Royal Meteorological Society

Ensemble prediction of transitions of the North Atlantic eddy-driven jet

T. H. A. Frame,[a*] M. H. P. Ambaum,[b] S. L. Gray[b] and J. Methven[b]

[a] *National Centre for Atmospheric Science (NCAS), University of Reading, Reading, UK*

[b] *Department of Meteorology, University of Reading, Reading, UK*

*Correspondence to: T. H. A. Frame, Department of Meteorology, The University of Reading, Earley Gate, PO Box 243, Reading, RG6 6BB, UK. E-mail: t.h.a.frame@reading.ac.uk

The consistency of ensemble forecasts from three global medium-range prediction systems with the observed transition behaviour of a three-cluster model of the North Atlantic eddy-driven jet is examined. The three clusters consist of a mid jet cluster taken to represent an undisturbed jet and south and north jet clusters representing southward and northward shifts of the jet. The ensemble forecasts span a period of three extended winters (October–February) from October 2007–February 2010. The mean probabilities of transitions between the clusters calculated from the ensemble forecasts are compared with those calculated from a 23-extended-winter climatology taken from the European Centre for Medium-Range Weather Forecasts 40-Year Re-analysis (ERA40) dataset. No evidence of a drift with increasing lead time of the ensemble forecast transition probabilities towards values inconsistent with the 23-extended-winter climatology is found. The ensemble forecasts of transition probabilities are found to have positive Brier Skill at 15 day lead times. It is found that for the three-extended-winter forecast set, probabilistic forecasts initialized in the north jet cluster are generally less skilful than those initialized in the other clusters. This is consistent with the shorter persistence time-scale of the north jet cluster observed in the ERA40 23-extended-winter climatology. Copyright © 2011 Royal Meteorological Society

Key Words: TIGGE; weather regimes; clustering

Received 21 December 2010; Revised 11 March 2011; Accepted 24 March 2011; Published online in Wiley Online Library

Citation: Frame THA, Ambaum MHP, Gray SL, Methven J. 2011. Ensemble prediction of transitions of the North Atlantic eddy-driven jet. *Q. J. R. Meteorol. Soc.* DOI:10.1002/qj.829

1. Introduction

The concept of weather regimes has long been invoked to explain the perception that weather conditions appear to persist longer than the passage of individual systems. This idea was initially closely related to the concept of weather analogues: the assumption that similar large-scale flow patterns are associated with similar weather types and evolve in a similar manner. In this vein, catalogues of regime classifications such as Grosswetterlagen (Hess and Brezowsky, 1952) aimed to provide a qualitative partitioning of the observed atmosphere into a discrete set of flow types, each associated with different weather conditions. The advent of dynamical systems theory and the discovery of chaos (Lorenz, 1963) both debunked the atmospheric analogues idea and appeared to provide an explanation for the existence of atmospheric regimes. In low-dimensional nonlinear systems, the regimes are associated with stable (or weakly unstable) equilibrium solutions to the dynamical equations to which the state remains close. The wings of the Lorenz (1963) 'butterfly' are the classic example of this behaviour. Whilst there have been attempts to explain atmospheric regimes through equilibrium solutions to low-dimensional atmospheric models (Charney and

DeVore, 1979; Crommelin, 2003), the link to high-dimensional atmospheric global circulation models and the actual atmosphere remains unclear. Regimes in such high-dimensional systems are usually diagnosed from output data by examination of probability density function estimates for evidence of multimodality (Silverman, 1981; Corti *et al.*, 1999; Ambaum, 2008; Woollings *et al.*, 2010b) and applying statistical techniques such as clustering (Smyth *et al.*, 1999; Hannachi, 2007; Cassou, 2008; Franzke *et al.*, 2009), rather than by analysis of the dynamical equations themselves.

One of the motivating factors for interest in regimes is their implications for predictability. These implications are something of a double-edged sword: on the one hand, knowing that you have entered a persistent regime may provide useful predictive skill for extended-range forecasting, but conversely failing to predict a change of regime accurately may lead to a significant loss in skill. One of the stated purposes of medium-range ensemble forecasting is to account for the possibility of small uncertainties in initial conditions leading to large differences in forecast outcomes, due to the nonlinear nature of the atmosphere. As such, if regimes (which are an inherently nonlinear phenomenon) exist, ensemble forecasts should, by design, be able to capture the transitions between them. Regardless of the existence (or not) of atmospheric regimes, cluster analysis provides a low-dimensional approximation to the atmospheric phase space, which optimally characterizes the broad characteristics of atmospheric data with respect to a chosen measure. This article addresses the question of whether operational medium-range ensemble forecasts replicate the statistics and predict the future state of such low-dimensional representations of the atmosphere. This is approached by examining the ability of the global 15 day ensemble forecasts from three different forecasting centres taken from the Thorpex Interactive Grand Global Ensemble (TIGGE) dataset (Park *et al.*, 2008) to replicate the transition statistics of a three-cluster model designed to characterize the behaviour of the North Atlantic eddy-driven jet (Woollings *et al.*, 2010a). The ensemble forecasts used in the study come from the European Centre for Medium-Range Weather Forecasts (ECMWF), the (UK) Met Office and the Meteorological Service of Canada (CMC). For details on the forecast models and data the reader is referred to http://tigge.ecmwf.int.

The rest of the article is divided into four sections. Section 2 provides an introduction to the three North Atlantic eddy-driven jet regimes and the clustering method used to identify them in forecast data. Section 3 contains an examination of the ability of the forecast models to replicate the climatological probabilities of regime transition. In section 4 the skill of the forecasts in predicting regime transitions is assessed. A summary and conclusions are contained in section 5.

2. Cluster and transition probability definition

Following Woollings *et al.* (2010a) we decompose the low-level zonal wind in the North Atlantic sector into three possible jet configurations. These three configurations are identified from low-level zonally averaged zonal wind in the North Atlantic sector and are designed to be representative of the North Atlantic eddy-driven jet. The use of low-level winds as a diagnostic is designed to separate the eddy-driven component from the subtropical jet, since the former is assumed to have a signal throughout the depth of the atmosphere, whereas the latter is assumed to be more confined to the upper levels. The physical motivation behind this assumption is the interpretation of the subtropical jet as a vertically confined upper-level baroclinic jet in vorticity balance with the meridional overturning circulation. By contrast, the eddy-driven jet is assumed to have a more barotropic structure reflecting the tendency of synoptic eddies to reduce baroclinicity by accelerating the westerly flow throughout the depth of the atmosphere (Hoskins *et al.*, 1983). We define the North Atlantic eddy-driven jet profile to be the zonally and vertically averaged zonal wind between $\lambda_1 = 300^\circ$ and $\lambda_2 = 360^\circ$E, and between the $p_1 = 700$ hPa and $p_2 = 925$ hPa pressure surfaces, i.e.

$$U(\phi, t) = \frac{1}{N_p N_\lambda} \sum_{p=p_1}^{p_2} \sum_{\lambda=\lambda_1}^{\lambda_2} u(\lambda, \phi, p, t),$$

where ϕ and t denote latitude and time, respectively, and N_p and N_λ are the number of levels and grid points between p_1 and p_2 and λ_1 and λ_2 respectively.

The three jet clusters are identified by K-means clustering (Jain, 2010) with three degrees of freedom, applied to daily mean jet profiles calculated for European Centre for Medium-Range Weather Forecasts 40-Year Re-analysis (ERA40) data (Uppala *et al.*, 2005) covering the extended winters (October–February) from October 1978–February 2002. The operation of K-means on the data may be summarized as follows. With the choice of three degrees of freedom, the K-means algorithm identifies three jet-profile cluster centroids which define a partitioning of the jet-profile data into three clusters. The partitioning is defined such that each jet profile, $U(\phi, t)$, is allocated to the cluster with centroid, $U_c(\phi)$, closest to it in the squared Euclidean norm,

$$|U - U_c|^2 = \sum_{\phi=\phi_1}^{\phi_2} \left[U(\phi, t) - U_c(\phi)\right]^2 .$$

The K-means algorithm identifies the three centroids that minimize the sum of the squared Euclidean distances of all jet profiles from their respective centroids.

Figure 1 shows the three cluster centroids, which are labelled south ('S'), mid ('M') and north ('N') to reflect the latitude of the wind maxima associated with each. Figure 2 shows composites of 500 hPa geopotential height anomalies obtained from the mean over all days allocated to each regime in the 23-extended-winter climatology. These composites also show a close qualitative similarity to those obtained by Woollings *et al.* (2010a) using the latitude of the maximum of the zonal jet profile to partition the data. The mid and south jet composites are reminiscent of the positive and negative North Atlantic Oscillation (NAO) regimes identified by Cassou (2008).

The choice of the number of degrees of freedom for the K-means algorithm can be somewhat arbitrary (Christiansen, 2007), particularly for atmospheric data that does not usually provide strong evidence of multimodality; see e.g. Stephenson *et al.* (2004) and Ambaum (2008). In the case of the work presented in this article, the choice of the number clusters is based on both the evidence of three preferred jet locations presented by Woollings *et al.* (2010a) and the more heuristic argument that the three clusters appear adequately to capture the qualitative behaviour observed in time series

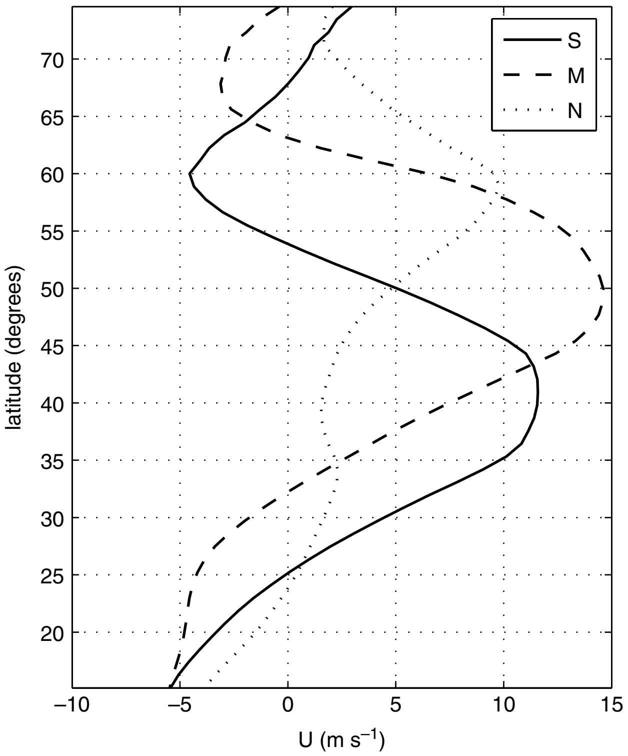

Figure 1. Jet-profile cluster centroids obtained by K-means clustering (with three degrees of freedom) of ERA40 jet-profile data covering 23 extended winters (ONDJF) from October 1978–February 2002.

of the jet profile U. Here the cluster analysis is not intended to provide evidence of multimodality but rather to provide a simple means of characterizing the variability of the jet which can be readily applied to forecast data.

The choice to use jet profiles to partition the data rather than a partitioning based on the jet maxima, as might be suggested by the work of Woollings *et al.* (2010a), is made because it is found to produce much greater consistency when applied to different datasets. As a test of consistency, the K-means algorithm was applied to jet profiles from the National Centers for Environmental Prediction (NCEP) reanalysis (Kalnay *et al.*, 1996) for the same 23 extended winters, producing cluster centroids with mean-squared difference (normalized by mean-squared amplitude) from the ERA40-derived clusters of ~ 0.01 (once the centroids were interpolated on to the same grid) and $\sim 95\%$ agreement in the allocation of data to clusters. By contrast, tests of K-means and Gaussian mixture models applied to the latitude of the jet maxima could only produce $\sim 75\%$ agreement between the allocation of the data to clusters. The greater consistency between the clustering of the two datasets when jet-profile data is used is probably attributable to the fact that the K-means algorithm picks out the large-scale structure of the jet profiles and is therefore less sensitive to noise and resolution.

The result of the K-means clustering is that the ERA40 jet-profile data, $U(\phi, t)$, is reduced to an indicator variable, X_t, which takes one of the values S, M or N depending on which cluster the jet belongs to at time t, i.e.

$$X_t(U(\phi,t)) = \underset{c=\mathrm{S,M,N}}{\operatorname{argmin}} \left\{ |U - U_{\mathrm{c}}|^2 \right\}.$$

The TIGGE dataset is reduced to a similar form using the cluster centroids obtained from the ERA40 data.

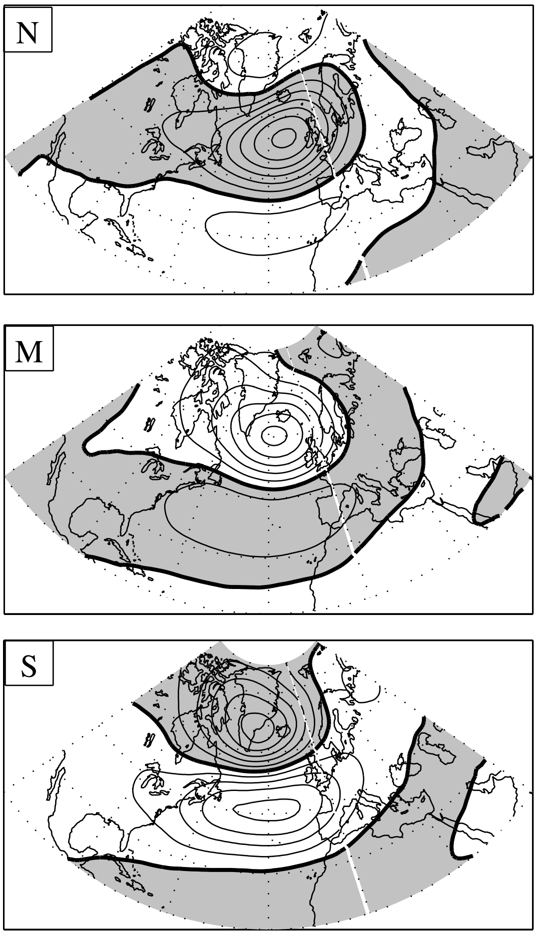

Figure 2. Composite of ERA40 500 hPa geopotential height anomalies calculated from all ONDJF days (October 1978–February 2002) allocated to each of the three jet profile clusters. From top to bottom: north jet days, mid jet days and south jet days. Contour interval is 25 m; shading indicates positive (anticyclonic) values; the thick line is the zero contour. Latitudinal grid lines are shown every 15°, longitudinal grid lines every 30°.

To gain some insight into the manner in which the jet moves between clusters in time, and to facilitate comparison between analyses and ensemble forecasts, we define a lagged conditional probability of cluster membership between two clusters A and B as

$$P_{A\to B}(\tau) = P\left(X_{t+\tau} = B \mid X_t = A\right).$$

This is the probability that the jet belongs to cluster B at time $t+\tau$ given that it belonged to cluster A at time t. This probability measure takes no account of the values taken by X in the time interval between t and $t+\tau$. Despite this, for small τ, one can loosely interpret the $P_{A\to A}(\tau)$ as the probability of A persisting for τ days, and $P_{A\to B}(\tau)$ as the probability of transition from A to B in time τ. For this reason and for concision we shall refer to the probabilities $P_{A\to B}$ loosely as transition probabilities.

Given a time series X_t, the transition probability $P_{A\to B}$ is estimated by the following steps: take the indices of the subset of all time points for which $X_t = A$; count the number, N_A, of data points in the the subset; shift the indices of the subset forward by τ; count the number of data points, N_B, in the forward-shifted subset that belong to cluster B; the transition probability is then given by $P_{A\to B} = N_B/N_A$.

Since a single ensemble forecast contains multiple estimates of the atmospheric state at a lead time τ, it is possible to use the ensemble to calculate probabilities of individual events. The simplest strategy for converting ensembles into a probabilistic prediction of a categorical event is to use the fraction of the ensemble for which the event occurs as an estimator. For TIGGE ensemble forecasts we define the predicted probability of membership of the cluster B at lead time τ to be the fraction of the ensemble in cluster B at lead time τ. For an ensemble forecast with initial analysis in cluster A, the predicted probability of membership of cluster B at lead time τ is taken as analogous to a predicted transition probability $P_{A\rightarrow B}(\tau)$; note that this definition ignores the fact that adding perturbations to the initial analysis to create the ensemble of initial conditions means that not all ensemble members are guaranteed to be in cluster A initially.

3. Comparison of 'climatological' transition probabilities from forecasts with reanalysis

The first question to ask when assessing whether the forecasting systems are able to replicate the observed clustering behaviour is whether their statistics lie within the bounds of the observed climatology. To answer this question, we compare transition probabilities calculated using the 23-extended-winter ERA40 climatology (ONDJF, October 1978–February 2002) with those calculated using three extended winters of TIGGE operational analyses (ONDJF, October 2007–February 2010) and those obtained by averaging the predicted transition probabilities from TIGGE ensemble forecasts for the same three extended winters.

Figure 3 shows the transition probabilities calculated from the ERA40 data (thick solid lines) and those calculated from TIGGE ECMWF operational analysis data (thick dashed lines); note that the use of Met Office and CMC analyses rather than those from ECMWF is found to make negligible difference to the results. To give an indication of how much transition probabilities calculated from a three-extended-winter subsample are expected to deviate from those of a longer term climatology, the grey shading shows a relative frequency histogram of the transition probabilities calculated using three-extended-winter subsamples of the ERA40 data. The three-extended-winter subsamples are overlapping. Each subsample comprises adjacent winters as this most closely resembles the nature of the three-extended-winter TIGGE dataset. For each transition probability $P_{A\rightarrow B}$, the thin horizontal black line indicates climatological occupancy, $P(X = B)$, of the cluster B calculated for the ERA40 data; i.e. the total fraction of the ERA40 data in cluster B.

The smallest values of τ for which the ERA40 transition probabilities (thick solid lines) intersect the climatological occupancy indicates the time-scale over which the transition probability converges to the climatological occupancy; this may be thought of as the time-scale over which knowing the state at time t provides no more information about the state at $t + \tau$ than could be inferred by the climatological occupancy. Comparing the climatological transition probabilities (thick solid lines) with the climatological occupancy (thin horizontal lines), it is evident that transition probabilities involving only the south and mid clusters ($P_{\mathrm{S}\rightarrow \mathrm{S}}$, $P_{\mathrm{S}\rightarrow \mathrm{M}}$, $P_{\mathrm{M}\rightarrow \mathrm{S}}$, $P_{\mathrm{M}\rightarrow \mathrm{M}}$) remain noticeably different from the climatological occupancy out to 15 days. In fact with further analysis (not shown) it is found that τ needs to be longer than ~ 30 days before the two lines intersect. This is consistent with the south and mid clusters being related to the negative and positive phases of the NAO, which is known to possess a long decorrelation time-scale (Ambaum and Hoskins, 2002; Keeley *et al.*, 2009). By contrast transitions involving the north cluster ($P_{\mathrm{S}\rightarrow \mathrm{N}}$, $P_{\mathrm{M}\rightarrow \mathrm{N}}$, $P_{\mathrm{N}\rightarrow \mathrm{S}}$, $P_{\mathrm{N}\rightarrow \mathrm{M}}$, $P_{\mathrm{N}\rightarrow \mathrm{N}}$) approach very close to or intersect the climatological occupancy within 15 days.

The variation in the transition probabilities calculated using different three-year periods of the ERA40 data (grey shading) is large. This large variation means that one can reasonably expect the transition probabilities calculated using three years of TIGGE data to differ significantly from those of the longer term ERA40 climatology. This is born out by the thick dashed lines in Figure 3, which show the transition probabilities calculated using the ECMWF operational analysis data from the TIGGE data. However, despite their deviation from the long-term climatology, the transition probabilities calculated using the TIGGE analysis do not lie beyond the grey shaded area and are therefore not unprecedented given the ERA40 climatological record. Whether the variation of the transition probabilities calculated for three-extended-winter periods should be interpreted as sampling error or as non-stationarity in the statistic itself is an issue beyond the scope of this work. The primary focus is the assessment of the consistency of the ensemble predicted transition probabilities with those of analysis/re-analysis.

To see clearly the relationship between the TIGGE ensemble predicted transition probabilities and those calculated from the ERA40 and TIGGE analysis data, Figure 4 shows the deviations

$$\Delta P_{A\rightarrow B} = P_{A\rightarrow B} - P^{\mathrm{clim}}_{A\rightarrow B}$$

of transition probabilities $P_{A\rightarrow B}$ from the values $P^{\mathrm{clim}}_{A\rightarrow B}$ calculated using the 23-extended-winter ERA40 climatology. The thick solid line showing $\Delta P_{A\rightarrow B} = 0$ is analogous to the thick solid line in Figure 3. Consistent with Figure 3, the grey shading shows a relative frequency histogram of transition probability deviations calculated from three-year subsamples of the ERA40 data, and the thick dashed line shows the transition probability deviations calculated from the TIGGE ECMWF operational analysis data. The crossed, circled and asterisked lines show the mean (over the TIGGE dataset) predicted transition probability deviations for the ECMWF, CMC and Met Office ensemble forecasts respectively. Two general points stand out in Figure 4. Firstly, at no point do the mean predicted transition probabilities deviate further from the ERA40 climatological transition probabilities than would be expected given variability associated with three-extended-winter subsamples, i.e. the mean deviation of the predicted transition probabilities remains on the grey shaded area. Secondly, large deviations ($\Delta P_{A\rightarrow B} \gtrsim 0.1$) of mean predicted transition probabilities from $P_{A\rightarrow B}$ are closely associated with large deviations of the TIGGE analysis transition probabilities; see for example $\Delta P_{\mathrm{S}\rightarrow \mathrm{S}}$ and $\Delta P_{\mathrm{S}\rightarrow \mathrm{M}}$. Considering both these points, Figure 4 provides no evidence of the forecast transition probabilities drifting towards unphysical climatological values over a 15 day lead time.

At short lead times the mean predicted transition probabilities tend to follow the TIGGE analysis transition probabilities, whereas at long lead times the mean ensemble predicted transition probabilities tend to be close to

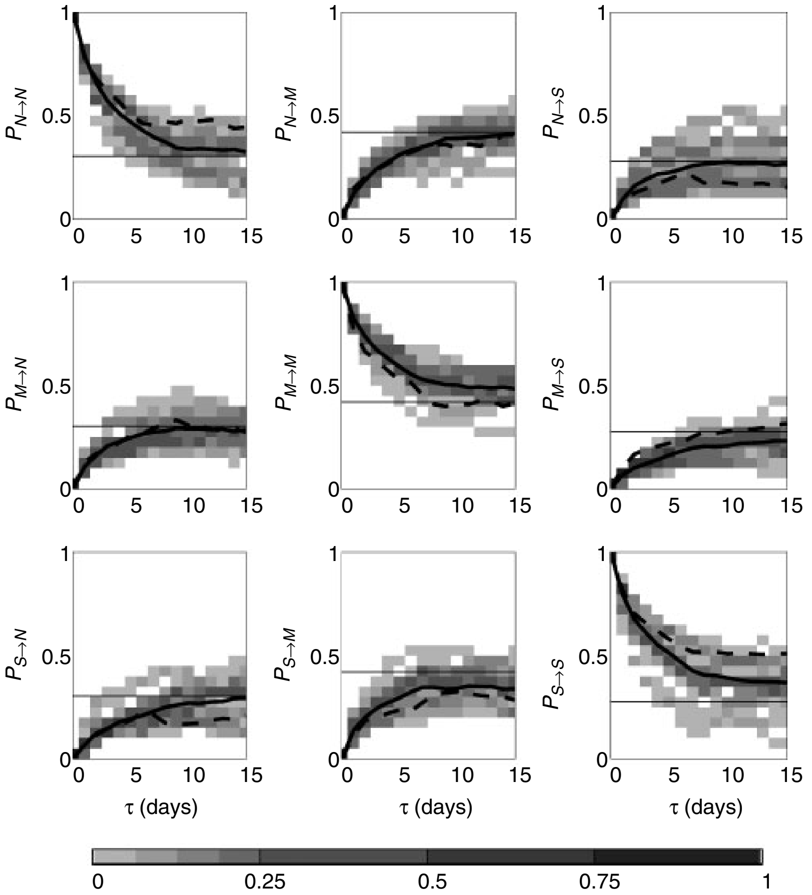

Figure 3. Estimated probabilities of transition between the clusters N, M and S versus lag. Thick solid line: 23-extended-winter (ONDJF, October 1978–February 2002) ERA40 climatological transition probabilities. Thin horizontal line: ERA40 23-extended-winter climatological cluster occupancy. Dashed line: operational analysis from TIGGE data (ONDJF, October 2007–February 2010). Grey shading: relative frequency histogram of transition probabilities obtained from three-extended-winter subsamples of the ERA40 data (zero values are shown in white).

or somewhere between the ERA40 climatological mean and the TIGGE analysis transition probabilities; see for example $\Delta P_{\mathrm{N}\to\mathrm{N}}$. This is consistent with a gradual loss of skill/predictability over the course of the forecast lead time. $\Delta P_{\mathrm{N}\to\mathrm{S}}$ is a particularly striking example in that the mean predicted transition probabilities from all three forecasting centres follow TIGGE analysis transition probabilities up to about $\tau = 7$ days, then drift back to the ERA40 climatological value by day 15.

4. Skill of TIGGE forecast transition probabilities

In section 3 it is shown that there is no evidence of a drift of the mean TIGGE forecast transition probabilities towards climatologically inconsistent values. It is found, rather, that the behaviour of the forecast transition probabilities with increasing lead time is consistent with a drift toward climatological values consistent with loss of predictability/forecast skill. To assess the skill of the TIGGE forecast transition probabilities, we will utilize the Brier Skill Score (Brier, 1950). The Brier Skill Score provides a means of assessing the quality of probabilistic forecasts of categorical ('yes/no') events relative to some baseline method of forecasting. This baseline forecasting method is usually taken to be repeatedly issuing the climatological probability of the event. The Brier Skill Score (BSS) is defined in terms of the ratio of the Brier Score (BS) for the two forecasting methods:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{clim}}}, \qquad (1)$$

such that a score of 1 implies perfect skill and scores less than or equal to zero imply that one would be better or no worse off simply by issuing the climatological probability of the event instead of attempting to produce a more informative forecast. The Brier Score is defined as

$$\mathrm{BS} = \frac{1}{N_{\mathrm{f}}} \sum_{i=1}^{N_{\mathrm{f}}} (f_i - o_i)^2,$$

where N_{f} is the number of forecasts, f_i is the ith forecast probability of the event and the outcome is

$$o_i = \begin{cases} 1, & \text{event occurs,} \\ 0, & \text{event does not occur.} \end{cases}$$

Forecasting high probabilities for events that occur and low probabilities for events that do not occur reduces the Brier Score. Note that BS is defined such that it is decreased by making better forecasts, whereas for BSS (Eq. (1)) the converse is true.

The Brier Skill Score for each of the possible forecast transition probabilities is calculated using the ERA40 climatological transition probabilities as the baseline method

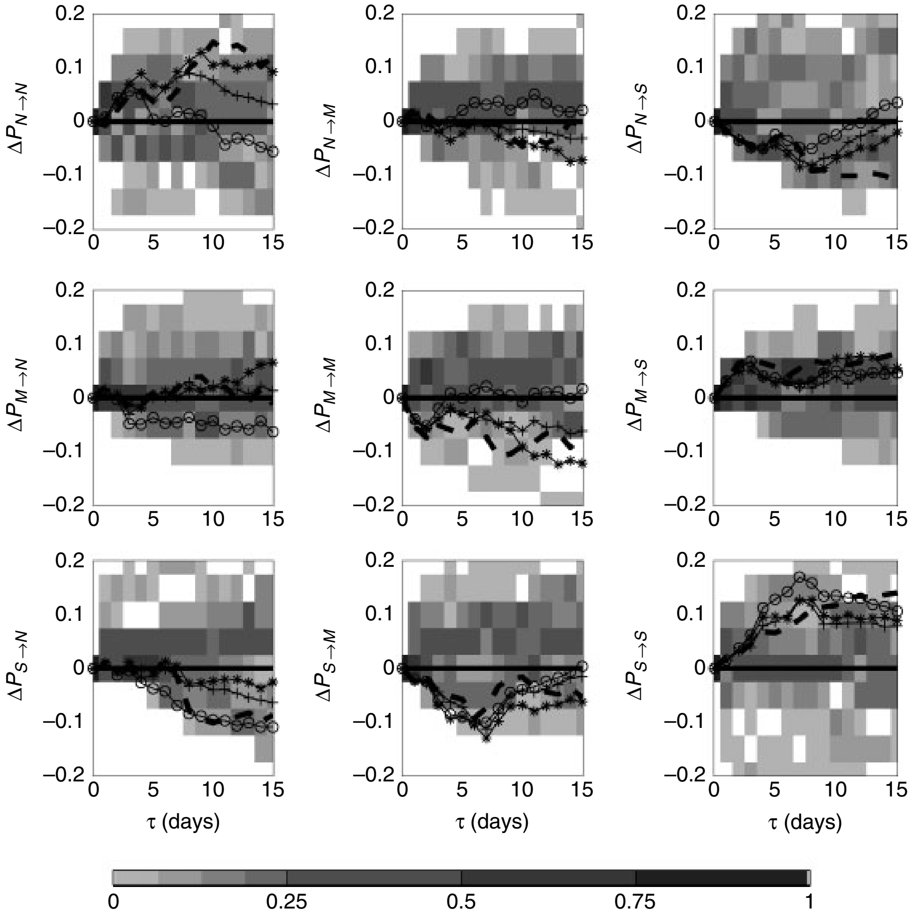

Figure 4. Deviation of transition probabilities from ERA40 climatological values. Dashed line: TIGGE operational analysis transition probabilities (ONDJF, October 2007–February 2010). Asterisked, crossed and circled lines: mean forecast transition probabilities from the three forecasting centres (ONDJF, October 2007–February 2010). Grey shading: relative frequency histogram of transition probabilities obtained from three-extended-winter subsamples of ERA40 data (zero values are shown in white).

of forecasting. The TIGGE forecast transition probabilities, $P_{A\to B}(\tau)$, are defined as described in section 2, but for clarity the method is briefly summarized here. The forecast probability of being in cluster B at lead time τ is calculated as the fraction of the ensemble in cluster B at lead time τ. The initial cluster (A) and the verifying outcome are defined from the ECMWF operational analyses. To avoid bias in favour of any one centre, only forecasts for which the initial and outcome cluster were the same for the analyses from all forecasting centres were used to assess the skill, although this decision was found to have negligible impact on the results. The Brier Skill Score versus lead time is shown in Figure 5.

A noticeable feature of Figure 5 is the high degree of similarity in the manner in which the skill of the three different forecasting centres changes with lead time. The similarity of the skill scores provides evidence for the general applicability of the results to recently/currently operational forecasting systems. The fact that they are so similar, even containing similar ‘bumps’ and ‘wiggles’ (e.g. at nine days for $P_{\mathrm{M}\to\mathrm{S}}$), is an indication that the scores may be strongly influenced by individual synoptic events that occurred during the TIGGE period. A clear example of this (not shown) is that removal of a large section of data from the winter of 2009–2010, during which the flow was characterized by a persistent southward shift of the jet or negative NAO (Cattiaux *et al.*, 2010), removes much of the skill of forecasts initialized in the south jet cluster ($P_{\mathrm{S}\to\mathrm{S}}$, $P_{\mathrm{S}\to\mathrm{M}}$, $P_{\mathrm{S}\to\mathrm{N}}$) beyond about seven days. The sensitivity to the removal of long persistent sections of the data serves to highlight the fact that the statistical degrees of freedom of the Brier Skill Score for the forecast transition probabilities is likely to be smaller than the number of forecasts in the TIGGE dataset. This means that we should not assume that the performance of the TIGGE forecasts is representative of a larger population of forecasts. However, using the Brier Skill Score to verify the TIGGE data allows us to distinguish between skilful forecast probabilities and ensembles constructed by drawing randomly from climatological statistics, as long as we remember that BS and hence BSS are conditionally distributed on the outcomes o_i (see e.g. Ferro, 2007).

The broad features of Figure 5 are that all different transitions are skilfully predicted in the first few days, with skill dropping off quite sharply after about 3–5 days. Several of the transitions show a distinct reduction in the rate at which skill falls off with lead time after about 7–10 days. This feature is most apparent in transitions involving the south and mid clusters, and least apparent in those involving the north cluster (particularly transitions between north and south). At long lead times (days 13–15), the forecasts initialized in the north jet cluster are less skilful than those initialized in the south and mid jet clusters. The skill of predictions of transition between the south and north clusters ($P_{\mathrm{S}\to\mathrm{N}}$) is also low relative to the other transitions.

To examine further the possible reasons for the differences between the skill of predictions of the different transitions,

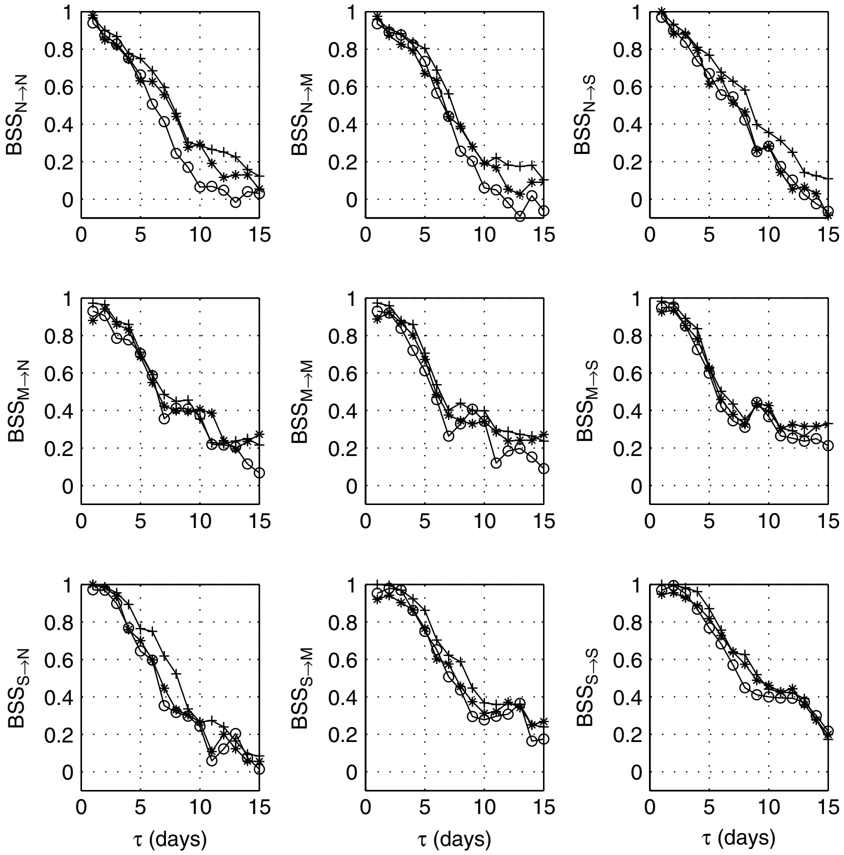

Figure 5. Brier Skill Score as a function of lead time computed for predicted transition probabilities derived from TIGGE ensemble forecasts (ONDJF, October 2007–February 2010). The asterisked, crossed and circled lines show the different forecasting centres.

Figure 6 shows a reliability (or attributes) diagram computed for day 10 (dashed line) and day 15 (solid line) of the ECMWF forecasts (similar diagrams for the other centres produce qualitatively similar results). The reliability diagram provides a graphical means of assessing whether the predicted probabilities of an event correspond to the observed frequency. To construct the diagram, each forecast is allocated into one of a set of discrete bins depending on the forecast probability. For each forecast probability bin, the observed relative frequency (the average of the outcome variable o_i for all the forecasts in the bin) is calculated. The observed relative frequencies are then plotted against the forecast probability so that (for a large enough sample) if the forecast probabilities are quantitatively accurate (calibrated) then the plotted points will lie exactly on the diagonal. Vertical and horizontal lines mark the climatological probability of each transition, and the grey shaded area marks regions associated with positive contribution to the Brier Skill Score. It should be noted that in Figure 6 the horizontal/vertical lines and grey shading are plotted for day 15 values, although from Figure 3 it can be seen that day 10 values would not be markedly different in most cases. The bar chart beneath each diagram shows the number of forecasts in each probability bin at day 10 (open bars) and day 15 (shaded bars). For a full discussion of reliability diagrams, the reader is referred to Murphy and Winkler (1977) and Hsu and Murphy (1986).

Looking first at the reliability diagrams for transitions from the north jet cluster, the forecast transition probabilities are more densely concentrated near the climatological values at day 15 (filled bars) than for other transitions. This greater contraction of the forecast transition probabilities to climatological values is consistent with the shorter time-scale over which the ERA40 climatological transition probabilities involving the north jet cluster become equal to the climatological occupancy (Figure 3). For transitions from the north to south jet clusters, the flatness of the day 15 reliability curve (solid line) between $P_{N\to S} = 0$ and $P_{N\to S} = 0.5$ relative to the day 10 curve (dashed) is consistent with overestimation of the transition probability in the forecasts compared with the analyses, and with the drift of mean TIGGE forecast transition probabilities to ERA40 climatological transition probabilities seen in Figure 4. For $P_{S\to S}$, the skill of the TIGGE forecasts is associated largely with accurately predicting very high transition probabilities for transitions that do occur or conversely very low probabilities for transitions that do not occur. The day 15 reliability curve (solid line) is, however, fairly flat between $P_{S\to S} = 0.1$ and $P_{S\to S} = 0.7$. This is consistent with skill in TIGGE forecasts of $P_{S\to S}$ being associated with a long-lived predictable southward shift of the jet in winter 2009/2010. A noticeable feature of Figure 6 is that (consistent with them being more skilful) the TIGGE day 15 forecasts of the probability of transition from the mid jet cluster ($P_{M\to S}$, $P_{M\to M}$, $P_{M\to N}$) more closely follow the diagonal than forecasts initialized in other clusters.

Another means of assessing the quality of probabilistic forecasts of categorical events is the Receiver Operating Characteristic (ROC) curve (Mason, 1982; Buizza and Palmer, 1998). The ROC curve provides a means of assessing the ability of a forecast system to discriminate between the occurrence and non-occurrence of an event that is largely

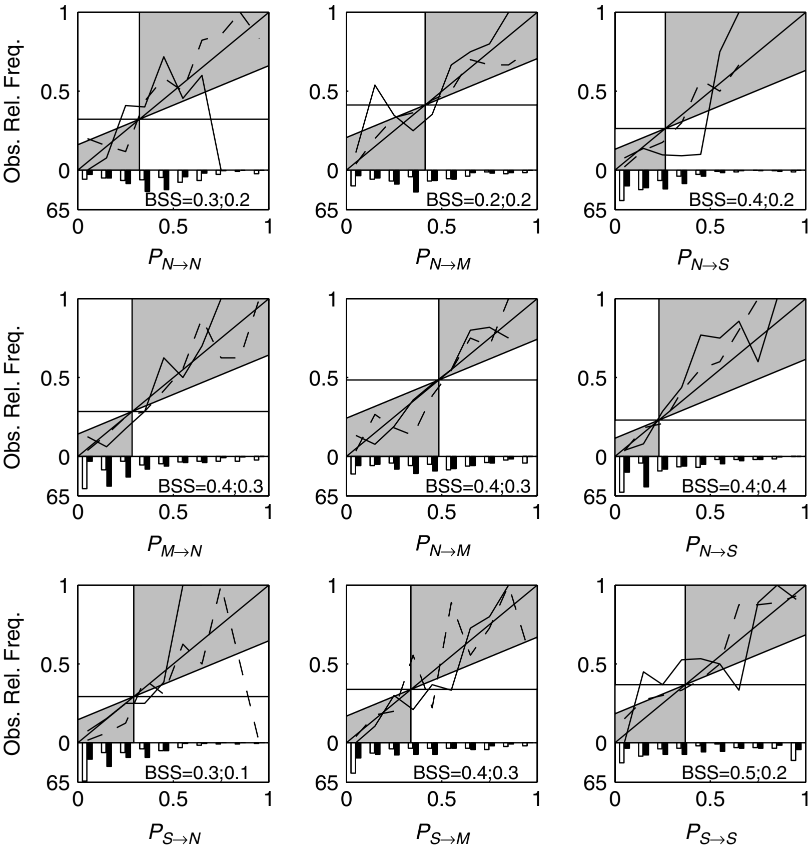

Figure 6. Reliability diagrams (see text) computed for 10 day (dashed line) and 15 day (solid line) lead ECMWF forecasts (ONDJF, October 2007–February 2010). The number of data points in each probability bin is shown in the bar charts beneath each diagram: unshaded bars show 10 day lead-time results, while shaded bars show 15 day results. BSS values shown are ordered as 10 day, 15 day. Grey shading indicates regions that contribute positively to the Brier Skill Score at 15 day lead times.

independent of forecast calibration (Viatcheslav and Zwiers, 2003), i.e. whether the forecast probability matches the observed relative frequency. To calculate a single point of the ROC curve, one selects a probability threshold that the forecast probability of the event must exceed before the event is predicted to occur. The hit rate (HR) and false alarm rate (FAR) for this threshold are then respectively defined as the frequency of occurrence and the frequency of non-occurrence of the event when it is predicted to occur. The ROC curve consists of HR plotted against FAR for all such probability thresholds in a discretization of the range $[0, 1]$. The area under the ROC curve (AUR) is an associated measure of forecast skill, with AUR $= 1$ corresponding to perfect skill and AUR $= 0.5$ corresponding to no skill.

Figure 7 shows the ROC curves calculated for the ECMWF ensemble 10 day (dashed) and 15 day (solid) predicted transition probabilities. The area under the ROC (AUR) is also shown on each panel. Looking first at transitions between the south and north jet clusters (S → N, N → S), it is noticeable that there is a much larger contraction of the area under the ROC from between $\tau = 10$ and $\tau = 15$ than for the other transitions. Consistent with Figures 5 and 6, the ROC curves and AUR values for forecasts initialized in the mid jet cluster and for S → M demonstrate a markedly smaller reduction in skill between the between days 10 and 15. As with the Brier Skill Score, the area under the ROC is smaller at day 15 for transitions originating in the north jet cluster (N → S, N → M, N → N) and S → N than other transitions.

5. Summary and conclusions

This article addresses the question of whether medium-range ensemble forecasts are consistent with and able to predict the transition probabilities associated with a low-dimensional cluster model of the North Atlantic eddy-driven jet. The jet is partitioned into three clusters: a mid jet cluster, which has been interpreted by Woollings *et al.* (2010a) as an undisturbed jet, and two clusters representing southward and northward shifts of the jet. The ability of ensemble forecasts from the TIGGE archive created in three forecasting centres (ECMWF, Met Office, CMC) during the period October 2007–February 2010 to recreate the observed transition probabilities of the three clusters is assessed. No evidence is found that the TIGGE ensemble forecast transition probabilities drift towards values inconsistent with climatological values calculated from ERA40 data. Furthermore it is found that the TIGGE forecast transition probabilities from all forecasting centres possessed significant skill out to 15 day lead times.

For the forecasts in the TIGGE dataset, probabilistic forecasts initialized in the north jet cluster are found generally to have lower day 15 Brier Skill than those initialized in the south and mid clusters. One exception is the prediction of transition from south to north clusters, which is also found to have lower day 15 Brier Skill. Forecasts initialized from the mid jet cluster are found to have the highest day 15 Brier Skill. Similar results are found for the area under the ROC. These results may point to generally lower predictability of the north jet cluster; however, due to the long time-scales associated with the clusters and

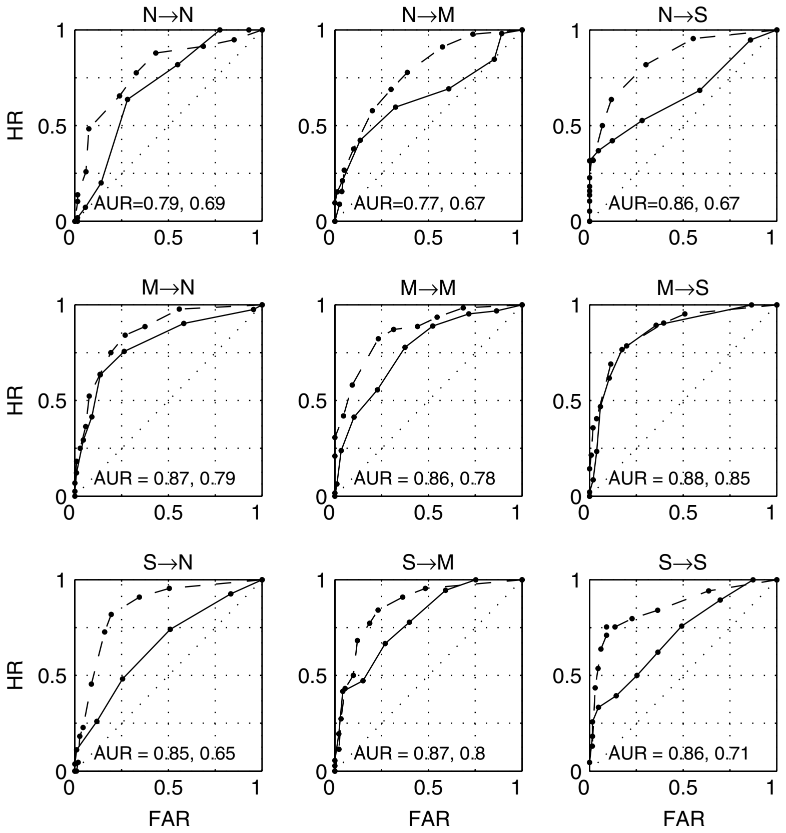

Figure 7. ROC curves computed for 10 (dashed line) and 15 (solid line) day lead ECMWF forecasts (ONDJF, October 2007–February 2010). AUR values shown are ordered as 10 day, 15 day.

the relatively short duration of the three-extended-winter forecast sample provided by the TIGGE dataset, one must be cautious when generalizing the results. Future studies into the predictability of atmospheric regime-like behaviour will certainly benefit from having longer forecast datasets available.

Acknowledgements

The authors gratefully acknowledge the help of the European Centre for Medium-Range Weather Forecasts for providing access to the TIGGE dataset. This work was supported via the National Centre for Atmospheric Science–Weather directorate, a collaborative centre of the Natural Environment Research Council. This article benefited from the suggestions of two anonymous reviewers.

References

Ambaum MHP. 2008. Unimodality of wave amplitude in the Northern Hemisphere. *J. Atmos. Sci.* **65**: 1077–1086.

Ambaum MHP, Hoskins BJ. 2002. The NAO troposphere–stratosphere connection. *J. Climate* **15**: 1969–1978.

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**: 1–3.

Buizza R, Palmer TN. 1998. Impact of ensemble size on ensemble prediction. *Mon. Weather Rev.* **126**: 2503–2518.

Cassou C. 2008. Intraseasonal interaction between the Madden–Julian Oscillation and the North Atlantic Oscillation. *Nature* **455**: 523–527.

Cattiaux J, Vautard R, Cassou C, Yiou P, Masson-Delmotte V, Codron F. 2010. Winter 2010 in Europe: A cold extreme in a warming climate. *J. Geophys. Res.* **37**: L20704. DOI: 10.1029/2010GL044613.

Charney JG, DeVore JG. 1979. Multiple flow equilibria in the atmosphere and blocking. *J. Atmos. Sci.* **36**: 1205–1216.

Christiansen B. 2007. Atmospheric regimes: can cluster analysis provide the number? *J. Climate* **20**: 2229–2250.

Corti S, Molteni F, Palmer TN. 1999. Signiture of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature* **398**: 799–802.

Crommelin DT. 2003. Regime transitions and heteroclinic connections in a barotropic atmosphere. *J. Atmos. Sci.* **60**: 229–246.

Ferro CAT. 2007. Comparing probabilistic systems with the Brier Score. *Weather and Forecasting* **22**: 1076–1088.

Franzke C, Horenko I, Majda AJ, Klein R. 2009. Systematic metastable atmospheric regime identification in an AGCM. *J. Atmos. Sci.* **66**: 1997–2012. DOI: 10.1175/2009JAS2939.1.

Hannachi A. 2007. Tropospheric planetary wave dynamics and mixture modeling: Two preferred regimes and a regime shift. *J. Atmos. Sci.* **64**: 3521–3541.

Hess P, Brezowsky H. 1952. Katalog der Grosswetterlagen Europas. *Ber. Dt. Wetterdienstes in der US-Zone* **33**: 39.

Hoskins BJ, James IN, White GH. 1983. The shape, propogation and mean-flow interaction of large-scale weather systems. *J. Atmos. Sci.* **40**: 1595–1612.

Hsu HR, Murphy AH. 1986. The attributes diagram: A geometrical frmework for assessing the quality of probability forecasts. *Int. J. Forecasting* **2**: 285–293.

Jain AK. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Lett.* **31**: 651–666.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iridell M, Saha S, White G, Woolen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropolewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D. 1996. The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Am. Meteorol. Soc.* **77**: 437–471.

Keeley SPE, Sutton RT, Shaffrey LC. 2009. Does the North Atlantic Oscillation show unusual persistence on intraseasonal timescales? *Geophys. Res. Lett.* **36**: L22706. DOI: 10.1029/2009GL040367.

Lorenz EN. 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**: 130–141.

Mason I. 1982. A model for assessment of weather forecasts. *Aust. Meteorol. Mag.* **30**: 291–303.

Murphy AH, Winkler RL. 1977. Reliability of subjective probability forecasts of precipitation and temperature. *Appl. Stat.* **26**: 41–47.

Park YY, Buizza R, Leutbecher M. 2008. TIGGE: Preliminary results on comparing and combining ensembles. *Q. J. R. Meteorol. Soc.* **134**: 2029–2050. DOI: 10.1002/qj.334.

Silverman BW. 1981. Using kernel density estimates to investigate multimodality. *J. R. Statist. Soc.* **43**: 97–99.

Smyth P, Ide K, Ghil M. 1999. Multiple regimes in northern hemisphere height fields via mixture model clustering. *J. Atmos. Sci.* **56**: 3704–3723.

Stephenson DB, Hannachi A, O'Neill A. 2004. On the existence of multiple regimes. *Q. J. R. Meteorol. Soc.* **130**: 583–605. DOI: 10.1256/qj.02.146.

Uppala SM, Kallberg PW, Simmons AJ, Andrae U, da Costa Bechtold V, Fiorino M, Gibson JK, Haseler J, Hernandez A, Kelly GA. 2005. The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131**: 2961–3012.

Viatcheslav VK, Zwiers FW. 2003. On the ROC score of probability forecasts. *J. Climate* **16**: 4145–4150.

Woollings TJ, Hannachi A, Hoskins BJ. 2010a. Variability of the North Atlantic eddy-driven jet. *Q. J. R. Meteorol. Soc.* **136**: 856–868.

Woollings TJ, Hannachi A, Hoskins BJ, Turner A. 2010b.. A regime view of the North Atlantic Oscillation and its response to anthropogenic forcing. *J. Climate* **23**: 1291–1307.