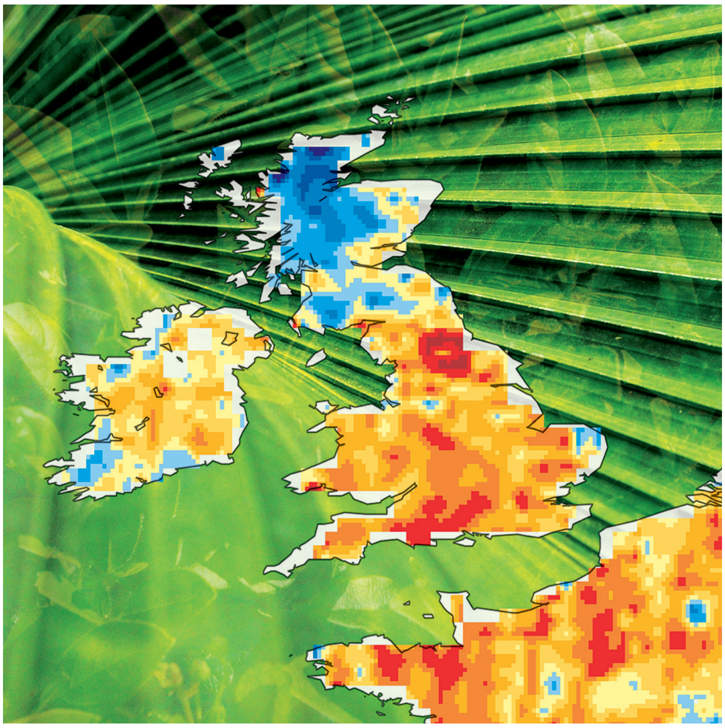
ECMWF Feature article

from Newsletter Number 127 – Spring 2011

METEOROLOGY

New clustering products



www.ecmwf.int/en/about/news-centre/media-resources

doi:10.21957/lr3bcise

This article appeared in the Meteorology section of ECMWF Newsletter No. 127 – Spring 2011, pp. 6–11.

New clustering products

Laura Ferranti, Susanna Corti

The ECMWF clustering is one of a range of products that summarise the large amount of information in the Ensemble Prediction System (EPS). The clustering gives an overview of the different synoptic flow patterns in the EPS. Based on the similarity between their 500 hPa geopotential fields over the North Atlantic and Europe, the members are grouped together.

EPS cluster products have been produced operationally since 1992. In recent years, due to the continuous improvements of the EPS (in particular reduced spread, consistent with decreasing ensemble mean error), these products only occasionally produced more than one cluster. The requirement for cluster products was recently reviewed with ECMWF's Member and Co-operating States, particularly during the annual Forecast Products Users' Meeting. Although some countries do their own clustering (for specific parameters or areas of interest), there was a clear requirement for ECMWF to continue providing a general cluster product from the EPS. Therefore, based on the feedback from the Member and Co-operating States, a new EPS clustering application was developed. The new clustering was endorsed by the TAC Subgroup on Verification Measures as part of their review of product development and user requirements.

The new system includes two components:

- A daily clustering of the forecast fields from the EPS, similar in principle to the original EPS clustering but using a different algorithm.
- A set of four fixed climatological regimes.

The daily clustering summarises the range of synoptic flow patterns in the current EPS. Each cluster is represented by the EPS member closest to its centre, referred to as the 'EPS scenario' for that cluster. Each EPS scenario is then attributed to one of the four climatological regimes. This shows the differences between scenarios in terms of the large-scale flow and provides information about the possible transitions between regimes during the forecast. This approach also enables the development of flow-dependent skill measures.

The new cluster products were implemented operationally in November 2010. This article describes the new clustering, introduces the new cluster products and provides information on how to use them. Validation of the new clustering is also addressed.

The new EPS clustering

The clustering algorithm takes the 51 forecasts (50 perturbed plus 1 control forecast) and groups together those that show a similar evolution of the 500 hPa geopotential over the North Atlantic and Europe (75°N–30°N, 20°W–40°E). For two EPS members to join the same cluster they must display similar synoptic development at 500 hPa throughout a given time window. Clustering in this way, rather than on individual forecast days, has the advantage that temporal continuity and synoptic consistency are retained. The clustering is made independently for four time windows: 72–96, 120–168, 192–240 and 264–360 hour forecast ranges.

The number of clusters can vary from case to case. In some cases the EPS will contain a number of well-separated groups of similar forecasts (a so called multimodal distribution). In other cases the EPS members will be rather more evenly spread out, with no clear grouping into separate clusters (a so called unimodal distribution): there is in effect just a single cluster containing all ensemble members. Since the clustering is intended as a summary of the ensemble information, the maximum number of clusters is limited; the maximum is six as in the previous clustering.

Details of the procedure to compute the clusters are given in the Appendix.

doi:10.21957/lr3bcise

Large-scale climatological regimes

To put the daily clustering in the context of the large-scale flow and to allow the investigation of regime changes, the new ECMWF clustering contains a second component. Each cluster is attributed to one of a set of four pre-defined climatological regimes:

- Positive phase of the North Atlantic Oscillation (NAO).
- Euro-Atlantic blocking.
- Negative phase of the North Atlantic Oscillation (NAO).
- Atlantic ridge.

The climatological regimes have been computed from 29 years of reanalysis data (ERA-Interim and ERA-40) using the same clustering algorithm as for the EPS scenarios (see the Appendix). They are consistent with those documented in the literature (see, for example, *Michelangeli et al.*, 1995).

Figure 1 shows the climatological regimes computed for the cold season (October to April). The climatological regimes for the warm season (May to September) have very similar patterns, but with lower amplitude. To account for this seasonal evolution, in the classification of the EPS scenarios the patterns and amplitudes of the climatological regimes are adjusted month by month. A pattern-matching algorithm assigns each EPS scenario to the closest climatological weather regime (in terms of the root mean square difference).

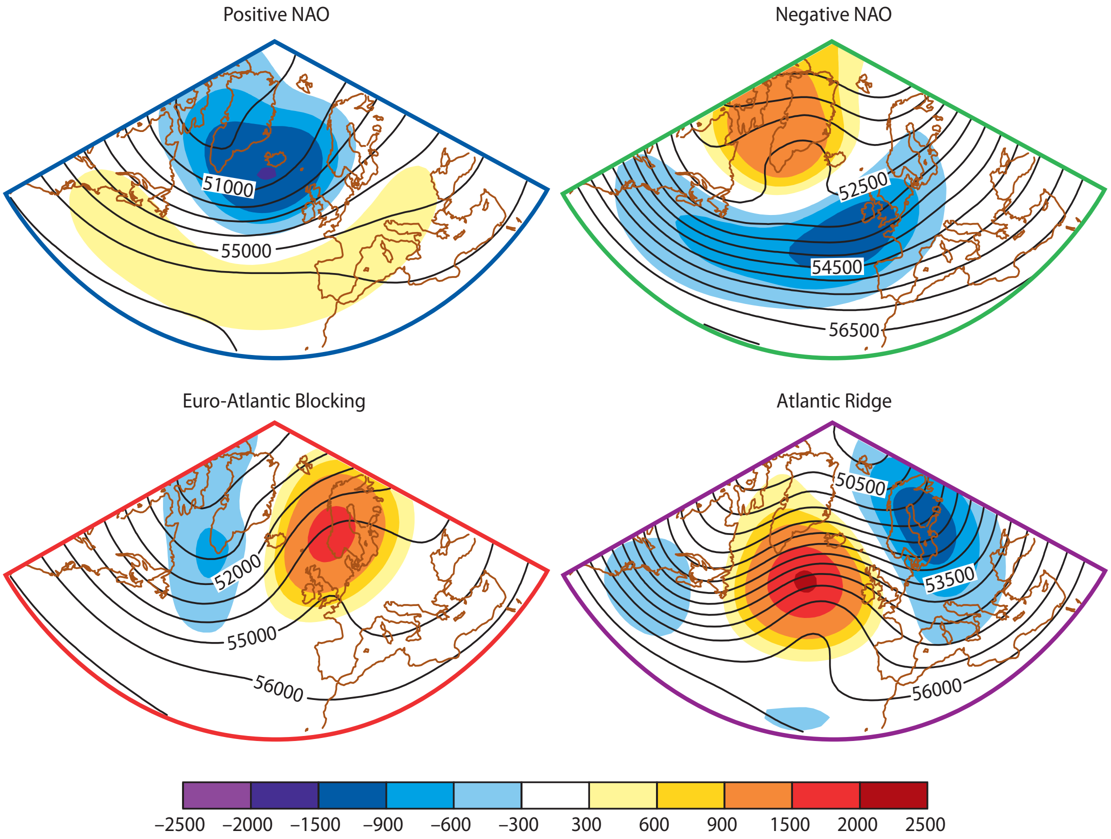


Figure 1 Geographical patterns of the four Euro-Atlantic climatological regimes (both anomalies and full fields) associated with the cold season climatological regimes computed as clusters in the phase-space spanned by the ten leading empirical orthogonal functions (EOFs). The geopotential anomalies (colour shading) and geopotential (contours) at 500 hPa are shown. The corresponding patterns for the warm season are available at http://www.ecmwf.int/products/forecasts/cluster_doc/era_cl4_mjjas_1980-2008.gif

doi:10.21957/lr3bcise

Using the new cluster products

The new cluster products are archived in MARS and available to forecast users through the operational dissemination of products. A graphical product using the new clustering is available for registered users on the ECMWF web site:

- http://www.ecmwf.int/products/forecasts/d/guide/medium/eps/newclusters/.

This web product is designed to provide forecasters with an overview of the EPS clusters for the current forecast. An example is shown in Figure 2 for the 120–168 hour (5–7 day) time window for the forecast from 00 UTC on 2 February 2011. There are three clusters with each cluster represented by one of its members: the forecast closest to the centre of the cluster. The representative members of the three clusters are referred to as the 'EPS scenarios'. These three EPS scenarios are shown at the beginning (120 hours, left column), middle (144 hours, centre) and end (168 hours, right) of the forecast time window. The top row shows the EPS scenario for the first cluster: there are 22 members in this cluster and the control forecast (labelled member 0) has been identified as the representative member of that cluster. The second row shows the EPS scenario for the second cluster: member 29, representing the 15 members of this cluster; member 46, the EPS scenario for the 14 members of cluster 3 is shown in the bottom row.

Users need to be careful in interpreting the number of clusters. There is no direct link between the number of clusters and the overall spread of the ensemble. For example, the EPS may contain a large range of solutions (large spread) but without forming distinct clusters. Alternatively there may be several distinct solutions, but all within the same general flow type (so small overall spread). Because the clustering is done separately for each time window, it is quite possible to have more clusters at the short range that at longer range.

The second component of the new clustering, the four fixed climatological weather regimes, is able to provide additional information. Each EPS scenario is attributed to one of the four climatological regimes. This attribution is shown in Figure 2 by the coloured border round each panel; the colours match those of the corresponding regime shown in Figure 1.

The synoptic forecast evolutions shown in Figure 2 can now be seen in the context the underlying large-scale pattern in which the synoptic features of the weather scenarios are embedded. To help with this interpretation, the panels in Figure 2 show both the forecast geopotential and the anomaly (the difference between the forecast and the climatological 500 hPa geopotential fields). At the beginning of the time window (120 hours into the forecast, left panels) all three scenarios are in the positive NAO regime (all have blue frames). 48 hours later (right panels), each EPS scenario has evolved towards a different large-scale flow regime.

- Scenario 1 (top panel), showing reinforced westerly flow crossing the Atlantic, is related to the positive NAO regime.
- Scenario 2 (middle panel), with a deeper low over the Azores and a further reduction of westerlies, exhibits the typical negative NAO circulation pattern.
- Scenario 3 (bottom panel), with an anticyclonic circulation penetrating UK, is consistent with the main features of the Euro-Atlantic blocking.

Since all members are equally likely, the number of members in each EPS cluster provides an indication of the scenario probability (see Box A). Cluster 1 (22 members) has the highest probability, while cluster 2 (15 members) and cluster 3 (14 members) are equally likely. The additional information from the climatological regimes shows that there is a growing uncertainty through the forecast in the large-scale flow pattern. By day 7 (168 hours into the forecast), although the most likely cluster (cluster 1) remains in the positive NAO regime, both the other clusters indicate a change in the large-scale flow. There is some uncertainty about which regime transition will occur, but it is more likely than not that the overall characteristics of the large-scale flow will change.

The web site includes the products equivalent to Figure 2 for each of the four time windows. For additional information, the 1000 hPa geopotential fields are also provided for each EPS scenario to show the corresponding near-surface evolution. The user should bear in mind that the clustering has been made on the 500 hPa fields; if the main focus of the user is the surface fields we suggest the users compute the clusters using surface parameters.

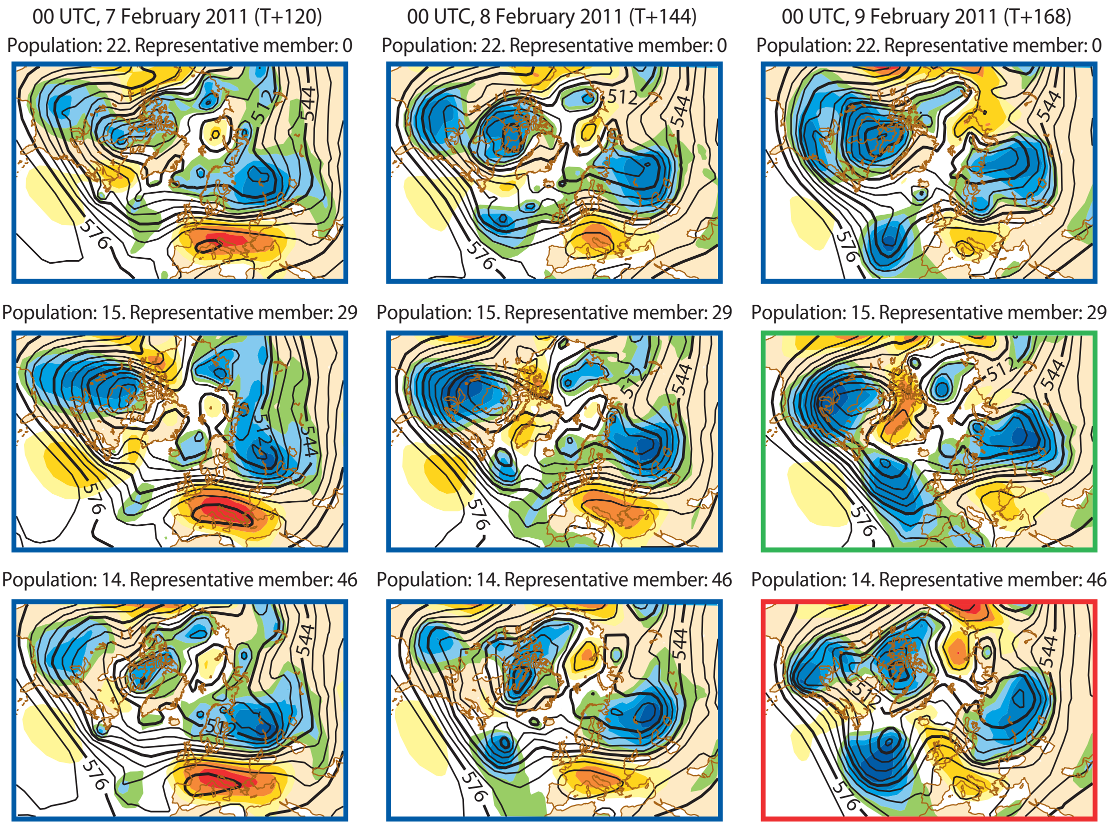

00 UTC, 7 February 2011 (T+120)
00 UTC, 8 February 2011 (T+144)
00 UTC, 9 February 2011 (T+168)
Population: 22. Representative member: 0
Population: 22. Representative member: 0
Population: 22. Representative member: 0
Population: 15. Representative member: 29
Population: 15. Representative member: 29
Population: 15. Representative member: 29
Population: 14. Representative member: 46
Population: 14. Representative member: 46
Population: 14. Representative member: 46
512
544
576

Figure 2 EPS scenarios for time window 120–168 hours for the forecast initiated 2 February 2011. Maps of geopotential at 500 hPa and anomalies from a 29-year reanalysis climate (colour shading: red positive, blue negative). The geopotential field is scaled by 100.

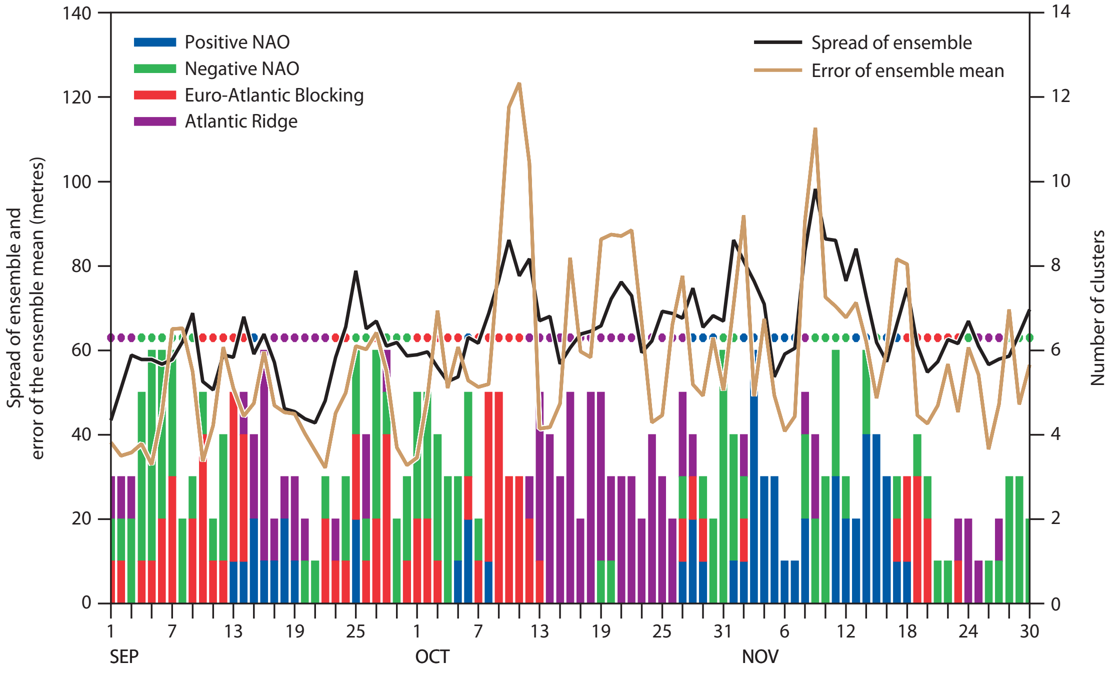


Figure 3 Daily time series for September to November 2010 of the number of EPS scenarios (the colours refer to the patterns shown in Figure 1), spread of the ensemble, error of the ensemble mean, and observed climatological regimes (coloured circles).

doi:10.21957/lr3bcise

Objective validation of the performance of the new products

A

An objective validation of the performance of the new products has been developed. Such evaluation will be routinely updated, including a larger amount of forecast data as it becomes available. The Continuous Ranked Probability Scores (CRPS) evaluates the performance of the EPS as a probabilistic forecast; it is negatively oriented so smaller values indicate better forecast performance. The CRPS computed for a deterministic forecast is the same as the mean absolute error.

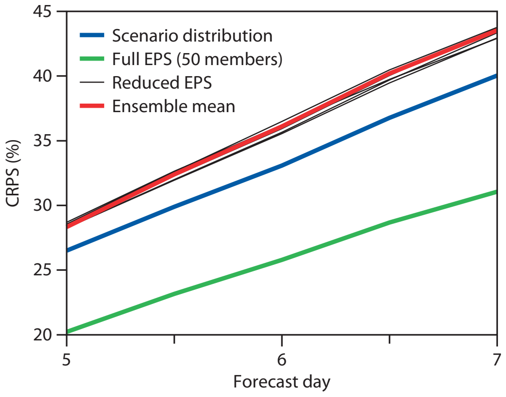
The figure compares the CRPS of the probabilistic forecast based on the full 51-member EPS distribution with the CRPS using probabilities derived from the number of members in each EPS cluster (this is referred to as the 'scenario distribution'). Also shown are the CRPS values for five reduced size ensembles with a maximum of six members and for the ensemble mean.

The scenario distribution is constructed by assuming that each scenario represents perfectly all the other members belonging to that cluster. Consider a case with three clusters with populations of 11, 22 and 18 members. The scenario distribution is then computed taking 11 times scenario 1, 22 times scenario 2 and 18 times scenario 3. The additional reduced size ensembles are constructed by extracting each day a number of random members equal to the number of clusters obtained for that day.

The performance of the scenario distribution is significantly better than that of any random reduced size ensembles. This indicates that the cluster scenarios are better at representing the whole EPS distribution. The results for the ensemble mean, being undistinguishable from the reduced ensemble with six members, indicates that a randomly chosen ensemble with maximum ensemble size of six has a CRPS equivalent to that of the ensemble mean. This implies that an EPS with a maximum of six randomly chosen members does not provide more information in a probabilistic sense than a deterministic forecast represented by the average of 51 members.

It is important to mention that the probabilistic scores depend largely on the ensemble size. The smaller the ensemble size, the more sensitive are the scores. For ensembles larger than 20–25 members the probabilistic scores start to be less responsive to the ensemble size. A large ensemble provides a more detailed and more reliable estimate of the forecast distribution. So it is not surprising that the full EPS provides a better probabilistic forecast than using just the scenarios. However the clustering products represent a compromise between the advantages of condensing forecast information using a few EPS scenarios against the disadvantage of losing information associated with the full 51 EPS members. The difference in CRPS between the whole EPS distribution and the CRPS of the scenario distribution reminds the users of the extent of such a compromise.

Similar skill estimates to those shown in the figure, but calculated by considering only cases when a selected climatological regime is observed, have also been calculated. The results from this flow-dependent verification are not shown here since currently the amount of data is not considered sufficient for a robust analysis. However, such verification results could be of great value to the users during the formulation of their forecast.



CRPS for the EPS as represented by the scenario distribution, the full 51 member EPS distribution, the five reduced EPS ensembles and ensemble mean. The period considered is January to November 2010.

Use of the climatological regimes in validating the EPS performance

The classification of each EPS scenario in terms of pre-defined climatological regimes provides an objective measure of the differences between scenarios in terms of large-scale flow patterns. This attribution enables flow-dependent verification and a more systematic analysis of EPS performance in predicting regimes transitions.

An example of the use of the climatological regimes in validating the EPS performance is given in Figure 3. For the period September to November 2010, this shows the number of EPS scenarios forecast each day (bars) and their classification (colour coded) with respect to the climatological regimes. The time window is 120–168 hours and the classification is made for the EPS scenarios at forecast range 168 hours. The sequence of coloured circles represents the 'observed' climatological regimes computed from the verifying analysis.

Figure 3 shows three distinct blocking events (red circles), each persisting for about a week (second week of September, beginning of October and late November), two periods with a persistent Atlantic ridge flow pattern (violet, mid September and second half of October) one of which extends up to two weeks, and a number of shorter-term regimes transitions.

During the two weeks of the persistent Atlantic ridge regime (13–26 October) the number of EPS scenarios ranges between 2 and 5, showing that the number of distinct synoptic evolutions within the EPS varies from day to day. However, throughout this period, all but two of the scenarios are attributed to the same Atlantic ridge regime. So the EPS is giving a strong signal that this is a period of enhanced large-scale predictability and transition to a different regime is unlikely. The beginning and end of this persistent period were both well forecast by the EPS. After four days in the blocking regime (7–10 October), over the next two days the EPS showed increasing probability of the change to the Atlantic ridge regime. At the end of October, the breakdown of this regime was also clearly signalled. Unlike the previous transition, the EPS indicated considerable uncertainty about what large-scale would follow this breakdown, and indeed in terms of the large-scale flow the end of October and beginning of November was rather changeable.

This verification shows that, as a general feature, cases where the EPS scenarios are associated with different climatological regimes (bars have several colours) occur at periods when regimes transitions take place. Conversely, the periods with a persistent observed regime are associated with a reduced level of forecast 'large-scale diversity' and consequently an enhanced level of predictability. Overall during the autumn 2010 season the EPS scenarios successfully captured the observed time evolution of the climatological regimes.

Figure 3 also shows the ensemble spread (ensemble standard deviation) and error of the ensemble mean forecast for each case. It is worth noting that the mean of both quantities over the season are comparable. This indicates that the EPS is well constructed and the verifying analysis will generally lie within the range of solutions predicted by the ensemble. However, as discussed in the previous section, Figure 3 shows that the cases with high spread do not always correspond to forecasts with a large number of scenarios.

Summary and future developments

After consultation with ECMWF's Member and Co-operating States, ECMWF has developed a new clustering application for the EPS. The new clustering extends the current clustering by providing additional information about the forecast in terms of large-scale climatological regimes. This gives the potential to prediction transitions between regimes and allows the development of flow-dependent verification of the EPS.

A new web product has been developed. This shows the EPS scenarios (the members representing each cluster) and also indicates the climatological regime associated with each scenario. Thus users are provided with a summary of the range of synoptic developments in the current EPS forecast and complementary information about the likelihood of transitions between climatological regimes. Initial validation results show that the EPS can predict these regime transitions. More objective verification of the regime transitions and of the flow-dependent behaviour of the EPS will be developed.

ECMWF will investigate extending the new clustering to cover the full time range of the monthly forecast, considering for example time windows of 2 to 4 weeks ahead. The new clustering methodology can also be adapted to produce additional products appropriate to the needs of individual users (e.g. different domains, variables and forecast ranges); the software could be implemented locally in Member States. In future it may be possible to provide the clustering as a flexible tool via a web interface to allow the user to make such tailored products.

Appendix. Methodology for identifying EPS scenarios and climatological regimes

The cluster algorithm used to identify both EPS scenarios and climatological regimes is based on the modified K-means method applied in *Straus et al.* (2007). This methodology can be summarized in the following four steps.

Identification of a suitable phase space for the cluster computation. Clustering techniques are effective only if applied in an L-dimensional phase space with $L<N$, where N is the number of elements in the dataset (for the EPS, N=51, the number of members in the ensemble). Because the North Atlantic and Europe area contains many more than 51 grid points, it is first necessary to transform the forecasts to a new much lower dimensional co-ordinate system. This is done using so-called empirical orthogonal functions (EOFs). The clustering is carried out in the reduced phase space defined by the L leading EOFs that explain at least 80% of the total variance of the dataset. The associated principal components (PCs) provide the new coordinates. For the EPS, the clusters are computed in an extended time window (e.g. the 120–168 hour forecast range), so EOFs extended in time are used.

Computation of the optimal partition of the data. For a given number k, the optimum partition of the data into k clusters is found. k members are allocated as (pseudorandom) ‘seed points’. An initial clustering is then made based on the distance from these seeds. The algorithm takes this initial cluster assignment and iteratively changes it by assigning each element to the cluster with the closest centroid, until a ‘stable’ classification is achieved. (A cluster centroid is defined by the average of the PC coordinates of all states that lie in that cluster.)

This process is repeated many times (using different seeds), and for each partition the ratio r^*_k of the variance among cluster centroids (weighted by the cluster population) to the average intra-cluster variance is recorded. The optimal partition is the one that maximises this ratio.

Assessment of the significance of a given k-partition. The goal is to assess the strength of the clustering compared to that expected from an appropriate reference distribution, such as a multi-dimensional Gaussian distribution. In assessing whether the null hypothesis of multi-normality can be rejected, it is necessary to perform Monte-Carlo simulations using a large number M of synthetic data sets. Each synthetic data set has precisely the same size (number of members) as the original data set against which it is compared. They are generated from a series of L dimensional Markov processes whose mean, variance and first-order auto-correlation are obtained from the observed data set. A cluster analysis is performed for each one of the simulated data sets. For each k-partition the ratio r_{mk} of variance among cluster centroids to the average intra-cluster variance is recorded.

Since the synthetic data are assumed to have a unimodal distribution, the proportion P_k of synthetic samples for which $r_{mk} < r^*_k$ is a measure of the significance of the k-cluster partition of the actual data, and $1-P_k$ is the corresponding confidence level for the existence of k clusters.

Choice of the most suitable number of clusters. The need to specify the number of clusters can be a disadvantage of the K-means method if we do not know in advance how many clusters to expect. However, there are three main criteria that can be used to choose the optimal number of clusters: (i) *Significance*: the partition with the highest significance (P_k) with respect to predefined multi-normal distributions (see previous step); (ii) *Reproducibility:* we can use as a measure of reproducibility the ratio of the mean-squared error of best matching cluster centroids from N pairs of randomly chosen half-length data sets from the full original data set. The partition with the best reproducibility (ratio closest to one) will be chosen; (iii) *Consistency:* this can be calculated both with respect to the choice of variable (for example comparing with clusters obtained from different dynamically linked variables) and with respect to the specified domain (test of sensitivities to changing the horizontal or vertical domain).

All three criteria have been used to identify the most suitable number of climatological regimes.

However, for the daily clustering only the statistical significance test is used; due to the limited sample size (51 ensemble members), reproducibility and consistency cannot be properly estimated. The number of clusters is determined as follows.

- If the significance (P_k) of all considered cluster partitions, from 2 to 6 clusters, is below a minimum threshold of 55%, it is assumed that there are no clusters.
- If the minimum significance threshold is achieved, then the partition with the highest significance is chosen.
- If more than one partition has a significance value higher than 95%, the one with the minimum number of clusters is chosen.

An additional criterion is used to evaluate the EPS cluster partition, based on the ratio between the average internal variance of clusters and the mean EPS variance of the season. For each time window all the partitions in which the average internal variance of the clusters is lower than 50% of the mean EPS variance are discarded. This condition limits the occurrence of large numbers of clusters (five or six) and, by taking into consideration that the ensemble spread is a function of forecast range, it adds some consistency in the cluster population between the four forecast time ranges.

However, due to well known features of the K-means methodology (*Michelangeli et al.*, 1995), there can still be too many cases with the maximum number of EPS clusters (here this is set to six). To avoid this, if the six cluster partition is selected and its significance is at least 93%, a check is made for other cluster partitions that have 90% or higher significance: the partition with the fewest clusters that satisfies this significance level is chosen instead of the six cluster partition.

Furthre reading

Michelangeli, P.-A., R. Vautard & **B. Legras,** 1995: Weather regimes: Recurrences and quasi-stationarity. *J. Atmos. Sci.*, **52,** 1237–1256.

Straus, D., M.S. Corti & **F. Molteni,** 2007: Circulation regimes: chaotic variability versus SST-forced predictability. *J. Climate*, **20,** 2251–2272.

European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, England

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error or omission or for loss or damage arising from its use.

doi:10.21957/lr3bcise