (ECMWF), the Met Office (MetO), Météo-France (MetFr), and Electricité de France (EDF). The different models and the experiment are fully described in Palmer et al. (2000), Doblas-Reyes et al. (2000), and Pavan and Doblas-Reyes (2000).

The atmospheric models were run for 120 days with 9-member ensembles for the period 1979 to 1993. The only difference between the EDF and MetFr runs is the horizontal resolution (T63 instead of T42). Initialisation was the same for all models. The atmosphere and soil variables were initialised on nine subsequent days from 12 GMT ERA (ECMWF Re-Analyses) analyses (Gibson et al., 1997), starting from the 10th day preceding the first month of the season covered (December, January, February, and March). This method is known as the *lagged average forecast* method

(Hoffman and Kalnay, 1983; Molteni et al., 1988). All hindcasts end on the last day of the fourth month of integration, so that the first integration was 128-day long, while the last one was 120-day long. Daily-observed SST and sea-ice extent were prescribed using either ERA-15 or GISST data, so that there was no interactive ocean model in this experiment, SST being updated daily in the integrations. It is likely that the model skill for the forced experiment can be regarded as an upper bound for the skill of current coupled ocean-atmosphere models (Latif and Barnett, 1996), at least as far as the hypothesis of a negligible feedback from the ocean at the seasonal timescale is accepted. However, Sutton and Mathieu (2003) suggest that ocean heat content anomalies may provide a better representation of the impact of the extratropical ocean on the atmosphere than SSTs.

The model bias, computed as the difference between the long-term climatology of the model and the verification, shows to be of the order of the anomaly being predicted. Some hindcast biases are described in Doblas-Reyes et al. (2000), Brankovic and Palmer (2000), Pavan and Doblas-Reyes (2000), and Pavan et al. (2000a,b). In short, over the Euro-Atlantic region ECMWF, MetFr, and EDF runs present a too strong meridional gradient of Z500 in midlatitudes, producing a strong zonality. MetO shows a more realistic circulation, with a zonal flow weaker than normal over North America and the Western Atlantic. There is also an overall excess of eastward penetration of the Atlantic storm track and a general underestimation of the intraseasonal variability, in particular blocking frequency.

Due to the model biases described above, the raw value of a forecast in long-range forecasting is in principle not useful (Déqué, 1991), so that anomalies have to be computed. Anomalies are expressed as departures from the corresponding long-term climatology. Given the short length of the time series available (14 years), calculation of both model and observed anomalies as well as the forecast verification has been carried out in cross-validation mode (Wilks, 1995). This implies eliminating from the computation the target year. Hindcasts have been verified using seasonal averages for the periods going from months 1 to 3 (December to February, DJF) and 2 to 4 (January to March, JFM), though for brevity the paper has been focused on the second period, less affected by the initial conditions (Pavan and Doblas-Reyes, 2000).

c. Forecast quality

Forecast verification is an important component of the forecast formulation process (Joliffe and Stephenson, 2003). It consists in summarising and assessing the overall forecast quality as a statistical description of how well the forecasts match the verification data. Forecast quality has been evaluated by assessing hindcast skill, including measures of reliability and accuracy. The ensemble mean has been considered as the deterministic hindcast of a continuous variable and time correlation has been used to evaluate its skill. The evaluation of probabilistic forecasts is a more complex task. Three sources of uncertainty in common scoring metrics of probabilistic forecasts are: improper estimates of probabilities from small-sized ensembles, insufficient number of forecast cases, and imperfect reference values due to observation uncertainties. A way to alleviate this problem consists in using several scoring measures. Forecast quality is a multidimensional concept described by several different scalar attributes, which provide useful information about the performance of a forecasting system. Thus, no single measure is sufficient for judging and comparing forecast quality. Consequently, a whole suite of verification measures to assess the probabilistic hindcasts quality has been used here: ranked probability skill score

(*RPSS*), receiver operating characteristic (*ROC*) area under the curve, Peirce skill score (*PSS*), and odds ratio skill score (*ORSS*). All the skill measures used in the paper are briefly described in the Appendix.

3. Model skill over Europe

Given our interest in the Euro-Atlantic region, the skill of the experiment has been assessed over a region extending from 35.N to 87.5.N and from 60.W to 42.5.E. Skill over the area is generally smaller than for the whole Northern Hemisphere, but the wintertime multi-model ensemble results are slightly better than or of the same order of the best single model, as shown by Pavan and Doblas-Reyes (2000). The geographical distribution of the ensemble-mean skill over this area has an uneven distribution. Figure 1 shows the JFM Z500 grid-point correlation for the four single-model ensemble mean hindcasts and the multi-model ensemble mean. In general, two maxima are present over the southwest Atlantic, northern Africa and northern Europe, while the lowest skill is found over Western Europe. The multi-model ensemble mean presents in general the best results. To better illustrate the multi-model ensemble improvement, Figure 2a shows the PDF of the Z500 grid-point correlation over the Euro-Atlantic region. All the models present a PDF biased towards positive values (mean value of 0.28, 0.29, 0.26, 0.26 and 0.33 for ECMWF, MetO, MetFr, EDF and the multi-model ensemble, respectively), though this bias is stronger for the multi-model ensemble.

The probabilistic hindcast skill has been assessed using the *RPSS* for three equiprobable categories. The categories are defined by the terciles of either the hindcast or verification anomalies. Tercile boundaries have been computed in cross-validation mode using two different methods. A simple way of estimating the quantiles of a sample consists in ranking the values and finding the boundaries that create equally populated bins. We will refer to this method as ‚counting„. A more sophisticated method relies upon a Gaussian-kernel estimator of the population PDF (Silverman, 1986) that allows for a high-resolution estimate. In this case, once the PDF has been estimated, the tercile boundaries are computed as the two values distributing the variable in three equiprobable sections. As mentioned above, there is an inherent uncertainty in these estimates that translates into an increased uncertainty in the skill assessment process. However, no *RPSS* differences have been found in the results obtained for each method. The geographical distribution of the *RPSS* displays essentially the same pattern as the ensemble mean skill, with larger areas of negative skill (not shown). However, the improvement provided by the multi-model ensemble is more evident than in the ensemble-mean results. Figure 2b shows the PDF of grid-point *RPSS* over the region for the multi-model and the single-model ensembles. The mean *RPSS* is -10.0, -4.2, -5.8, -.8.3 and 3.4 for ECMWF, MetO, MetFr, EDF and the multi-model ensemble, respectively. The main *RPSS* improvement for the multi-model ensemble consists in a reduction of the very negative values and an important increase of the probability in the range of positive *RPSS*, which explains the increase of the potential value of the multi-model ensemble hindcasts (Palmer et al., 2000). No clear improvement appears in the high *RPSS* range. The low scores of the single models in the probabilistic case may be due to the difficulty in obtaining good estimates of the PDF with a 9-member ensemble. Thus, part of the multi-model ensemble improvement with the probabilistic hindcasts may well be due to the increase in ensemble size.

The poor skill found over the European region on a grid-point basis may

strongly reduce the value of the hindcasts. Pavan and Doblas-Reyes (2000) have suggested that an alternative way of extracting information with significant skill might consist in using large-scale patterns as predictors. This hypothesis is checked in the next section in the case of the NAO. It is nevertheless important to bear in mind that other modes as the Eastern Atlantic or the Scandinavian also have a strong impact on European climate (Qian et al., 2000; Massacand and Davies, 2001; Castro-Díez et al., 2002) and their predictability should also be assessed in future studies.

4. NAO hindcast skill

a. Tropospheric anomalies associated with the NAO

Simple indices of the NAO have been defined as the difference between the normalized monthly sea level pressure at subtropical and subpolar locations (Hurrell, 1995a; Hurrell and van Loon, 1997; Luterbacher et al., 1999). An example of this kind of index is the one defined by Jones[1] (Jones et al., 1997), computed as the difference in sea level pressure at Gibraltar and Stykkisholmur, in Iceland. Positive (negative) values of the index are linked to the positive (negative) phase of the NAO. An example of the corresponding NAO signature for the geopotential height field is shown in Figure 3. These plots have been constructed by averaging Z500 anomalies from the NCEP reanalyses for the three winters (JFM) with the highest (lowest) NAO value in the period of the experiment based upon JonesÂindex. The positive (negative) index years are 1983, 1984, and 1989 (1985, 1987, and 1988). The positive phase pattern presents a negative anomaly over Iceland, eastern Greenland and the Arctic, and a positive one over the central subtropical Atlantic and Western Europe. This is the kind of pattern that we expect the models to simulate. It is associated with a cold anomaly over Greenland and North Africa and a warm anomaly over the extratropical North Atlantic and Europe (not shown). The negative phase shows a similar pattern with reversed sign. As for the precipitation signature (not shown), the NAO positive phase shows a positive anomaly over Iceland and Scandinavia and a negative one over the Iberian Peninsula and western Mediterranean, eastern Greenland and eastern subtropical Atlantic (Hurrell, 1995a; Hurrell and van Loon, 1997). This corresponds to an increase of storm track activity (not shown) over northern Europe (Serreze et al., 1997; Rodwell et al., 1999). The negative phase shows an increase of cyclone activity over the central North Atlantic leading towards the Bay of Biscay.

b. NAO indices: reference and hindcasts

A simple NAO index for the hindcasts is defined here in a similar way to Pavan and Doblas-Reyes (2000). An empirical orthogonal function (EOF) analysis has been carried out using the December to March monthly Z500 NCEP reanalysis data from 1948 to 2000 over the region 87.5°N-20°N and 90°W-60°E. The seasonal cycle and the long-term mean have been previously removed to create monthly anomalies, which have been weighted by the cosine of the latitude. The first EOF, shown in Figure 4, explains 28.6% of the variance. Ambaum et al. (2001) have discussed the physical consistency of defining the NAO based on regional EOF analysis and recommended this regional approach. The principal component of the leading EOF (PC1 henceforth) has been used as a surrogate for the NAO index, along with JonesÂindex. The

[1] These data are available at http://www.cru.uea.ac.uk/cru/data/nao.htm

correlation of PC1 with JonesÂindex for JFM is 0.93 (0.91 for DJF). The slope of the linear regression of PC1 against JonesÂindex is close to 1 (0.94 and 0.97 for DJF and JFM, respectively). This indicates that the results presented hereafter may depend on the verification index used, though not strongly.

A first example of hindcast NAO index consists in the projection of the monthly grid point anomalies for each model and ensemble member onto the NCEP leading EOF described above. This method is referred to as Pobs in the following. The resulting covariances are then seasonally averaged. The set of ensemble hindcasts are displayed using open dots in Figure 5a. Each dot corresponds to the JFM hindcast of a member of a single-model ensemble for a given year. Since the interannual variance of single-model anomalies is generally underestimated, single-model hindcasts have been standardized in cross-validation mode using a separate estimate of the standard deviation for each model. The verification time series have also been standardized. The multi-model ensemble hindcast is built up as the ensemble of all the single-model ensemble hindcasts. The 2-4 month multi-model ensemble mean (solid dots) has a correlation with PC1 of 0.46, and 0.33 with JonesÂindex, both not statistically significant with 95% confidence based on 14 degrees of freedom (correlation would be statistically significant at 5% level if larger than 0.50). Single-model ensemble-mean skill is similar to or lower than for the multi-model ensemble (not shown). Pavan and Doblas-Reyes (2000) have shown that an increase in correlation up to statistically significant values (0.55) may be obtained if a linear combination of each model ensemble mean is taken instead of pooling the ensemble-mean hindcasts using equal weights.

A different set of NAO ensemble hindcast indices has been defined using the first principal component of an EOF analysis performed on each model. This method will be referred to as Pmod henceforth. The corresponding spatial patterns obtained in the EOF analysis are shown in Figure 4. The first EOF explains 25.5%, 24.3%, 29.3%, and 30.4% of the total variance for ECMWF, MetO, MetFr and EDF, respectively. They present a spatial distribution similar to the NCEP NAO pattern, the pattern for MetO being the most realistic, though some spatial shifting can be noted. The spatial correlation of the single-model leading EOF with the corresponding NCEP EOF is 0.87, 0.99, 0.86 and 0.78 for ECMWF, MetO, MetFr and EDF, respectively. The use of single-model principal components as NAO hindcasts has the advantage of taking into account the spatial biases in the NAO patterns in the different models. The spatial error, illustrated in Figure 4, can reduce the NAO signal estimated when using projections of model anomaly. The NAO hindcasts were standardized as described above and the multi-model constructed in the same way. This approach corresponds to using a Mahalanobis metric, which has some good invariance properties (Stephenson, 1997), to assess the model ability to simulate the NAO. The corresponding multi-model ensemble-mean hindcasts turn out to be very similar to the ones obtained by projecting the model anomalies (Figure 5b). However, correlation with the verification time series is now higher (the same result applies for the seasonal hindcasts for 1-3 month hindcasts) rising up to 0.57 (PC1) and 0.49 (Jones). These values are already statistically significant at 95% confidence. Other measures of error, as the root mean square error or the mean absolute deviation, are also reduced. This implies an improvement in NAO skill with regard to that obtained with the Pobs method. Additionally, the multi-model ensemble spread does not change when considering either anomaly projections or single-model principal components (not shown).

Two additional NAO hindcast indices have been tested. The corresponding

results will be discussed very briefly. In the first one, the geopotential anomalies of the verification and the individual ensemble members have been averaged over pre-defined regions and their differences computed, following Stephenson et al. (2000). The boundaries of the two areas are (90°W-22°E, 55°N-33°N) and (90°W-22°E, 80°N-58°N) for the southern and northern boxes, respectively. These boundaries have been chosen on the basis of the correlation between the DJFM-mean JonesÂindex and the Z500 NCEP reanalyses from 1959 to 1998[2]. An areal average was chosen instead of a simple difference between two grid points because it avoids some of the subjectivity inherent to the selection of the reference grid points. The results are quite similar to those discussed above. The multi-model ensemble-mean skill is 0.37 using PC1 as verification. Secondly, an NAO temperature index has been defined based on the temperature seesaw over Europe and Greenland (Loewe, 1937; van Loon and Rogers, 1978; Stephenson et al., 2000). When winters in Europe are unusually cold and those in west Greenland are mild (Greenland above mode), the Icelandic Low is generally weak and located around the southern tip of Greenland. In the opposite mode, when Europe is mild and west Greenland is cold (Greenland below mode), the Atlantic westerlies are strong, the Icelandic Low is deep, and a strong maritime flow extends into Europe (Hurrell and van Loon, 1997; Serreze et al., 1997). The areas selected are (90°W-0°, 72°N-50°N) and (0°-90°E, 72°N-50°N). As expected, a strong anticorrelation between this temperature and the geopotential indices described above is found. A higher correlation of the multi-model ensemble-mean hindcasts (0.47) is obtained with the temperature index, but this might be just due to the prescription of observed SSTs in the experiment. The skill of the two areal-average indices confirms that the positive multi-model ensemble-mean correlation is a robust feature. In the rest of this paper only results from the more successful Pobs and Pmod methods will be discussed.

Some statistical properties of the NAO ensemble hindcasts have also been analysed. Skewness is a measure of the asymmetry of a distribution about its mean. Distributions with positive and negative skewness represent asymmetric distributions with a larger tail to the right or left respectively. Positive kurtosis indicates a relatively peaked distribution with long tails (leptokurtic). Negative kurtosis indicates a relatively flat distribution with short tails (platykurtic). Measures of skewness for the 2-4 month NAO hindcasts are negative for both the multi-model ensemble and some of the single models. This is because more negative than positive NAO hindcasts are found in the ensemble. Nevertheless, this is not the case for the 1-3 month hindcasts. Instead, the hindcast time series present a negative kurtosis at both lead times and for all the models. This platykurtic behaviour indicates that the tails of the hindcast distribution present a low probability. Finally, Figure 5 gives hints of the ensemble distribution being skewed to values with the same sign as the observed anomaly when the multi-model ensemble mean is far from zero. This can be interpreted as an indication of predictive skill, especially for the years 1988 and 1989, although longer samples are needed to extract more definite conclusions.

c. Probabilistic NAO hindcasts

Deterministic predictions based on only the ensemble mean do not include all the information provided by the individual members within the ensemble. Instead, it

[2] Correlations are available at http://www.cdc.noaa.gov/Correlation/

is more useful to provide hindcasts for given categories in terms of probability forecasts. The skill scores described in the appendix have been used to assess the skill of these hindcasts. Table 1 summarizes the results for the hindcasts obtained from the single-model principal components with JonesÂindex as verification. Similar results have been found for the different sets of hindcasts and verification data available.

Three events have been considered in this paper: anomalies above the upper tercile, above the mean, and below the lower tercile. The hindcast probability bias was in the range [0.8, 1.2] for the three categories, which for the short length of the sample corresponds to low-biased hindcasts. This indicates that a simple bias correction by standardizing the values provides quite reliable hindcasts. Nevertheless, longer time series would allow for a systematic correction of the conditional biases.

RPSS is an appropriate measure of the quality of multiple category forecasts as it takes into account the way the probability distribution shifts toward the extremes within a particular category. The *RPSS* for NAO hindcasts is very low, as shown in Table 1. This should be expected for an event with a low signal-to-noise ratio (Kumar et al., 2001), as in the case of the seasonal NAO. However, the values tend to be positive, indicating that the ensemble hindcasts provide slightly better estimates of the tercile probabilities than climatology. The multi-model ensemble shows the highest skill score. More interestingly, the *RPSS* is generally not significantly different from zero for the single models, but it turns out to be statistically significant at 5% level for the multi-model ensemble, regardless of the verification data used (not shown).

An assessment of hindcast quality for binary events has also been undertaken. Several events had to be considered because the measures of accuracy of binary events do not take into account the severity of errors across categories (Joliffe and Stephenson, 2003). For instance, if category one were observed, calculations of the false alarm rate would not discriminate if either tercile two or three were forecast, the second case being less desirable. Besides, the estimation of the scores fore various events allows for an evaluation of the robustness of hindcast quality. The values of the *ROC* area under the curve are in most of the cases above the no-skill value of 0.5. The multi-model ensemble does not always have the highest score, single models showing a higher value for some events. However, the multi-model skill is similar for the different events taken into account, which is not the case for the single models. The homogeneous *ROC* area values for the multi-model ensemble might partly be a consequence of the *ROC* score being almost invariant with the set of probability thresholds (Stephenson, 2000). Thus, the *ROC* area shows that, as in the case of the ensemble-mean correlation, there is a consistent positive skill in the NAO multi-model hindcasts, though it does not tend to be statistically significant at the 5% level (it appears to be the case only for the upper tercile event). Similar conclusions are drawn for the other skill measures. Table 1 also shows the results for *PSS*, *OR*, and *ORSS*. The multi-model ensemble displays again the best results. Although the *PSS* is a measure that could be affected by the hindcast bias, it shows statistically significant skill in the same cases as the other measures do, proving that the NAO hindcasts are not only accurate, but also reliable. The similarity between *ROC* area and *OR* values can be explained through the parameterisation of the *ROC* curve described in Stephenson (2000). As a general rule, *ORSS* seems to be the most stringent skill score. *ORSS* is independent of the marginal distributions, so that it is able to strongly discriminate the cases with and without association between hindcasts and observations. It is important to note that the skill for the event , above

the mean„ seems to be always quite low. This might be due to the lack of robustness of the estimated mean as a consequence of the short sample used (Kharin and Zwiers, 2002).

5. Concluding remarks

Given the low seasonal hindcast skill at grid-point scale over the Euro-Atlantic region, a means of extracting greater predictability by reference to larger scale features should prove to be useful. This paper suggests that predictions of the NAO may provide an alternative to relying on GCM direct output. A comprehensive assessment of the NAO seasonal hindcast skill has been carried out. This approach to the assessment of predictive skill presents advantages over the analysis of the predictability of a few case studies (Dong et al., 2000; Elliott et al., 2001), although the analysis of specific cases allows for the identification of sources of predictability. The skill evaluation has been done using a multi-model framework. The multi-model approach used here consisted in merging a set of ensemble hindcasts from four atmospheric models.

Both deterministic (ensemble mean) and probabilistic (categorical) hindcasts have been considered and evidence of the multi-model ensemble skill being superior to that of the single model ensembles has been presented. It has been shown that the NAO multi-model ensemble-mean hindcasts may have significant skill (at the 5% level) when the NAO indices are defined as the standardized leading principal component of the single-model ensembles (Pmod method). The skill for probabilistic hindcasts of the NAO indices falling in the outer tercile categories and also for the ‚above-the-mean„ event has been computed. Because different verification scores measure different aspects of the forecasting system, a set of probabilistic skill measures has been used to estimate probabilistic forecast quality in this paper. A consistent positive skill to forecasting probabilistically NAO terciles has been found. A strong agreement has been observed in the results obtained using two independent verification sets: the sea level pressure NAO index defined by Jones and the principal component of the 500-hPa geopotential height leading EOF computed with NCEP monthly-mean data for the period 1948-2000. In addition, the two methods described above to compute the tercile boundaries (counting and Gaussian-kernel PDF estimate) also presented similar results (not shown). As for the ensemble mean, an overall degradation of the probabilistic forecast quality has been observed for the NAO hindcasts computed as projections of monthly anomalies of the individual ensemble members onto the leading EOF of the 500-hPa geopotential height monthly mean NCEP analyses (Pobs method) when compared to those of the Pmod method. This may be due to the model anomaly projection method being less suitable because of the model systematic error leading to spatial shifts of the simulated NAO patterns, which would generate different values of the index for a similar type of signal in each model.

Given the shortness of the sample, some of the skill might be due to decadal variability in the initial conditions or from the strong predictability of particular winters. A simple way of removing artificial skill due to long-term trends is to verify year-to-year differences in the time series. Differences remove the low-frequency variability from the time series, so that the skill in year-to-year changes can be assessed (Stephenson et al., 2000). The correlation of the backward differences for the ensemble-mean hindcast based on the standardized first principal component is 0.19, 0.62, -0.01, 0.33 and 0.35 for ECMWF, MetO, MetFr, EDF and the multi-model

ensemble, respectively. Correlations are much higher for the 1-3 month hindcasts. Thus, despite these low correlations, the NAO skill in these experiments is partly due to the correct simulation of year-to-year variations.

The contribution to the positive skill from specific years is an important issue because of the small sample used in this study. The case of JFM 1989 is particularly interesting. The multi-model ensemble shows an extraordinarily good forecast for this winter. When this year is removed from the time series, the skill is substantially reduced (the multi-model ensemble-mean correlation drops to 0.12 from values close to 0.5 and similar reductions in correlation are found for the single models). The probabilistic score measures seem to be less affected though. For instance, *RPSS* takes the value 7.1 (compared with 13.1 in Table 1) and is marginally significant at the 5% level, whilst the odds ratio takes values around 1.4, which are not significant. Scores are quite insensitive to the removal of other years. For instance, the correlation is 0.46 when the year 1985 is not taken into account. Thus, a substantial part of the skill presented in this paper comes from the correct simulation of the atmospheric circulation over the North Atlantic in JFM 1989. It is then important to try to understand the reasons why some years are so well forecast while others are not.

One of the possible dynamical reasons for 1989 anomalies being correctly predicted may be depicted using an estimate of the barotropic refraction index for the 1985 and 1989 JFM hindcasts as done in Pavan et al. (2000). These two years have an opposite sign NAO (Figure 5), which is mainly due to the presence of positive (negative) geopotential anomalies (not shown) over the subtropical Atlantic in 1989 (1985). The multi-model ensemble mean displays the right pattern north of 45.N in both cases. Nevertheless, anomalies are misplaced in 1985 over the region south of 45.N, which is not the case in 1989 (not shown). This explains the bad NAO prediction for the former year (Figure 5), with the multi-model ensemble evenly distributed around zero, and the highly satisfactory 1989 hindcasts, with a positively skewed multi-model ensemble. Figure 6 shows the barotropic refraction index for the verification and the multi-model ensemble for both years. This index, which gives an indication of the propagation of large-scale waves in the extratropics, corresponds to the critical wavenumber separating the meridionally confined waves from those with a propagating structure profile. That is, the minima of the function give an indication of possible meridional confinement of the waves. The refraction index for the verification (solid line) has a very similar behaviour north of 45.N in both years. This feature agrees well with the strong resemblance of the anomaly patterns over that sector. However, a clear confinement of the waves with wavenumber greater than 3 in the latitudinal range 35.N-45.N is evidenced in the analyses for 1989, though not for 1985. This implies that all sort of large-scale waves can propagate into the subtropical Atlantic in 1985, the confinement of waves with wavenumber greater than 2 being found just north of 60.N. Figure 6b presents evidences of a substantial number of members of the multi-model-ensemble correctly showing some sort of confinement south of 45.N in 1989 (31 out of 36 members have a local minimum in this region; they are depicted using dashed lines). Nevertheless, an important latitudinal spread of the minima is found, so that the ensemble mean of the index may give the misleading impression of confinement not taking place. In the other hand, a substantial amount of ensemble members (26 out of 36, depicted using dashed lines) seem to unrealistically confine large-scale waves in a wide range of latitudes south of 45.N in 1985 (Figure 6a). This emphasises the importance of correctly simulating the

structure of large-scale waves, in order to produce skilful NAO hindcasts. In addition, this diagnostic only makes sense if an ensemble is used; in other words, a single-member simulation or the use of the ensemble mean would not have led to the same conclusions. Furthermore, the use of a multi-model ensemble also takes partly into account the different systematic errors of the single models when simulating the processes dynamically related to the NAO. As a consequence, the probabilistic formulation of multi-model seasonal NAO hindcasts may be able to make a better use of all this information than a deterministic formulation based on a single model.

Although there is some evidence of NAO skill in the hindcasts presented here, the skill is quite small. More research should be carried out to better understand the physical reasons for the positive skill found. The high agreement among the different accuracy and skill measures for the multi-model seems to be encouraging and deserves further investigation with improved models and larger samples. The results presented in this paper strengthen the prospects and expected utility of the present-day state-of-the-art seasonal forecast systems. It is also interesting to emphasize that the methodology described here may provide even better results when applied to other large-scale phenomena, either over the North Atlantic region (given the limited amount of variability explained by the NAO) or elsewhere. At present, the method is being used to assess the skill of the multi-model ensemble hindcasts carried out in the framework of the DEMETER project[3] (Palmer et al., 2003) and in the operational seasonal forecasts at ECMWF.

Acknowledgements

This study was undertaken when the first author worked at the Centre Nationale de Recherches Météorologiques, Météo-France (Toulouse, France). VP has received support from the Progetto Strategico SINAPSI funded by the Ministero dell'Istruzione, dell'Universita' e della Ricerca (MIUR) and Consiglio Nazionale di Ricerca (CNR). The authors wish to thank David Anderson, Magdalena Balmaseda, Michel Déqué, Thomas Jung, Alexia Massacand, Laura Ferranti, and Tim Palmer for reviews of early drafts and constructive advice. Richard Graham and an anonymous reviewer are especially acknowledged for their significant contribution to the improvement of the scientific quality and readability of the paper. This work was in part supported by the EU-funded DEMETER project (EVK2-1999-00197).

[3] DEMETER is a EU-funded project intending to assess the hindcast skill and potential value of multi-model ensemble-based system for seasonal-to-interannual prediction, including innovative examples of the application of multi-model seasonal ensemble information in malaria incidence and crop yield forecasting. Full information about the project can be found at http://www.ecmwf.int/research/demeter.

Appendix: Scoring rules

A tool commonly used to evaluate the association between ensemble-mean hindcasts and verification is the time correlation coefficient. This measure is independent of the mean and variance of both variables. As in the rest of the paper, different climatologies for hindcasts and verification were computed using the cross-validation technique, making the correlation estimator unbiased (Déqué, 1997).

A set of verification measures has been used to assess the quality of the probabilistic hindcasts: the ranked probability skill score (*RPSS*), the receiver operating characteristic (*ROC*) area under the curve, the Peirce skill score (*PSS*), and the odds ratio skill score (*ORSS*). Most of them, along with estimates of the associated error, are described in Stephenson (2000), Zhang and Casey (2000), and Thornes and Stephenson (2001), where the reader is referred to for more specific definitions and properties.

The accuracy measure for *RPSS* is the ranked probability score (*RPS*). *RPS* was first proposed by Epstein (1969b) and simplified by Murphy (1971). This score for categorical probabilistic forecasts is a generalisation of the Brier score for ranked categories. For J ranked categories, the *RPS* can be written:

$$RPS(\mathbf{r},\mathbf{d}) = \frac{1}{J-1}\sum_{i=1}^{J}\left(\sum_{k=1}^{i} r_k - \sum_{k=1}^{i} d_k\right)^2 \qquad \text{(A.1)}$$

where the vector $\mathbf{r}=(r_1,...,r_J)$ ($\sum_{k=1}^{J} r_k = 1$) represents an estimate of the forecast PDF and $\mathbf{d}=(d_1,...,d_J)$ corresponds to the verification PDF where d_k is a delta function which equals to 1 if category k occurs and 0 otherwise. By using cumulative probabilities, it takes into account the ordering of the categories, though for finite ensemble sizes, the estimated probabilities for the event to be in different categories strongly depend on the estimate of the category thresholds. *RPS* can be accumulated for several time steps or grid points over a region, or both. The *RPSS* expresses the relative improvement of the forecast against a reference score. The reference score used in this paper has been the climatological probability hindcast, which, under the assumption of a Gaussian distribution of the observations, is the forecast without any skill that minimises the *RPS* (Déqué et al., 1994). The *RPSS* is defined as:

$$RPSS = 100\left(1 - \frac{RPS_{\text{forecast}}}{RPS_{\text{climatol}}}\right) \qquad \text{(A.2)}$$

Such skill score is 100 for a perfect forecast, 0 for a probabilistic forecast which is no more accurate than a trivial forecast using long-term climatology, and negative for even worse forecasts, as random or biased values. To provide an estimate of the skill score significance, the calculations were repeated 100 times for a given time series (either a grid point or the NAO index). Each time, the order of the individual hindcasts was scrambled (this preserves the PDF of the variable), then computing the skill score, and finally taking the 5% upper threshold of the resulting skill distribution.

RPSS can be a too stringent measure of skill by requiring a correct estimate of a simplified PDF. Then, a set of simple accuracy measures for binary events is made based upon the hit rate H, or the relative number of times an event was forecast when it occurred, and the false alarm rate F, or the relative number of times the event was forecast when it did not occur (Jolliffe and Stephenson, 2003). It is based on the

likelihood-base rate factorisation of the joint probability distribution of forecasts and verifications (Murphy and Winkler, 1987). To derive it, a contingency table is computed, wherein the cells are occupied by the number of hits (*a*, number of cases when an event is forecast and is also observed), false alarms (*b*, number of cases the event is not observed but is forecast), misses (*c*, number of cases the event is observed but not forecast), and correct rejections (*d*, number of no-events correctly forecast) for every ensemble member. Then, the hit rate and the false alarm rate take the form:

$$H = \frac{a}{a+c} \quad F = \frac{b}{b+d} \tag{A.3}$$

The previous scheme allows for the definition of a reliability measure, the bias *B*. Reliability is another attribute of forecast quality and corresponds to the ability of the forecast system to average probabilities equal to the frequency of the observed event. The bias indicates whether the forecasts of an event are being issued at a higher rate than the frequency of observed events. It reads:

$$B = \frac{a+b}{a+c} \tag{A.4}$$

A bias greater than 1 indicates over-forecasting, i.e., the model forecasts the event more often than it is observed. Consequently, a bias lower than 1 indicates under-forecasting.

The Peirce skill score (*PSS*) is a simple measure of skill that equals to the difference between the hit rate and the false alarm rate:

$$PSS = H - F \tag{A.5}$$

When the score is greater than zero, the hit rate exceeds the false alarm rate so that the closer the value of *PSS* to 1, the better. The standard error formula for this score assumes independence of hit and false alarm rates and, for large enough samples, it is computed as:

$$\sigma_{PSS} = \sqrt{\frac{H(1-H)}{a+c} + \frac{F(1-F)}{b+d}} \tag{A.6}$$

The odds ratio (*OR*) is an accuracy measure that compares the odds of making a good forecast (a hit) to the odds of making a bad forecast (a false alarm):

$$OR = \frac{H}{1-H}\frac{1-F}{F} \tag{A.7}$$

The ratio is greater than one when the hit rate exceeds the false alarm rate, and is unity when forecast and reference values are independent. It presents the advantage of being independent of the forecast bias. Furthermore, it has the property that the natural logarithm of the odds ratio is asymptotically normally distributed with a standard error of $1/(n_h)^{1/2}$, where

$$\frac{1}{n_h} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \tag{A.8}$$

To test whether there is any skill, one can test against the null hypothesis that the forecasts and verifications are independent with log odds of zero. A simple skill score, the odds ratio skill score (*ORSS*), ranging from ‡1 to +1, where a score of zero represents no skill, may be obtained from the odds ratio through the expression:

$$ORRS = \frac{OR - 1}{OR + 1} = \frac{H - F}{H + F - 2HF} \tag{A.9}$$

Thornes and Stephenson (2001) provide a useful table with the minimum values of *ORSS* needed to have significant skill at different levels of confidence depending on the value of n_h.

The *ROC* (Swets, 1973) is a signal-detection curve plotting the hit rate against the false alarm rate for a specific event over a range of probability decision thresholds (Evans et al., 2000; Graham et al., 2000; Zhang and Casey, 2000). Basically, it indicates the performance in terms of hit and false alarm rate stratified by the verification. The probability of detection is a probability decision threshold that converts probabilistic binary forecasts into deterministic binary forecasts. For each probability threshold, a contingency table is obtained from which the hit and false alarm rates are computed. For instance, consider a probability threshold of 10%. The event is forecasted in those cases where the probability is equal to or greater than 10%. This calculation is repeated for thresholds of 20%, 30%, up to 100% (or whatever other selection of intervals, depending mainly on the ensemble size). Then, the hit rate is plotted against the false alarm rate to produce a *ROC* curve. Ideally, the hit rate will always exceed the false alarm rate and the curve will lie in the upper-left-hand portion of the diagram. The hit rate increases by reducing the probability threshold, but at the same time the false alarm rate is also increased. The standardized area enclosed beneath the curve is a simple accuracy measure associated with the *ROC*, with a range from 0 to 1. A system with no skill (made by either random or constant forecasts) will achieve hits at the same rate as false alarms and so its curve will lie along the 45° line and enclose a standardized area of 0.5. As the *ROC* is based upon a stratification by the verification it provides no information about reliability of the forecasts, and hence the curves cannot be improved by improving the climatology of the system. The skill score significance was assessed, as in the case of *RPSS*, by Monte Carlo methods.

References

Ambaum MHP, Hoskins BJ, Stephenson DB (2001) Artic Oscillation or North Atlantic Oscillation? J Clim 14: 3495-3507

Ambaum MHP, Hoskins BJ (2002) The NAO troposphere-stratosphere connection. J Clim 15: 1969-1978

Atger F (1999) The skill of ensemble prediction systems. Mon Weather Rev 127: 1941-1953

Bacon S, Carter DJT (1993) A connection between mean wave height and atmospheric pressure gradient in the North Atlantic. Int J Climatol 13: 423-436

Barnett TP (1985) Variations in near-global sea level pressure. J Atmos Sci 42: 478-501

Barnett TP (1995) Monte Carlo climate forecasting. J. Climate 8: 1005-1022

Barnston AG, Livezey RE (1987) Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. Mon Weather Rev 115: 1083-1126

Beniston M, Rebetez M (1996) Regional behavior of minimum temperatures in Switzerland for the period 1979-1993. Theor Appl Climatol 53: 231-243

Brankovic C, Palmer TN (1997) Atmospheric seasonal predictability and estimates of ensemble size. Mon Weather Rev 125: 859-874

Brankovic C, Palmer TN (2000) Seasonal skill and predictability of ECMWF PROVOST ensembles. Quart J Roy Meteorol Soc 126: 2035-2068

Brankovic C, Palmer TN, Ferranti L (1994) Predictability of seasonal atmospheric variations. J Clim 7: 217-237

Bretherton CS, Battisti DS (2000) An interpretation of the results from atmospheric general circulation models forced by the time history of the observed sea surface temperature distributions. Geophys Res Lett 27: 767-770

Castro-Díez Y, Pozo-Vázquez D, Rodrigo FS, Esteban-Parra MJ (2002) NAO and winter temperature variability in southern Europe. Geophys Res Lett 29 8: 1 1-4

Cayan DR (1992) Latent and sensible heat flux anomalies over the northern oceans: the connection to monthly atmospheric circulation. J Clim 5: 354-369

Czaja A, Frankignoul D (1999) Influence of the North Atlantic SST on the atmospheric circulation. Geophys Res Lett 26: 2969-2972

Déqué M (1991) Removing the model systematic error in extended range forecasting. Ann Geophys 9: 242-251

Déqué M (1997) Ensemble size for numerical seasonal forecasts. Tellus 49A: 74-86

Déqué M, Royer JF, Stroe R (1994) Formulation of gaussian probability forecasts based on model extended-range integrations. Tellus 46A: 52-65

Deser C, Blackmon ML (1993) Surface climate variations over the North Atlantic Ocean during winter: 1900-1989. J Clim 6: 1743-1753

Doblas-Reyes FJ, Déqué M, Piedelièvre JP (2000) Multi-model spread and probabilistic seasonal forecasts in PROVOST. Quart J Roy Meteorol Soc 126: 2069-2088

Dong BW, Sutton RT, Jewson SP, O'Neill A, Slingo JM (2000) Predictable winter climate in the North Atlantic sector during the 1997-1999 ENSO cycle. Geophys Res Lett 27: 985-988

Drévillon M, Terray L, Rogel P, Cassou C (2001) Mid latitude Atlantic SST influence on European winter climate variability in the NCEP reanalysis. Clim Dyn 18: 331-344

Drévillon M, C Cassou, Terray L (2002) Model study of the wintertime atmospheric response to fall tropical Atlantic sea-surface-temperature anomalies. Submitted to Quart J Roy Meteorol Soc

Elliott JR, Jewson SP, Sutton RT (2001) The impact of the 1997/98 El Niño event on the Atlantic Ocean. J Clim 14: 1069-1077

Epstein ES (1969a) Stochastic dynamic prediction. Tellus 21: 739-759

Epstein ES (1969b) A scoring system for probability forecasts of ranked categories. J Appl Meteor 8: 985-987

Evans RE, Harrison MSJ, Graham RJ, Mylne KR (2000) Joint medium-range ensembles from the Met. Office and ECMWF systems. Mon Weather Rev 128: 3104-3127

Fang Z, Wallace JM (1994) Arctic sea-ice variability on a time-scale of weeks and its relation to atmospheric forcing. J Clim 7: 1897-1914

Fritsch JM, Hilliker J, Ross J, Vislocky RL (2000) Model consensus. Wea Forecasting 15: 571-582

Gibson JK, Kallberg P, Uppala S, Hernandez A, Nomura A, Serrano E (1997) ERA description. ECMWF re-analysis project report series 1. ECMWF Tech. Report, 872pp.

Glowienka-Hensa R (1985) Studies on the variability of the Icelandic Low and Azores High between 1881 and 1983. Contrib Atmos Phys 58: 160-170

Graham RJ, Evans ADL, Mylne KR, Harrison MSJ, Robertson KB (2000). An assessment of seasonal predictability using atmospheric general circulation models. Quart J Roy Meteorol Soc 126: 2211-2240

Harrison MSJ, Palmer TN, Richardson DS, Buizza R, Petroliagis T (1995) Joint ensembles from the UKMO and ECMWF models. In Proceedings of the Seminar Predictability, ECMWF, Reading, UK, 2: 61-120.

Hastenrath S (2002) Dipoles, temperature gradients, and tropical climate anomalies. Bull Amer Meteorol Soc 83: 735-738

Hoerling MP, Hurrell JW, Xu T (2001) Tropical origins for recent North Atlantic climate change. Science.292: 90-92.

Hoffman NR, Kalnay E (1983) Lagged average forecasting, an alternative to Monte Carlo forecasting. Tellus 35A: 100-118

Honda M, Nakamura H, Ukita J, Kousaka I, Takeuchi K (2001) Interannual seesaw between the Aleutian and Icelandic lows. Part I: Seasonal dependence and life cycles. J Clim 13: 1029-1042

Hurrell JW (1995a) Decadal trends in the North Atlantic Oscillation regional temperatures and precipitation. Science 269: 676-679

Hurrell JW (1995b) Transient eddy forcing of the rotational flow during northern winter. J Atmos Sci 52: 2286-2301

Hurrell JW, van Loon H (1997) Decadal variations in climate associated with the North Atlantic oscillation. Clim Change 36: 301-326

Jolliffe IT, Stephenson DB, eds (2003) Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley and Sons, 240 pp

Jones PD, Jónsson T, Wheeler D (1997) Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and South-West Iceland. Int J Climatol 17: 1433-1450

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds B, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year reanalysis project. Bull Amer Meteorol Soc 77: 437-472

Kharin VV, Zwiers FW (2002) Climate prediction with multimodel ensembles. J Clim 15: 793-799

Kodera K, Chiba M, Koide H, Kitoh A, Nikaido Y (1996) Interannual variability of winter stratosphere and troposphere in the Northern Hemisphere. J Meteorol Soc Japan 74: 365-382

Krishnamurti TN, Kishtawai CM, LaRow T, Bachiochi DR, Zhang Z, Williford CE, Gadgil S, Surendran S (1999) Improved weather and seasonal climate forecasts from multimodel superensemble. Science 285: 1548-1550

Krishnamurti TN, Kishtawai CM, Zhang Z, LaRow T, Bachiochi D, Williford CE (2000) Multimodel ensemble forecasts for weather and seasonal climate. J Clim13: 4196-4216

Kumar A, Barnston AG, Hoerling MP (2001) Seasonal prediction, probabilistic verifications, and ensemble size. J Clim 14: 1671-1676

Lamb PJ, Peppler RA (1987) North Atlantic Oscillation: Concept and an application. Bull Amer Meteorol Soc 68: 1218-1225

Latif M, Barnett TP (1996) Decadal variability over the North Pacific and North America: dynamics and predictability. J Clim 9: 2407-2423

Leith CE (1974) Theoretical skill of Monte Carlo forecasts. Mon Weather Rev 102: 409-418

Loewe F (1937) A period of warm winters in western Greenland and the temperature see-saw between western Greenland and Europe. Quart J Roy Meteorol Soc 63: 365-371

Luterbacher J, Schmutz C, Gyalistras D, Xoplaki E, H Wanner (1999) Reconstruction of monthly NAO and EU indices back to AD 1675. Geophys Res Lett 26: 2745-2748

Marsh R (2000) Recent variability of the North Atlantic thermohaline circulation inferred from surface heat and freshwater fluxes. J Clim 13: 3239-3260

Marshall JC, Molteni F (1993) Towards a dynamical understanding of weather regimes J Atmos Sci 50:

1792-1818

Marshall JC, Johnson H, Goodman J (2001) A study of the interaction of the North Atlantic Oscillation with ocean circulation. J Clim 14: 1399-1421

Martineu C, Caneill JY, Sadourny R (1999) Potential predictability of European winters from the analysis of seasonal simulations with an AGCM. J Clim 12: 3033-3061

Massacand AC, Davies HU (2001) Interannual variability of European winter weather: The potential vorticity insight. Atmos Sci Lett doi:10.1006/asle.2001.0026, http://www.idealibrary.com/links/toc/asle/0/0/0.

Molteni F, Cubash U, Tibaldi S (1988) 30- and 60-day forecast experiments with the ECMWF spectral models. In Persistent meteo-oceanographic anomalies and teleconnections. Eds. C. Chagas and G. Puppi, vol 69, Pontificae Academiae Scientiarum Scripta Varia, Vatican City: 505-555

Moulin C, Lambert CE, Dulac F, Dayan U (1997) Atmospheric export of dust from North Africa: Control by the North Atlantic Oscillation. Nature 387: 691-694

Murphy AH (1971) A note on the ranked probability score. J Appl Meteor 10: 155-156

Murphy AH (1992) Climatology, persistence, and their linear combination as standards of reference in skill scores. Wea Forecasting 7: 692-698

Murphy AH, Winkler RL (1987) A general framework for forecast verification. Mon Weather Rev 115: 1330-1338

Orsolini Y, Doblas-Reyes FJ (2003) Ozone signatures of climate patterns over the Euro-Atlantic sector in spring. Quart J Roy Meteorol Soc in press

Palmer TN, Sun Z (1985) A modelling and observational study of the relationship between sea surface temperature anomalies in the northwest Atlantic and the atmospheric general circulation. Quart J Roy Meteorol Soc 111: 947-975

Palmer TN, Anderson DLT (1994) The prospect for seasonal forecasting - a review paper. Quart J Roy Meteorol Soc 120: 755-793

Palmer TN, Shukla J (2000) Editorial to DSP/PROVOST special issue. Quart J Roy Meteorol Soc 126: 1989-1990.

Palmer TN, Brankovic C, Richardson DS (2000) A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. Quart J Roy Meteorol Soc 126: 2013-2034

Palmer TN, Alessandri A, Andersen U, Cantelaube P, Davey M, Déqué M, Díez E, Doblas-Reyes FJ, Feddersen H, Graham R, Gualdi S, Guérémy JF, Hagedorn R, Hoshen M, Keenlyside N, Latif M, Lazar A, Maisonnave E, Marletto V, Morse AP, Orfila B, Rogel P, Terres JM, Thomson M (2003) Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). Submitted to Bull. Amer. Meteorol. Soc.

Pavan V, Doblas-Reyes FJ (2000) Multi-model seasonal forecasts over the Euro-Atlantic: skill scores and dynamic features. Clim Dyn 16: 611-625

Pavan V, Molteni F, Brankovic C (2000) Wintertime variability in the Euro-Atlantic region in observations and in ECMWF seasonal ensemble experiments. Quart J Roy Meteorol Soc 126: 2143-2173

Peng S, Whitaker JS (1999) Mechanisms determining the atmospheric response to midlatitude SST anomalies. J Clim 12: 1393-1408

Perlwitz J, Graf HF (1995) The statistical connection between tropospheric and stratospheric circulation of the Northern Hemisphere in winter. J Clim 8: 2281-2295

Qian B, Corte-Real J, Xu H (2000) Is the North Atlantic Oscillation the most important atmospheric pattern for precipitation in Europe? J Geophys Res 105: 11901-11910

Richardson DS (2001) Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. Quart J Roy Meteorol Soc 127: 2473-2489

Rodwell MJ, Rowell DP, Folland CK (1999) Oceanic forcing of the wintertime North Atlantic oscillation and European climate. Nature 398: 320-323

Rogers JC (1990) Patterns of low-frequency monthly sea level pressure variability (1899--1986) and associated wave cyclone frequencies. J Clim 3: 1364-1379.

Serreze MC, Carse F, Barry RG, Rogers JC (1997) Icelandic low cyclone activity: Climatological features, linkages with the NAO, and relationship with recent changes in the Northern Hemisphere circulation. J Clim 10: 453-464

Shabbar A, Huang J, Higuchi K (2001) The relationship between the wintertime North Atlantic

Oscillation and blocking episodes in the North Atlantic. Int J Climatol 21: 355-369

Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York.

Stefanova L, Krishnamurti TN (2002) Interpretation of seasonal climate forecast using Brier skill score, the Florida State University superensemble, and the AMIP-I dataset. J Clim 15: 537-544

Stephenson DB (1997) Correlation of spatial climate/weather maps and the advantages of using the Mahalanobis metric in prediction. Tellus 49A: 513-527

Stephenson DB (2000) Use of the "odds ratio" for diagnosing forecast skill. Wea Forecasting 15: 221-232

Stephenson DB, Pavan V (2003) The North Atlantic Oscillation in coupled climate models: a CMIP1 evaluation. Clim Dyn (in press)

Stephenson DB, Pavan V, Bojariu R (2000) Is the North Atlantic Oscillation a random walk? Int J Climatol 20: 1-18

Stephenson DB, Wanner H, Brönnimann S, Luterbacher J (2003) The history of scientific research on the North Atlantic Oscillation. The North Atlantic Oscillation, J. W. Hurrell, Y. Kushnir, G. Ottersen, M. Visbeck, (Eds.), AGU, (in press)

Stern W, Miyakoda K (1995) Feasibility of seasonal forecast inferred from multiple GCM simulations. J Clim 8: 1071-1085

Sutton R, Mathieu PP (2003) Response of the atmosphere-ocean mixed layer system to anomalous ocean heat flux convergence. Submitted to Quart J Roy Meteorol Soc

Swets JA (1973) The relative operating characteristic in psychology. Science 182: 990-1000

Thompson PD (1977) How to improve accuracy by combining independent forecasts. Mon Weather Rev 105: 228-229

Thornes JE, Stephenson DB (2001) How to judge the quality and value of weather forecast products. Meteorol Appl 8: 307-314

Tracton MS, Kalnay E (1993) Operational ensemble prediction at the National Meteorological Center. Practical aspects. Wea Forecasting 8: 379-398

van Loon H, Rogers JC (1978) The seesaw in winter temperatures between Greenland and Northern Europe. Part I: general description. Mon Weather Rev 106: 296-310

Vislocky RL, Fritsch JM (1995) Improved model output statistics forecast through model consensus. Bull Amer Meteor Soc 76: 1157-1164

Walker GT (1924) Correlations in seasonal variations of weather IX. Mem. 24, 275-332, Indian Meteorol Dep, Pune

Wallace JM, Gutzler DS (1981) Teleconnections in the geopotential height field during the Northern Hemisphere winter. Mon Weather Rev 109: 784-812

Wilks DS (1995) Statistical Methods in the Atmospheric Sciences. Academic Press, 1st ed

Zhang H, Casey T (2000) Verification of categorical probability forecasts. Wea Forecasting 15: 80-89

Figure captions

Figure 1: 500-hPa geopotential height correlation coefficient for a) the multi-model, b) ECMWF, c) MetO, d) MetFr and e) EDF over the Euro-Atlantic region for the 2-4 month ensemble mean hindcasts. Contour interval is 0.2 and the zero line has been skipped. Negative values are dashed.

Figure 2: Probability distribution function of grid-point a) ensemble-mean correlation, and b) *RPSS* for tercile categories of the 500-hPa geopotential height over the Euro-Atlantic region. Skill has been computed separately for each single grid point in the region before estimating the distribution. The thick line corresponds to the multi-model and the thin lines to the single-model ensembles. Tercile boundaries have been computed using a kernel-based method (see text for details).

Figure 3: 500-hPa geopotential height signature for the a) positive and b) negative phase of the NAO. Contour interval is 0.2 and the zero line has been skipped. Negative values are dashed. See text for details.

Figure 4: Leading empirical orthogonal function of the 500-hPa geopotential height winter (DJFM) monthly mean anomalies for a) NCEP reanalyses, b) ECMWF, c) MetO, d) MetFr, and e) EDF. Negative values are dashed. Contour interval is 0.5 units and the zero line has been skipped.

Figure 5: a) Winter NAO hindcast (JFM seasonal average) index defined as the projection of 500-hPa geopotential height ensemble anomalies from individual ensemble members onto the first EOF of the NCEP reanalyses (Pobs method). Solid triangles and squares represent the two verifications: NCEP PC1 and JonesÂindices, respectively. The dots correspond to the multi-model ensemble-mean hindcasts, while the small open dots represent the individual ensemble members. b) Same as a), but for the NAO hindcasts obtained from the single-model leading principal component (Pmod method). All NAO index values have been standardized to correct the underestimation of each single-model interannual variance using cross-validation.

Figure 6: Refraction index as a function of the latitude for a) JFM 1985 and b) JFM 1989. The solid line corresponds to the verification, the dashed lines to the ensemble members having a local minimum between 30.N and 45.N and the dotted lines to the rest of the ensemble members.

Table 1: Ranked probability skill score (*RPSS*), area under the *ROC* curve (*ROC*), Peirce skill score (*PSS*), odds ratio (*OR*), and odds ratio skill score (*ORSS*) for the JFM NAO probabilistic hindcasts. 95% statistically significant values appear in bold (see the text for information about the tests applied). The three events considered are: hindcasts above the upper tercile (1), above the mean (2), and below the lower tercile (3).

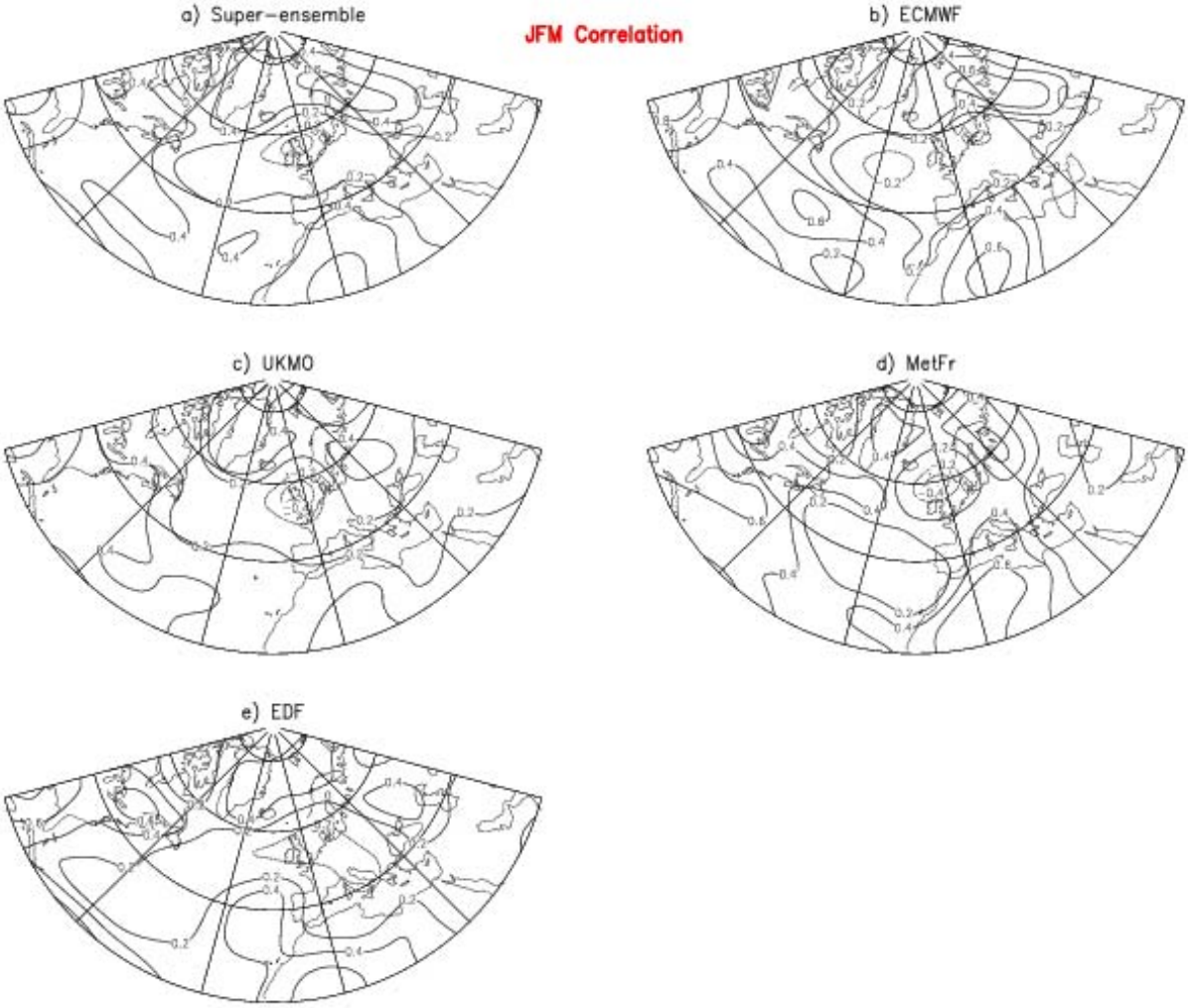

Figure 1: 500-hPa geopotential height correlation coefficient for a) the multi-model, b) ECMWF, c) MetO, d) MetFr and e) EDF over the Euro-Atlantic region for the 2-4 month ensemble mean hindcasts. Contour interval is 0.2 and the zero line has been skipped. Negative values are dashed.

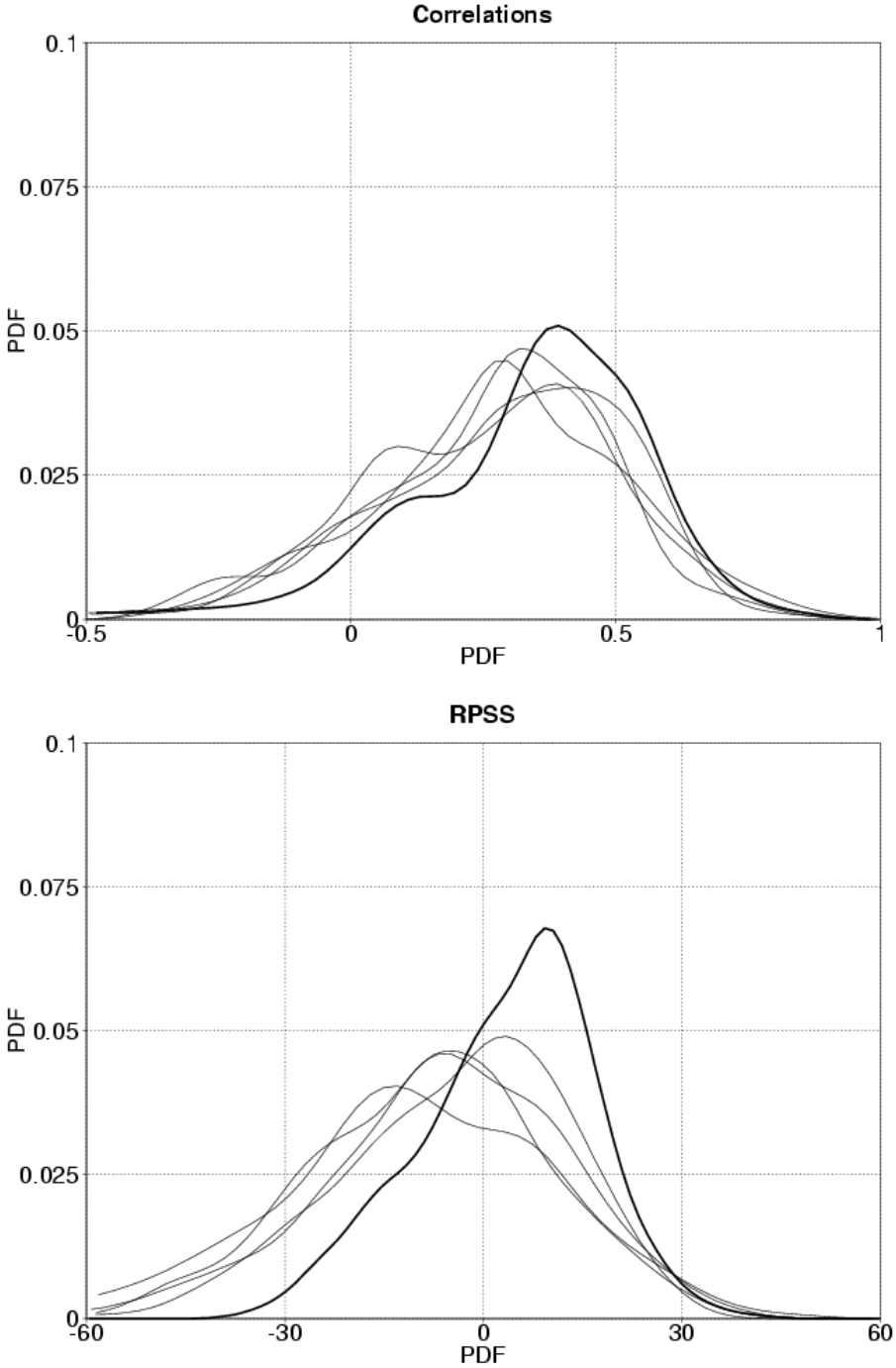

Figure 2: Probability distribution function of grid-point a) ensemble-mean correlation, and b) *RPSS* for tercile categories of the 500-hPa geopotential height over the Euro-Atlantic region. Skill has been computed separately for each single grid point in the region before estimating the distribution. The thick line corresponds to the multi-model and the thin lines to the single-model ensembles. Tercile boundaries have been computed using a kernel-based method (see text for details).

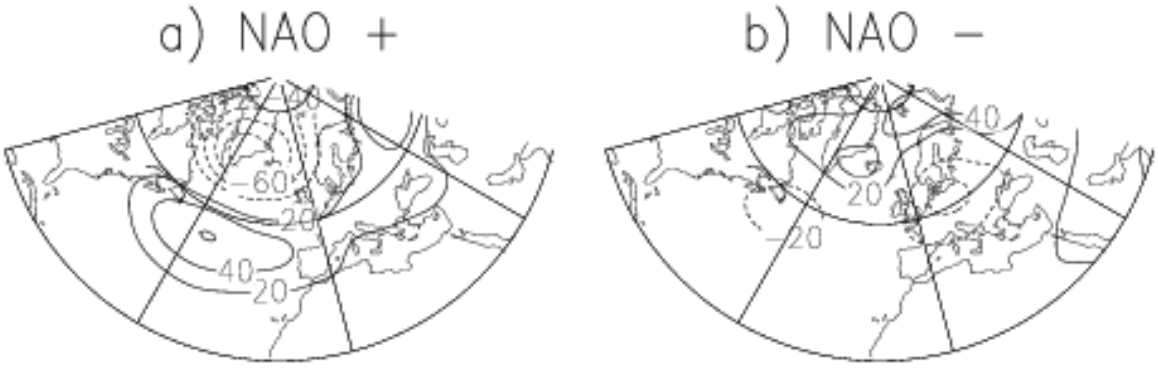

Figure 3: 500-hPa geopotential height signature for the a) positive and b) negative phase of the NAO. Contour interval is 0.2 and the zero line has been skipped. Negative values are dashed. See text for details.

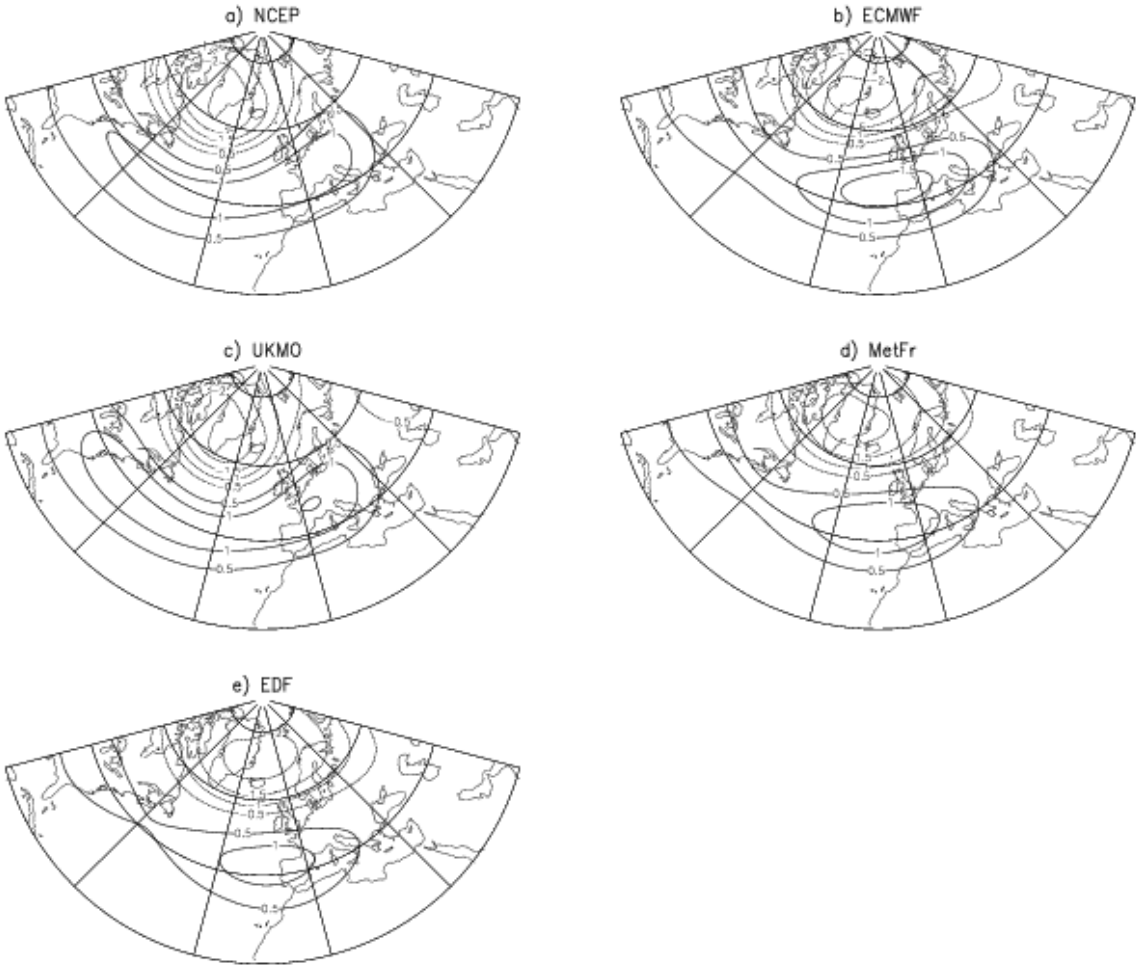

Figure 4: Leading empirical orthogonal function of the 500-hPa geopotential height winter (DJFM) monthly mean anomalies for a) NCEP reanalyses, b) ECMWF, c) MetO, d) MetFr, and e) EDF. Negative values are dashed. Contour interval is 0.5 units and the zero line has been skipped.

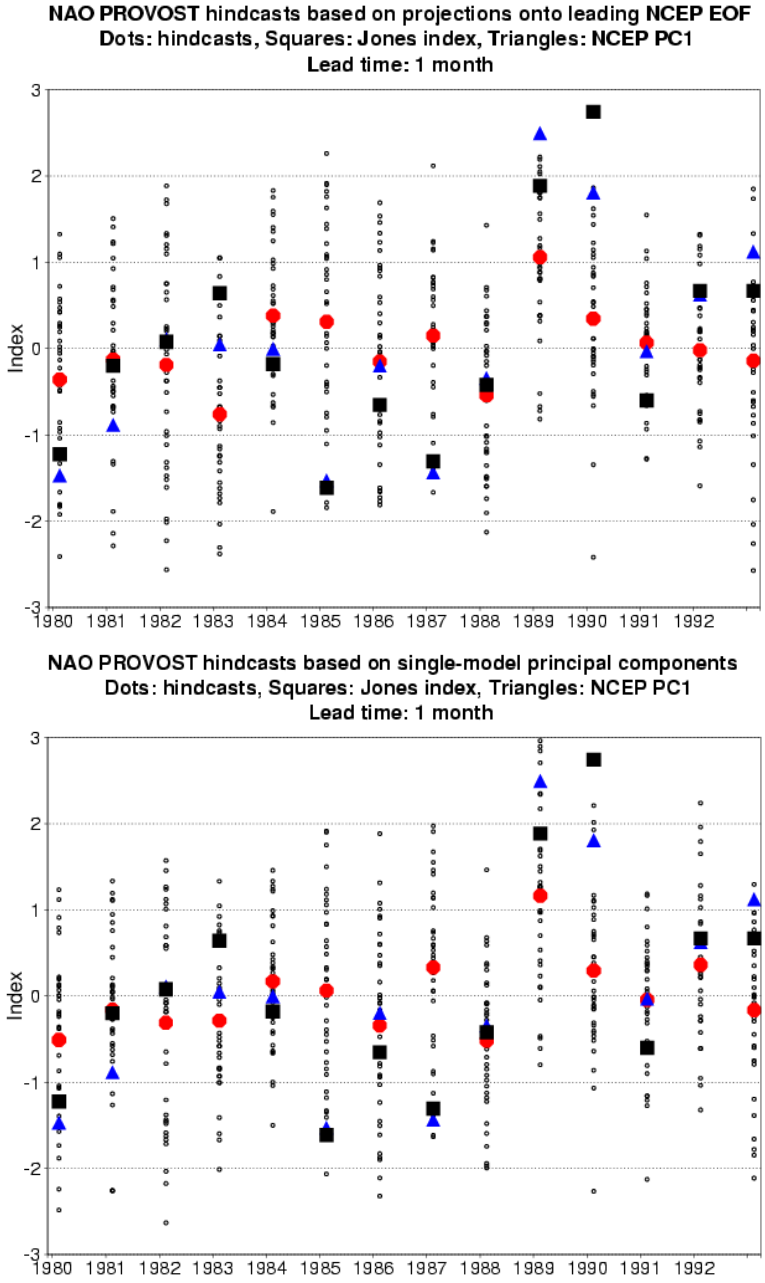

Figure 5: a) Winter NAO hindcast (JFM seasonal average) index defined as the projection of 500-hPa geopotential height ensemble anomalies from individual ensemble members onto the first EOF of the NCEP reanalyses (Pobs method). Solid triangles and squares represent the two verifications: NCEP PC1 and JonesÂindices, respectively. The dots correspond to the multi-model ensemble-mean hindcasts, while the small open dots represent the individual ensemble members. b) Same as a), but for the NAO hindcasts obtained from the single-model leading principal component (Pmod method). All NAO index values have been standardized to correct the underestimation of each single-model interannual variance using cross-validation.

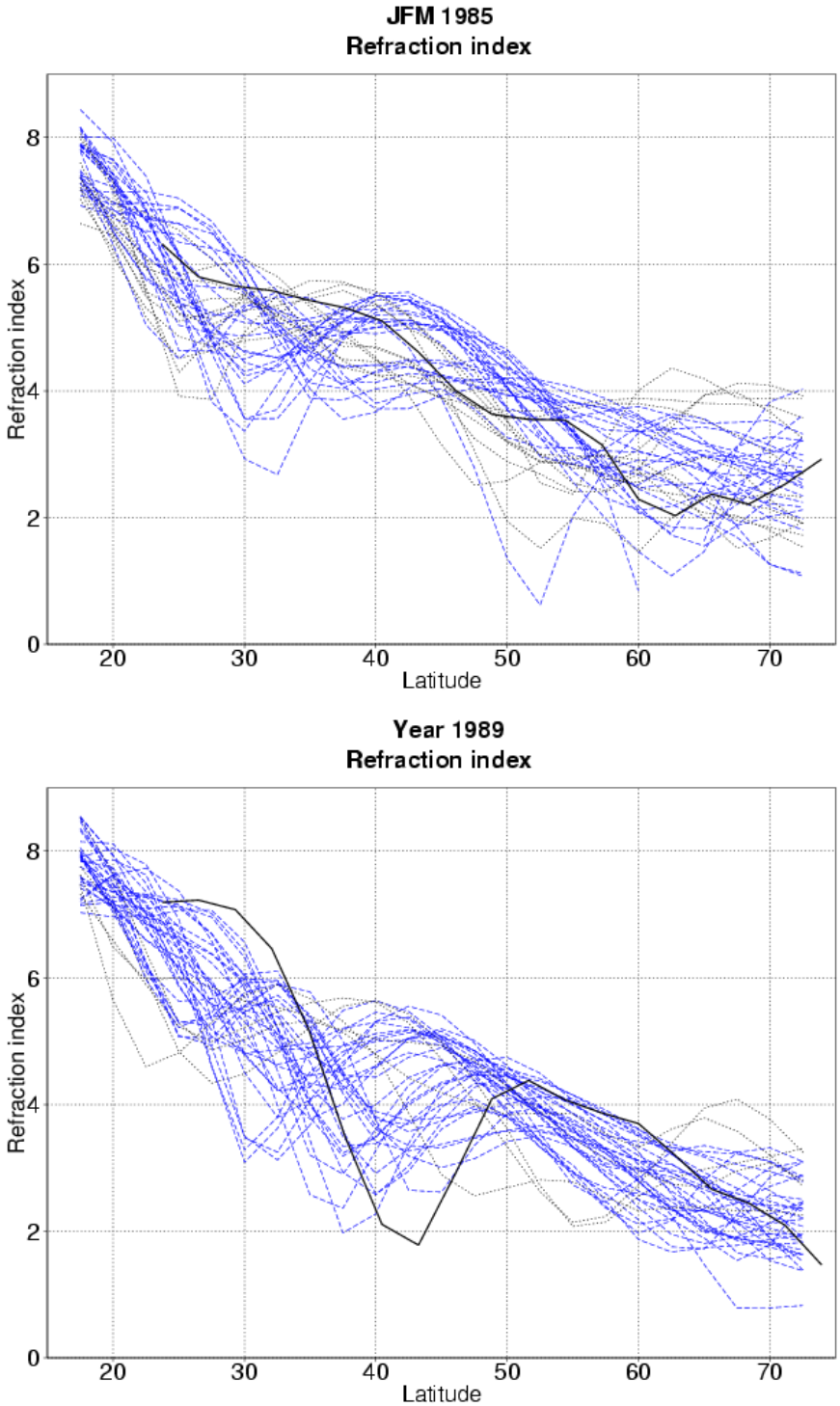

Figure 6: Refraction index as a function of the latitude for a) JFM 1985 and b) JFM 1989. The solid line corresponds to the verification, the dashed lines to the ensemble members having a local minimum between 30.N and 45.N and the dotted lines to the rest of the ensemble members.

Table 1

	ECMWF			**MetO**			**MetFr**			**EDF**			**S-E**		
RPSS	-18.2			**9.3**			4.0			**3.8**			**13.1**		
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
ROC	.63	.56	.46	.57	.50	**.74**	.53	.52	.70	**.80**	.63	.59	.66	.60	**.66**
PSS	.12	.06	-.03	.11	.03	.13	.04	.06	**.19**	**.23**	.12	.05	**.13**	.07	**.09**
OR	1.68	1.28	.88	1.63	1.12	1.76	1.12	1.28	**2.32**	**2.86**	1.63	1.25	**1.83**	1.31	**1.52**
ORSS	.25	.12	-.06	.24	.06	.28	.08	.12	.40	**.48**	.24	.11	.29	.13	.21

Table 1: Ranked probability skill score (*RPSS*), area under the *ROC* curve (*ROC*), Peirce skill score (*PSS*), odds ratio (*OR*), and odds ratio skill score (*ORSS*) for the JFM NAO probabilistic hindcasts. 95% statistically significant values appear in bold (see the text for information about the tests applied). The three events considered are: hindcasts above the upper tercile (1), above the mean (2), and below the lower tercile (3).