Clustering Techniques and their applications at ECMWF

Laura Ferranti

European Centre for Medium-Range Weather Forecasts

Outline

- Cluster analysis - Generalities
- Cluster product at ECMWF
- Predictability of Euro-Atlantic regimes at different forecast ranges
- Flow dependent verification

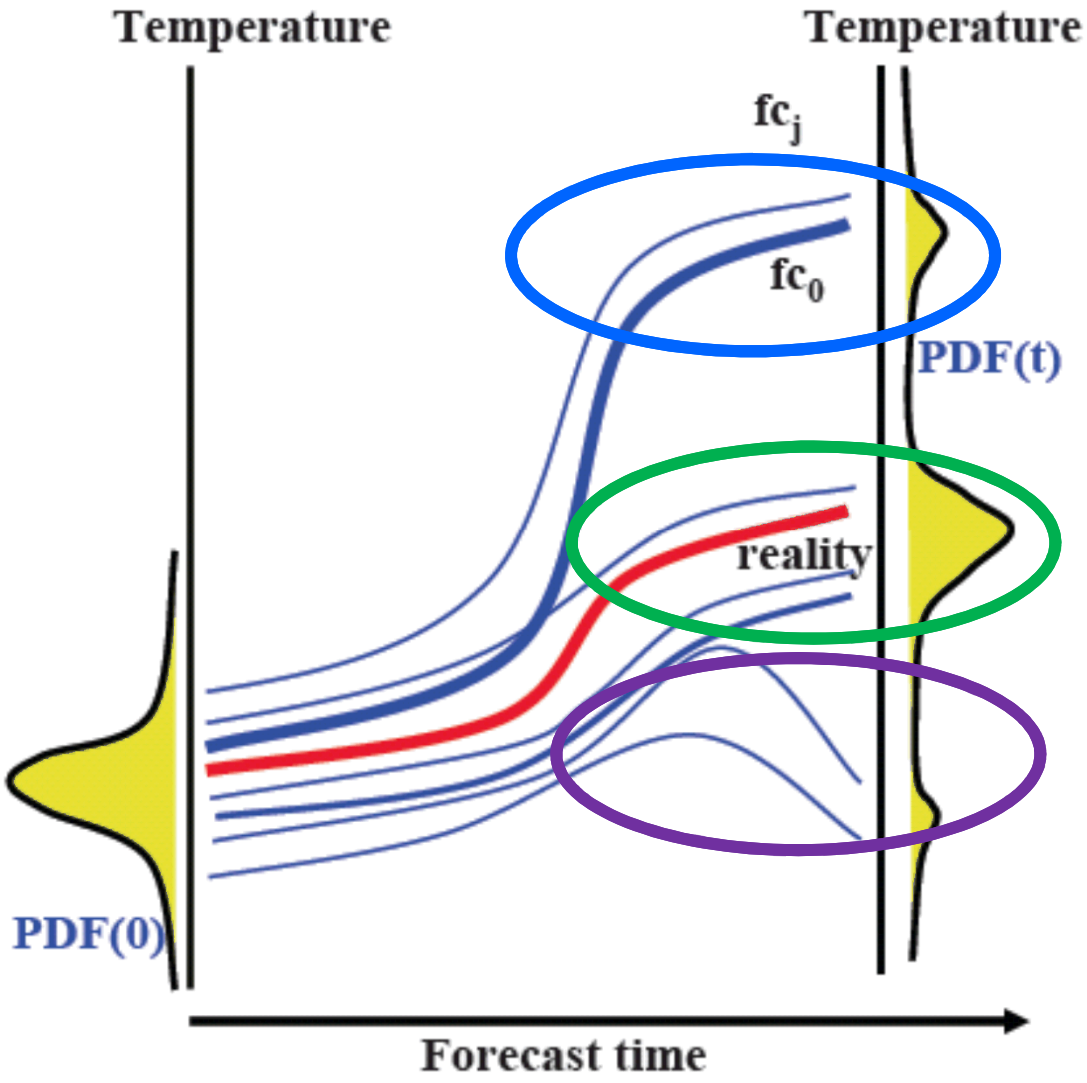
Cluster analysis - Generalities

"Cluster analysis deals with **separating data into groups whose identities are not known in advance**. In general, even the "correct number" of groups into which the data should be sorted is not known in advance." *Daniel S. Wilks*

Examples of use of cluster analysis in weather and climate literature:

- Grouping daily weather observations into synoptic types (Kalkstein et al. 1987)
- Defining weather regimes from upper air flow patterns (Mo and Ghil 1998; Molteni et al. 1990)
- Grouping members of forecast ensembles (Tracton and Kalnay 1993; Molteni et al 1996; Legg et al 2002)

Example – Grouping members of Forecast Ensembles



Clustering analysis - Metrics

"Central to the idea of the clustering of the data points is the idea of distance. Clusters should be composed of points separated by small distances, relative to the distances between clusters." *Daniel S. Wilks*

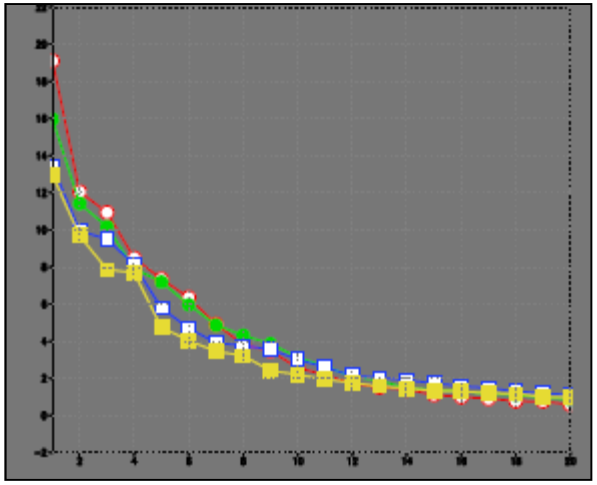
$$d_{i,j} = \left[\sum_{k=1}^{K} w_k (x_{i,k} - x_{j,k})^2 \right]^{1/2}$$

Weighted Euclideian distance between two vectors x_i and x_j

Clustering analysis – Suitable (Sub)spaces of states

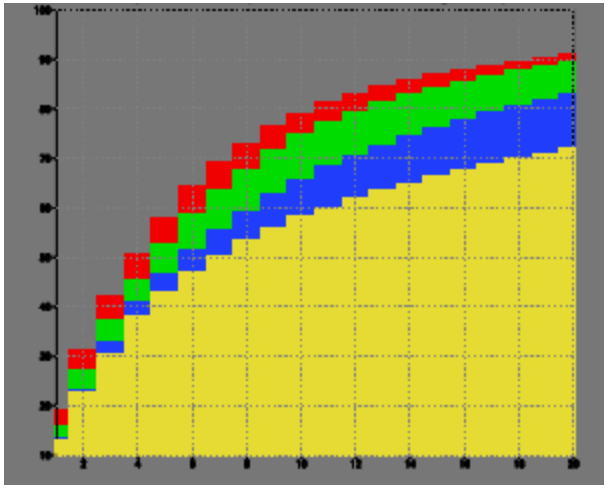
Clustering techniques are effective only if applied in a L-dimensional phase space with L << N (N=number of elements in the data set in question). If the actual space of states is too large (ex: 500 maps with 25x45 grid points) it is advisable to compute the clusters in a suitable sub-space.

EOF decomposition. The first EOF expresses the maximum fraction of the variance of the original data set. The second explains the maximum amount of variance remaining with a function which is orthogonal to the first, and so on. To be useful EOF analysis must result in an decomposition of the data in which a big fraction of the variance is explained by the first few EOFs.



Explained variance

Accumulated variance



Clustering techniques:

- Exclusive Clustering - data are grouped in an exclusive way
- Overlapping Clustering - fuzzy set of clusters data
- Hierarchical Clustering – based on the union between the 2 nearest clusters starting with N clusters for a dataset of N points
- Probabilistic Clustering

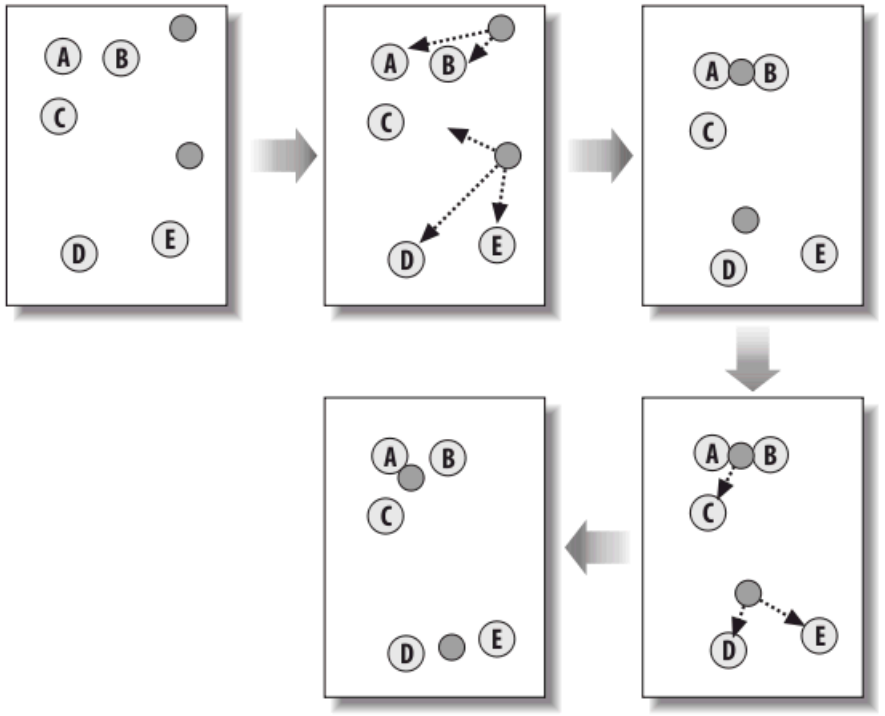
The most widely used exclusive clustering approach is called K-means method. K is the number of clusters into which the data will be grouped (this number must be specified in advance).

Cluster analysis - K-means method

- For a given number k of clusters, the optimum partition of data into k clusters is found by an algorithm that takes an initial cluster assignment (based on the distance from random seed points), 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 variance among cluster centroids (weighted by the population) to the average intra-cluster variance is recorded.

- The partition that maximises this ratio is the optimal one.



Cluster analysis - Significance

The goal is to assess the strength of the clustering compared to that expected from an appropriate reference distribution, such as a multidimensional Gaussian distribution.

- In assessing whether the null hypothesis of multi-normality can be rejected, it is therefore necessary to perform Monte-Carlo simulations using a large number M of synthetic data sets.

- Each synthetic data set has precisely the same length as the original data set against which it is compared, and it is generated from a series of *n* 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 red-noise 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.

Cluster analysis - How many clusters?

The need of specifying the number of clusters can be a disadvantage of K-means method if we don't know in advance what is the best cluster partition of the data set in question. However there are some criteria that can be used to choose the optimal number of clusters.

- Significance: partition with the highest significance with respect to predefined Multinormal distributions
- Reproducibility: We can use as a measure of reproducibility the ratio of the mean-squared error of best matching cluster centroids from a N pairs of randomly chosen half-length datasets from the full actual one. The partition with the highest reproducibility will be chosen.
- Consistency: The consistency can be calculated both with respect to variable (for example comparing clusters obtained from dynamically linked variables) and with respect to domain (test of sensitivities with respect to the lateral or vertical domain).

Cluster product at ECMWF

- 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. The members are grouped together based on the similarity between their 500 hPa geopotential fields over the North Atlantic and Europe.
- These cluster products were implemented in operations in November 2010. They 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/charts/medium/eps/newclusters/newclusters/

Cluster product at ECMWF: large scale climatological regimes

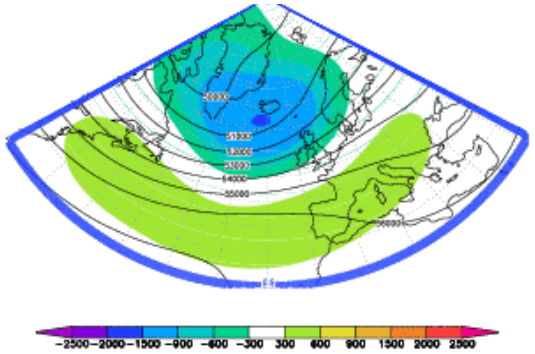
To put the daily clustering in the context of the large-scale flow and to allow the investigation of regime changes, the new ECMWF clustering contains **a second component**. Each cluster is attributed to one of a set of four pre-defined climatological regimes

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

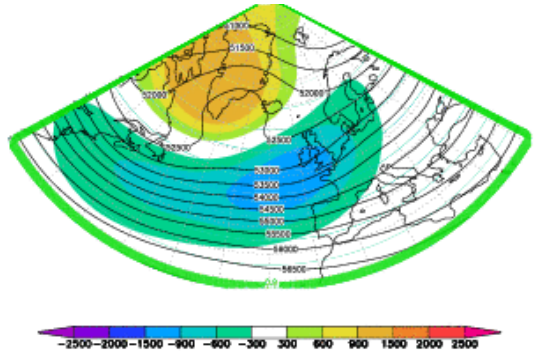
Climatological Regimes in the cold season Euro-Atlantic Region

500 hPa Geopotential height – 29 years of ERA INTERIM ONDJFM

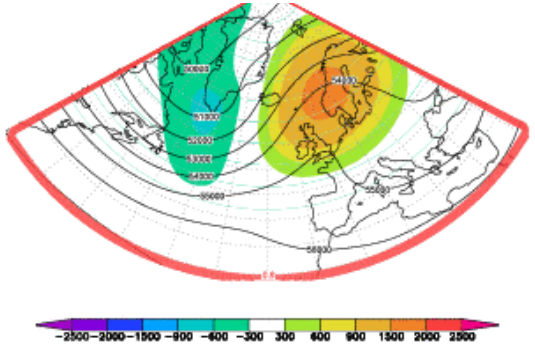
Positive NAO 32.3%



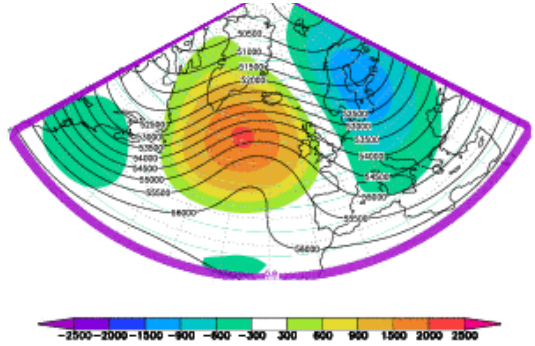
Negative NAO 21.4%



Euro-Atlantic Blocking 26.1%



Atlantic Ridge 20.2%



Cluster product at ECMWF: 2-stage process

1st step: (to be done once per season)

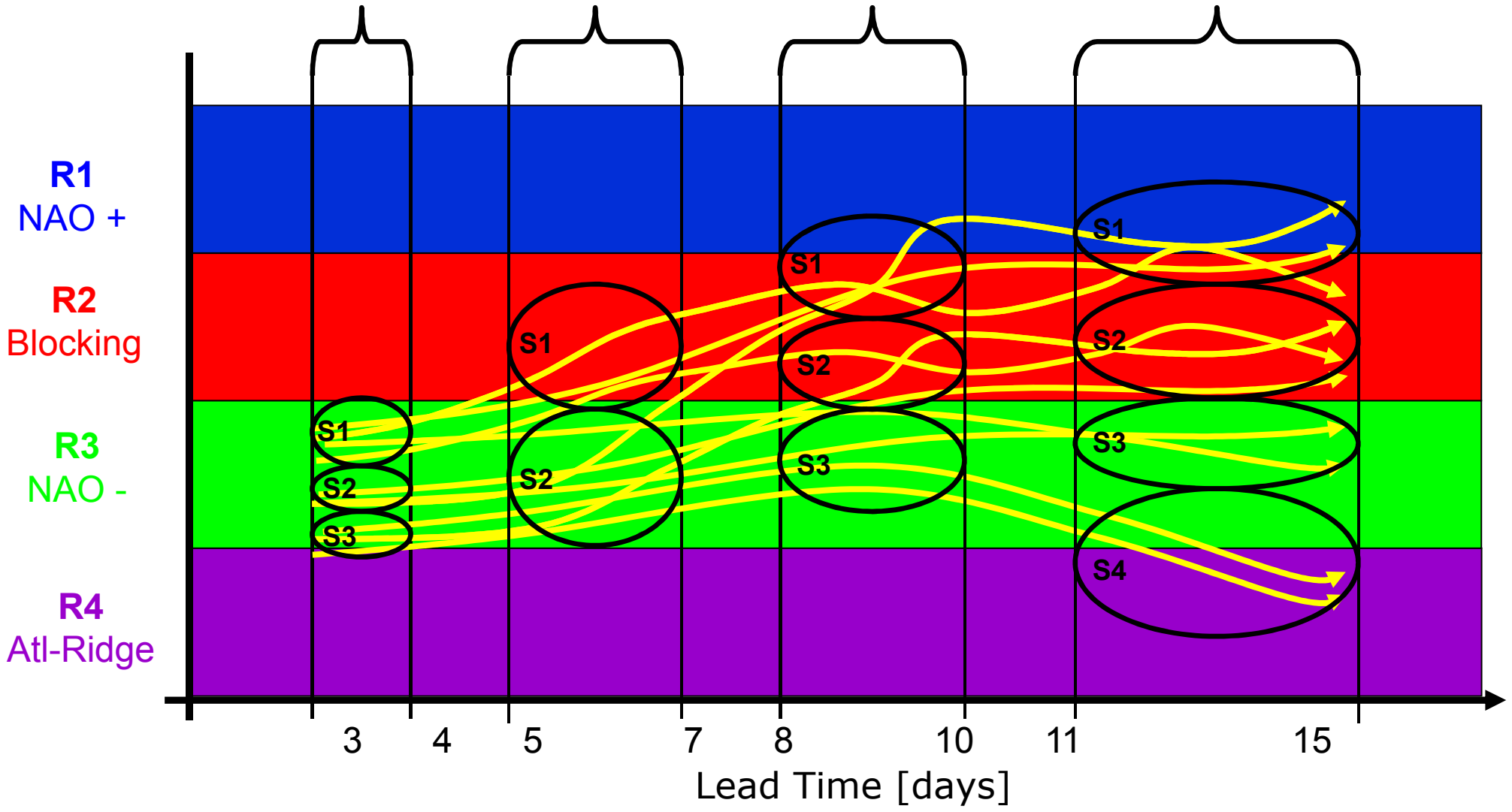
- Identification of the climatological weather regimes over selected regions for every season.

2nd step: (to be done for every forecast)

- Identification of forecast scenarios from the real-time EPS forecasts.

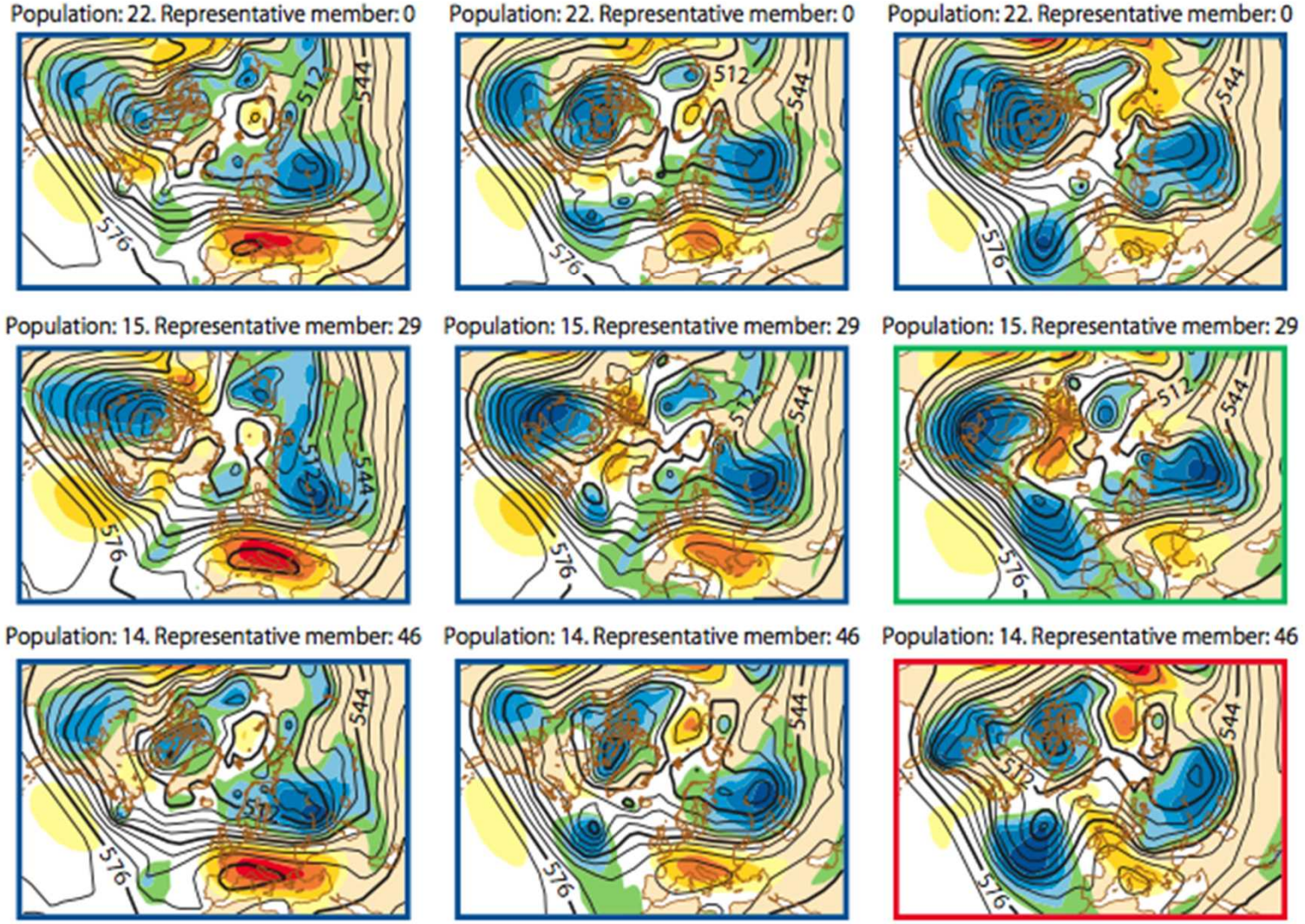
- Association of each forecast scenario to the closest climatological weather regime.

Regimes & Scenarios



Regime transitions within a time window

Day 5 to day 7 - 9 February 2011 – 3 scenarios 2 possible transitions



Show guide

Method

Cluster scenario

Parameter

500

1000

Your Room

Add this product

Show overview

Parameter

Cluster

Forecast base time

Download...

PDF (1.2 Mbytes)

Postscript (1.8 Mbytes)

Cluster scenario

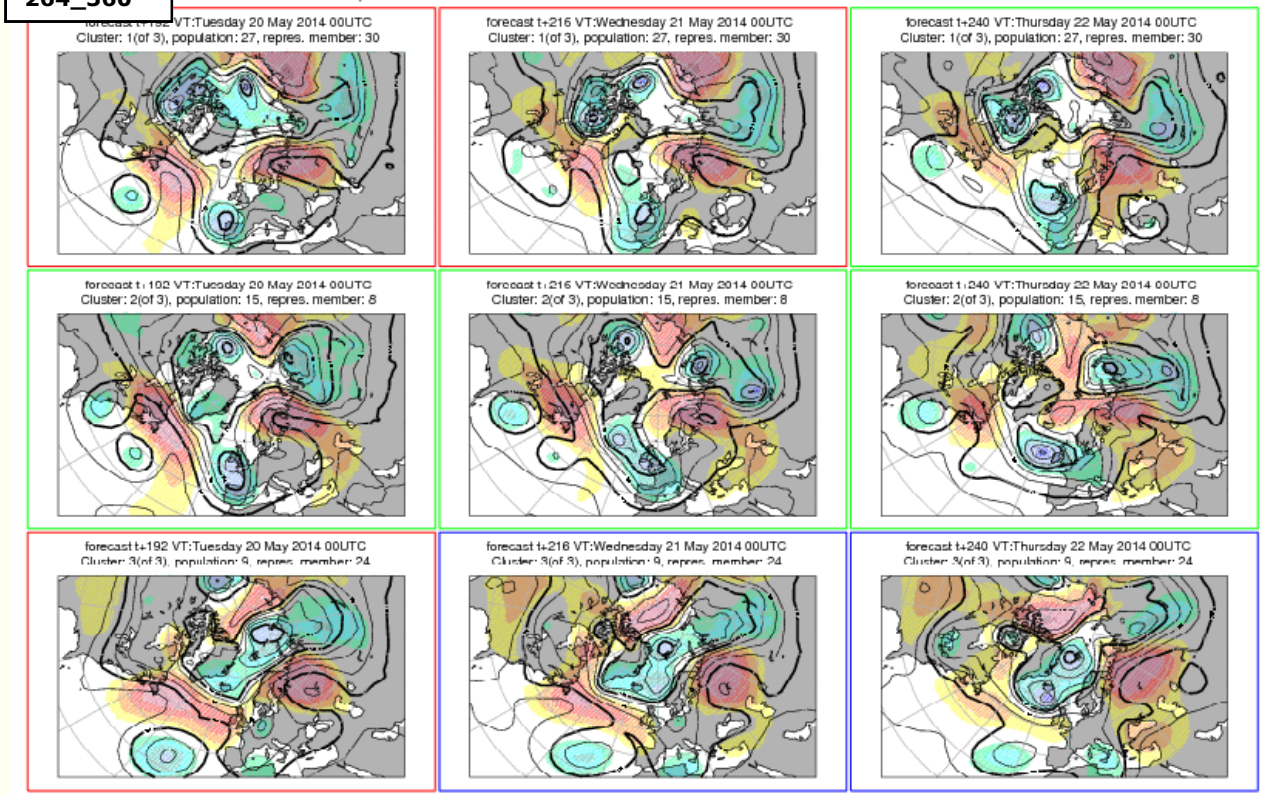
Maps of geopotential height: at 500hPa full field (black contours) and anomalies (colour shading), at 1000hPa full field only. Click on show guide for the full description.

Cluster 192_240

72_96
120_168
192_240
264_360

Forecast base time Mon 12 May 2014 00UTC

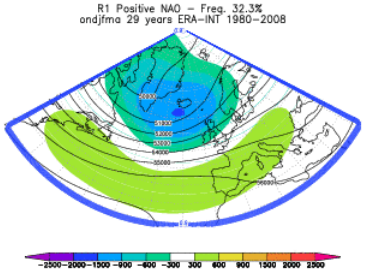
Monday 12 May 2014 00UTC ECMWF EPS Cluster scenario - 500 hPa Geopotential
Reference step t+192-240 Domain 75/340/30/40 Cont. in cluster=1 Det. in cluster=1

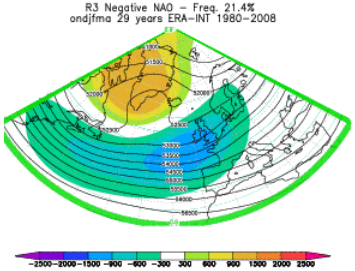


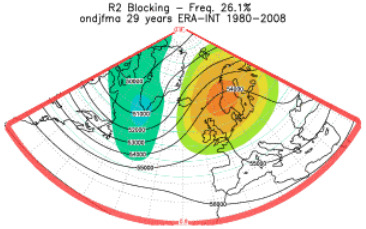
Maps of geopotential height: at 500hPa full field (black contours) and anomalies (colour shading), at 1000hPa full field only. Click on show guide for the full description.

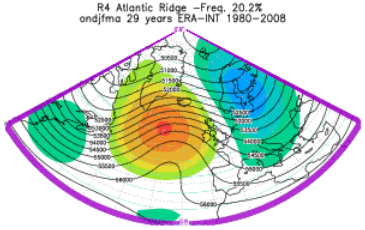
Verification & spread

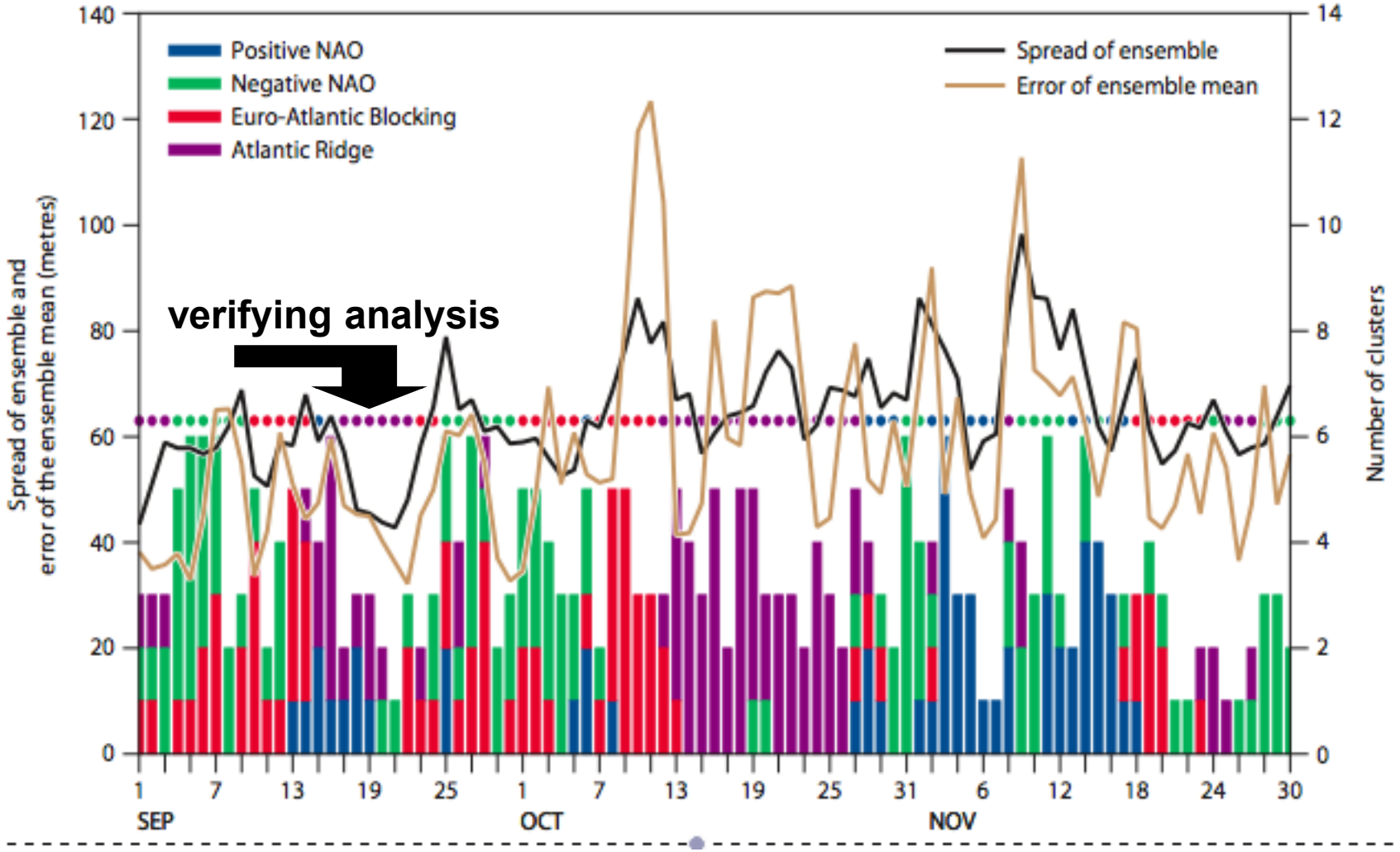
Climatological regimes





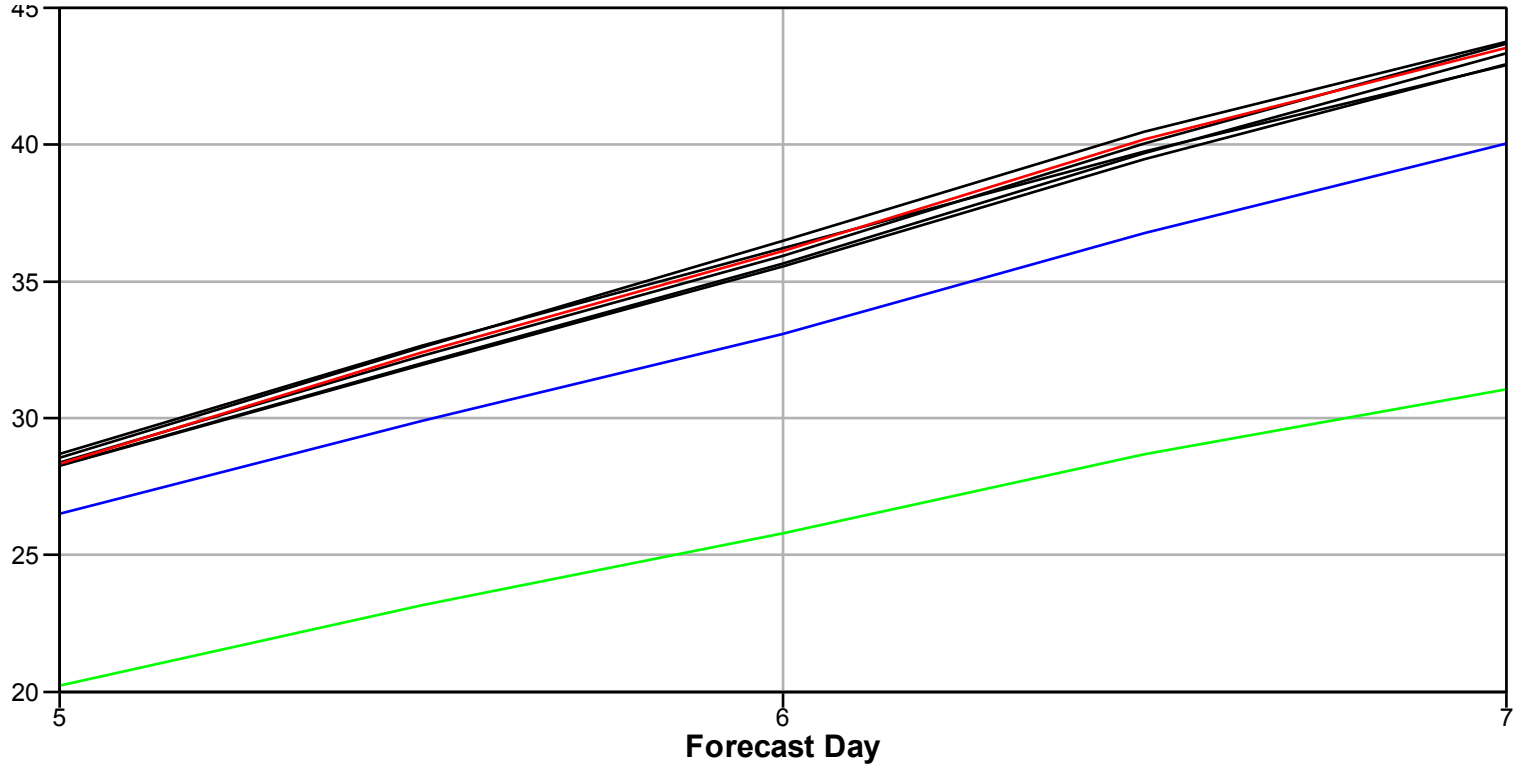






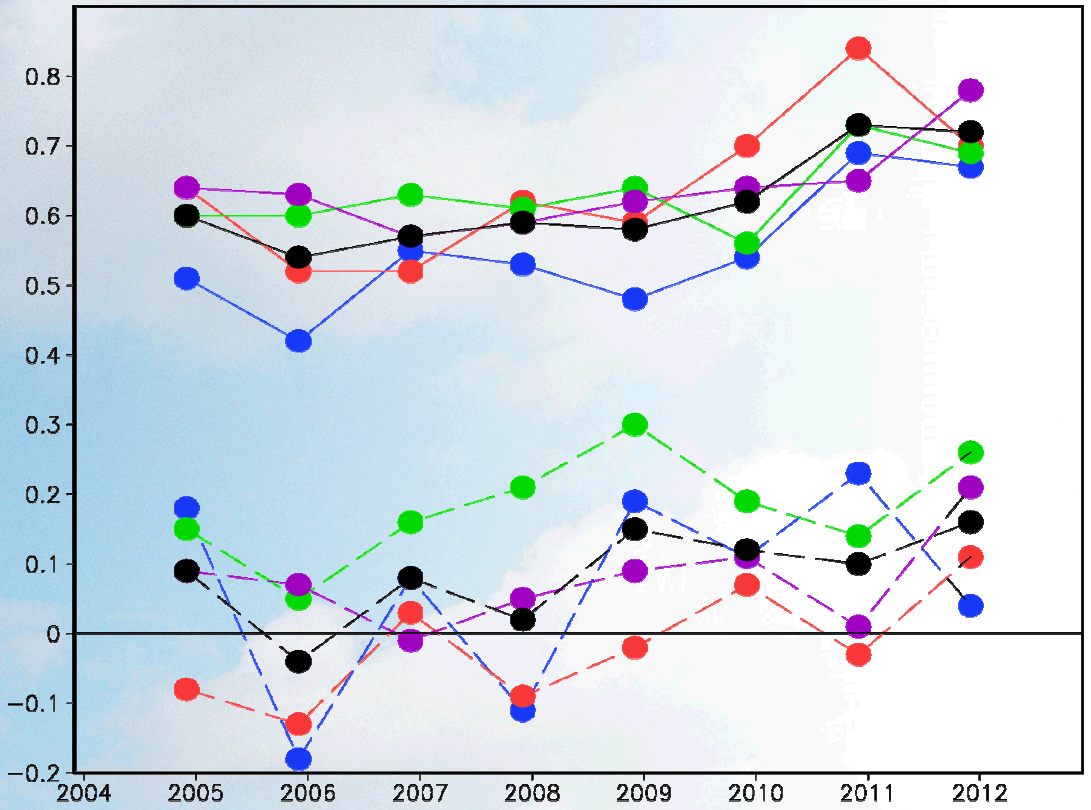
Verification: Continuous Ranked Probability Score (CRPS)

Scenario distribution
Full EPS (50 members)
Reduced EPS
Ensemble Mean



Which flow regime is more predictable ? (medium range)

Evolution of probabilistic skill in predicting the occurrence/not occurrence of the climatological regimes. The black das-dotted lines indicate the scores aggregated over the 4 regimes.



BSS at day 5

BSS at day 7

NAO+

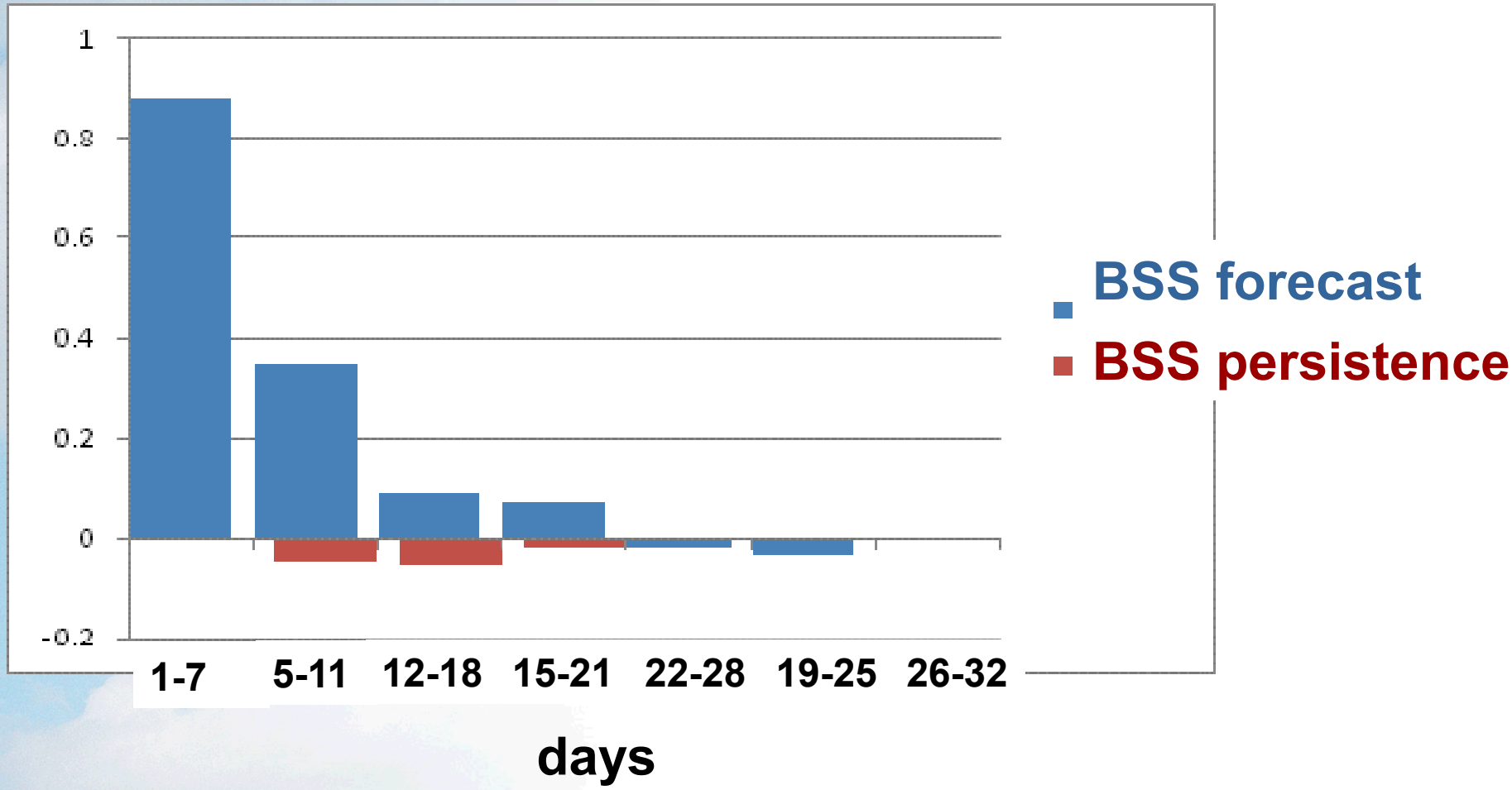
BL

NAO-

AR

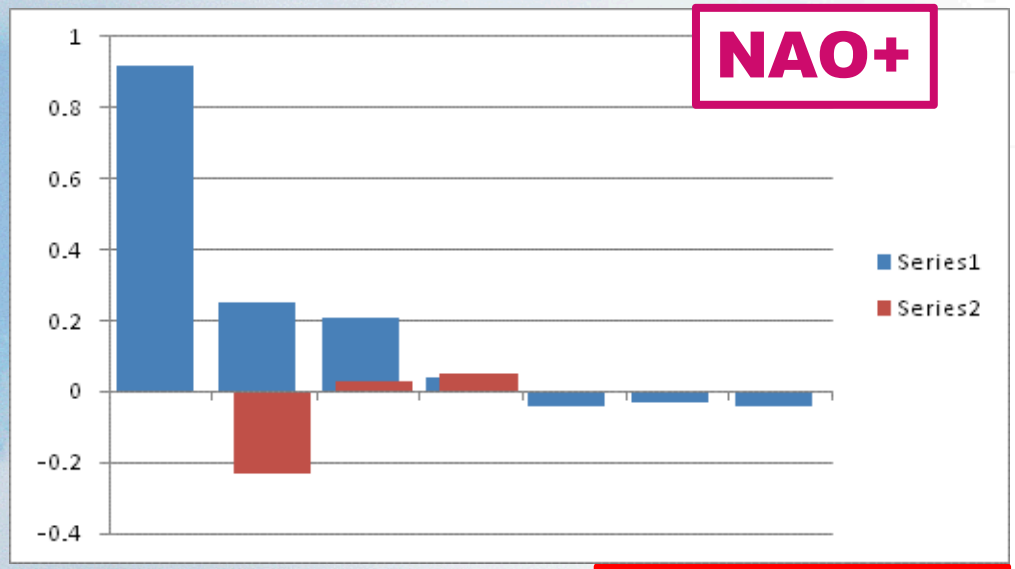
Euro-Atlantic regimes predictability

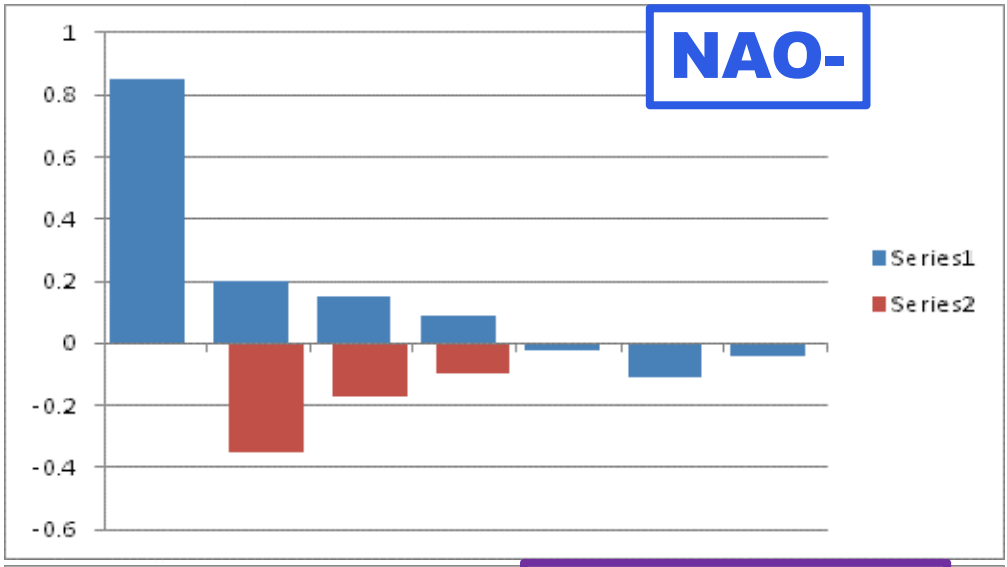
Brier Skill Scores :

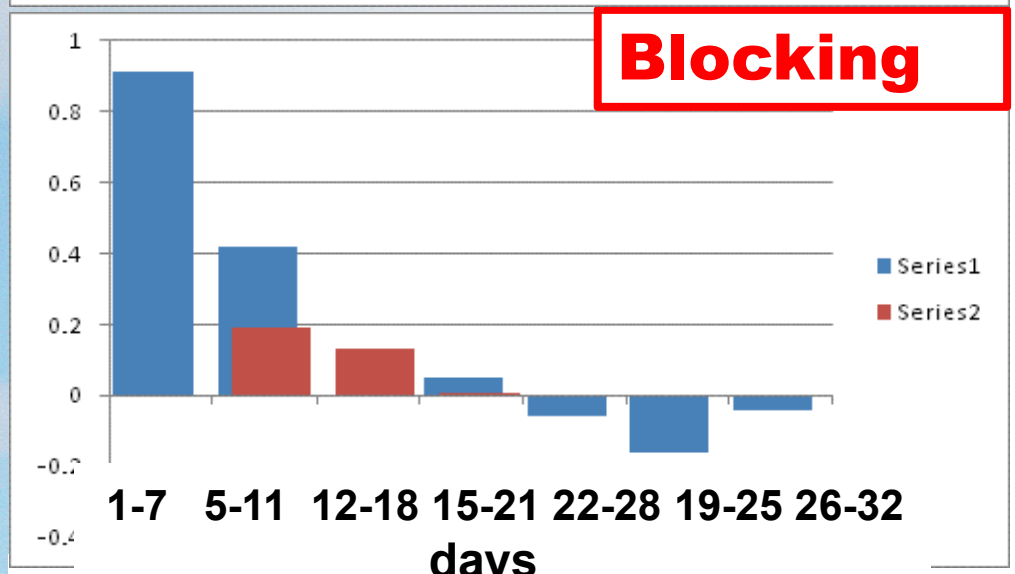


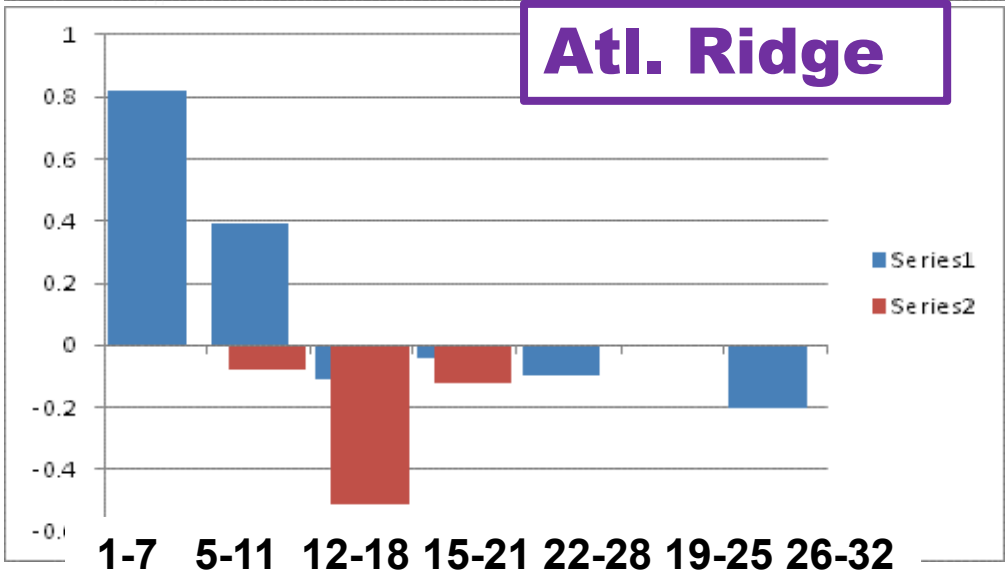
Euro-Atlantic regimes predictability

Brier Skill Scores :







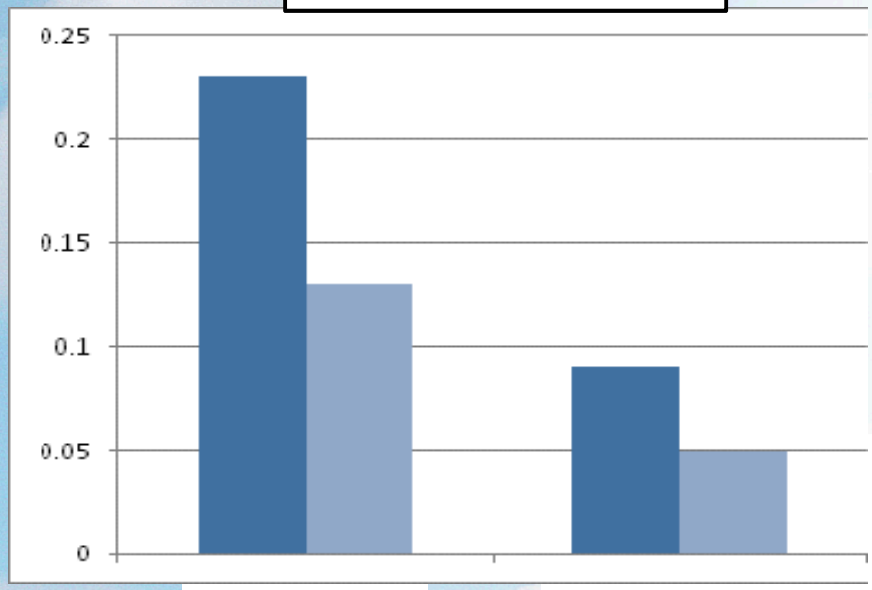


BSS forecast BSS persistence

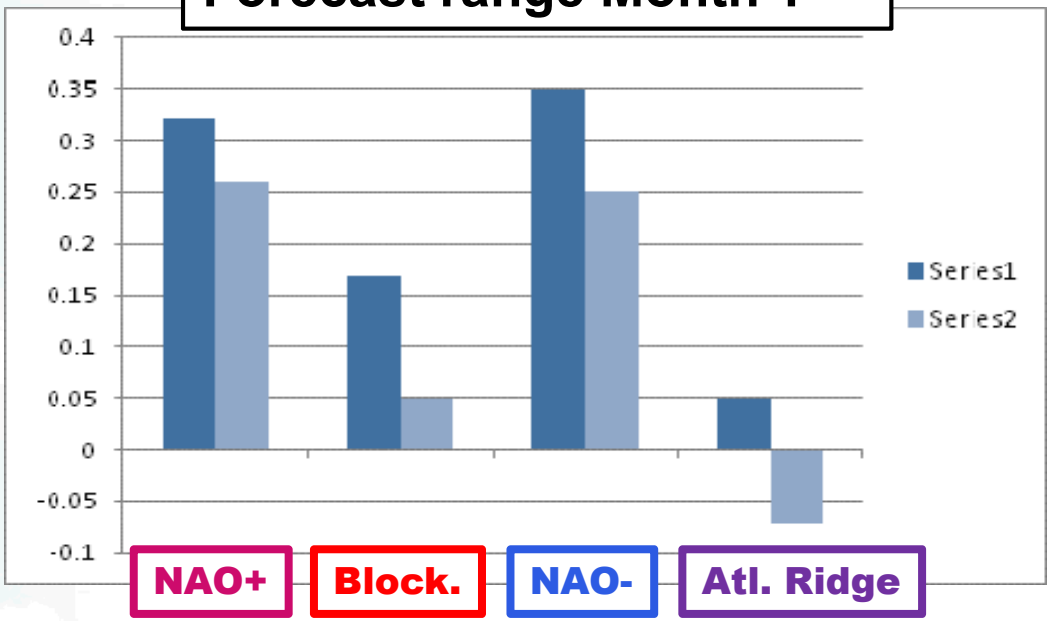
Euro-Atlantic regimes predictability

Brier Skill Scores :

All regimes



Forecast range Month 1

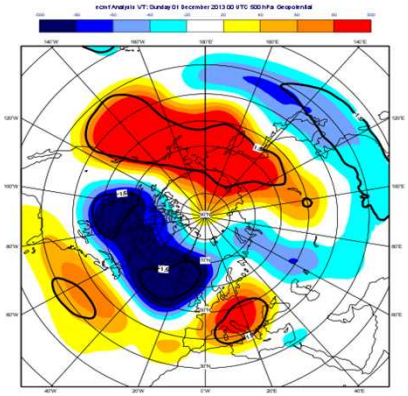


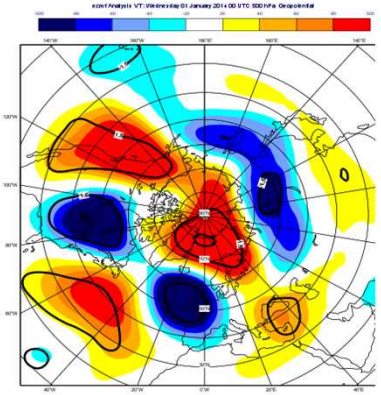
BSS Seasonal for. Sys 4

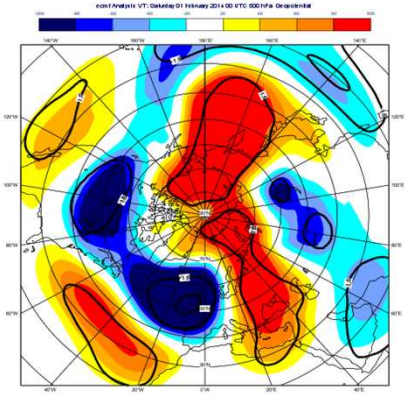
BSS Seasonal for. Sys 3

Regime analysis for DJF 2013-14 :
Monthly means

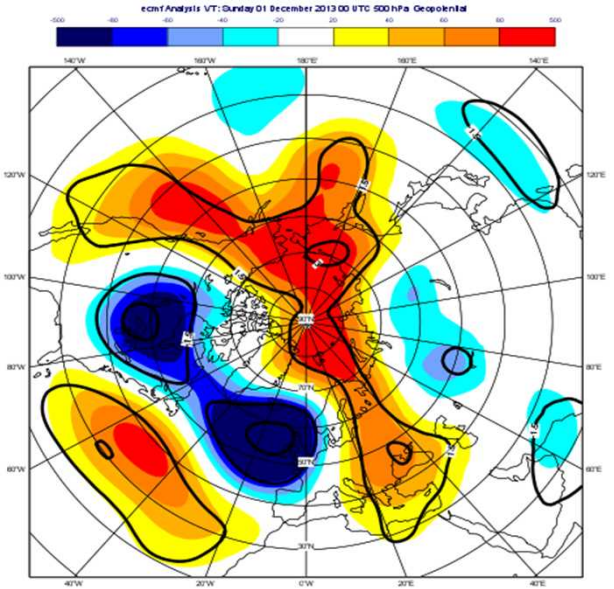
December 2013 January 2014 February 2014

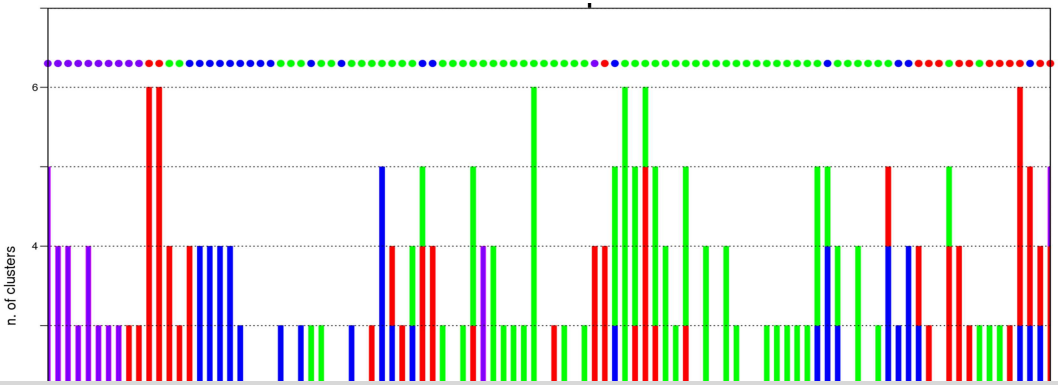


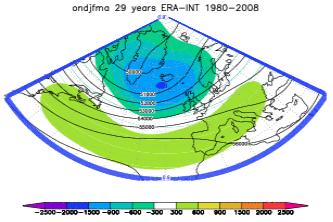


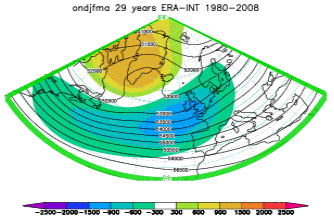


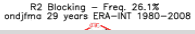
DJF 213-14 anomalies

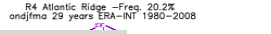


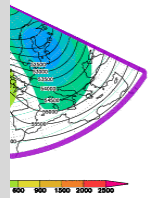








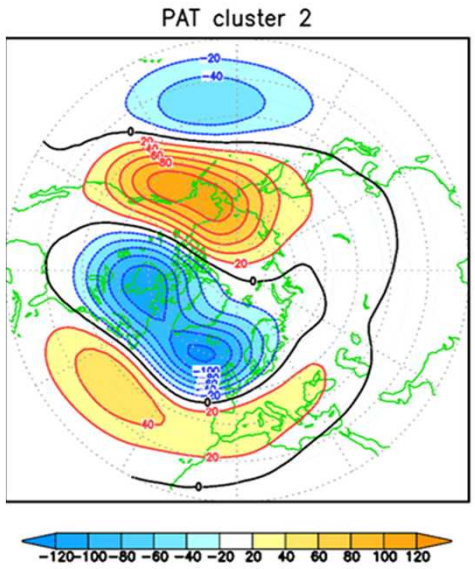


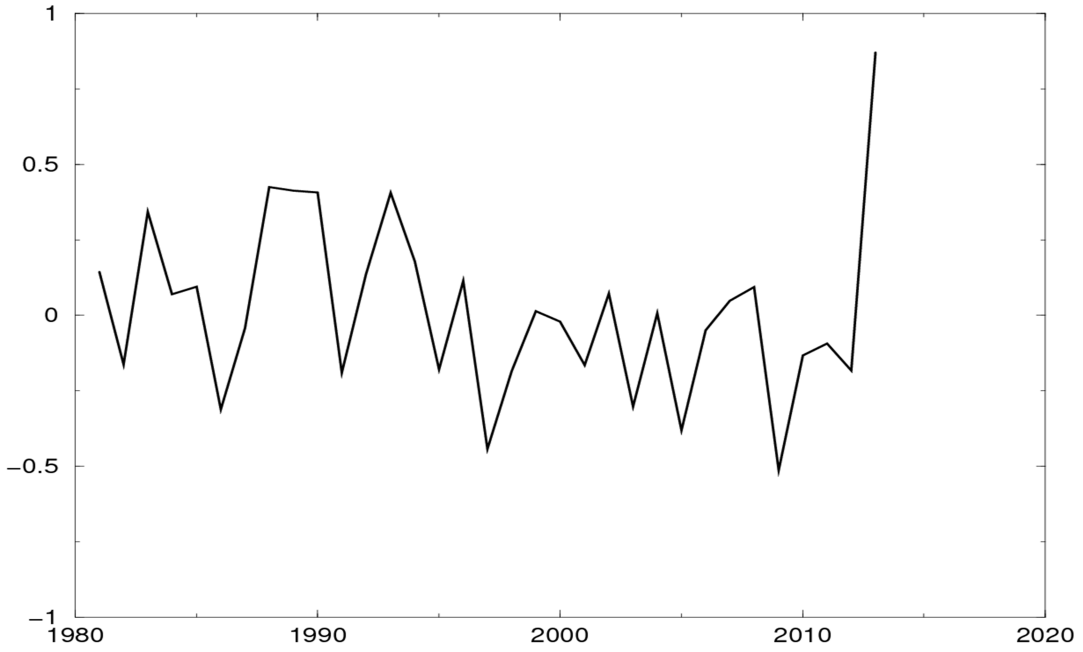


- This winter circulation had a rather hemispheric nature so that it was difficult to describe it by using the 4 climatological Euro-Atlantic regimes

- This winter circulation is well described by the hemisperic regimes (number 2 and 3) see Franco's lecture.

DJF 2013/14 was a record winter: Projections onto NH regime 2

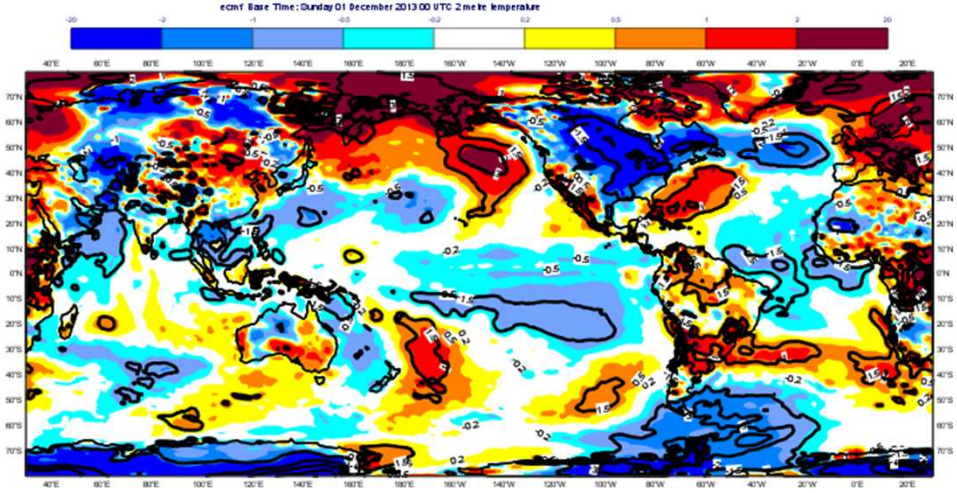


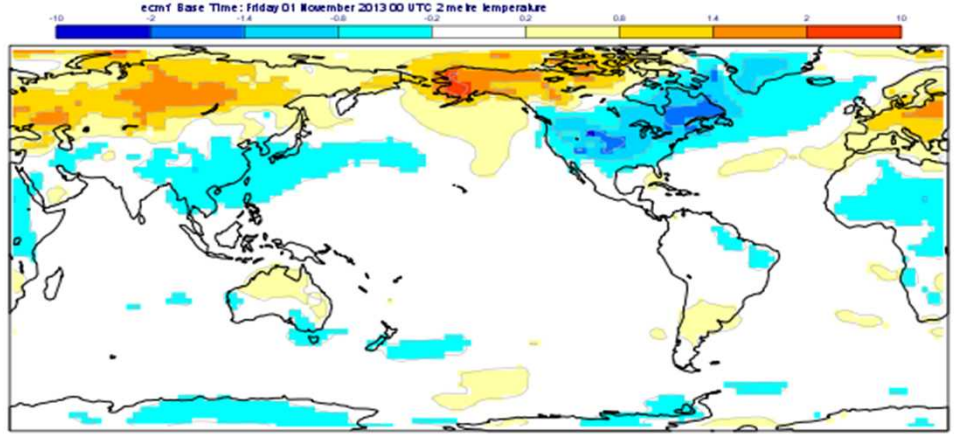


DJF 2014 anomalies:

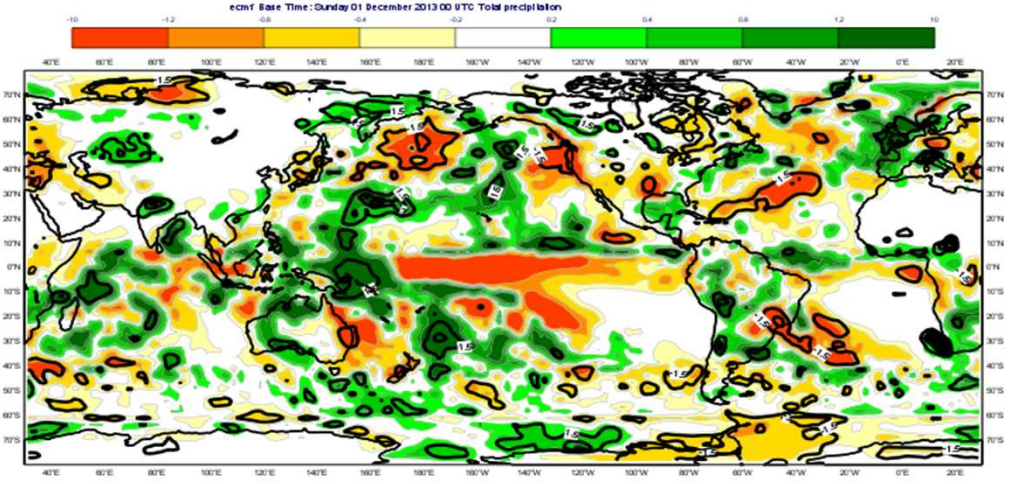
verifying analysis: Composites for proj. onto NH Cl2

2m temp.

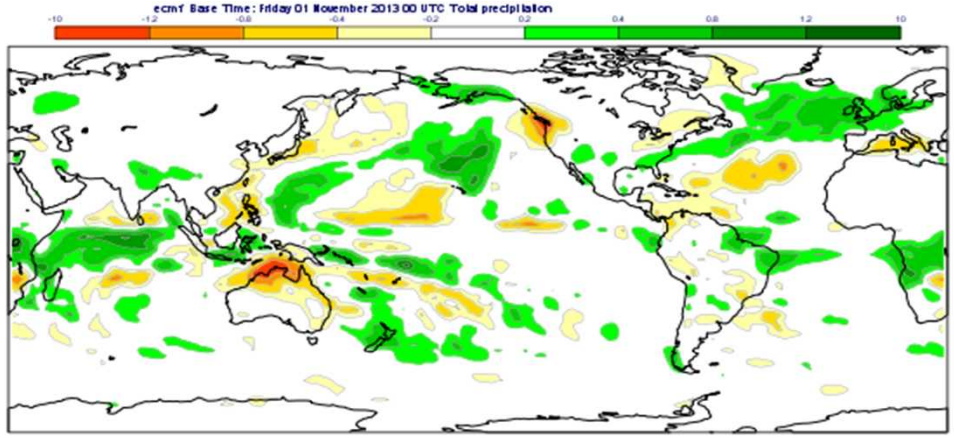




GPCP



precip



Flow dependent verification over the Atlantic sector

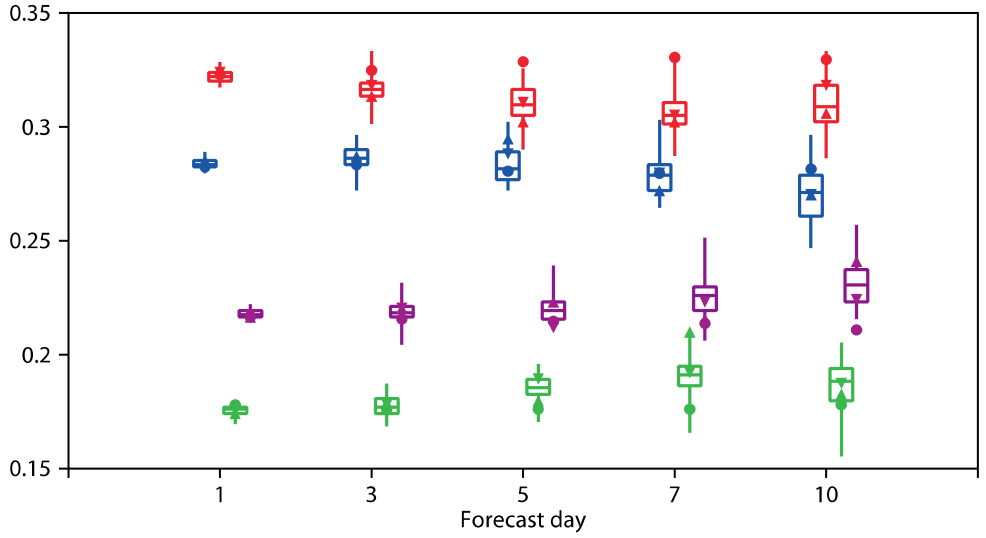
- Identifying the flow configurations that lead to a more/less accurate forecast and quantifying the skill changes.

- The concept of weather regimes is used to classify different flow configurations.

- Oper. Forecast data: ENS cold season (Oct to April) 2007-2012 operational analysis

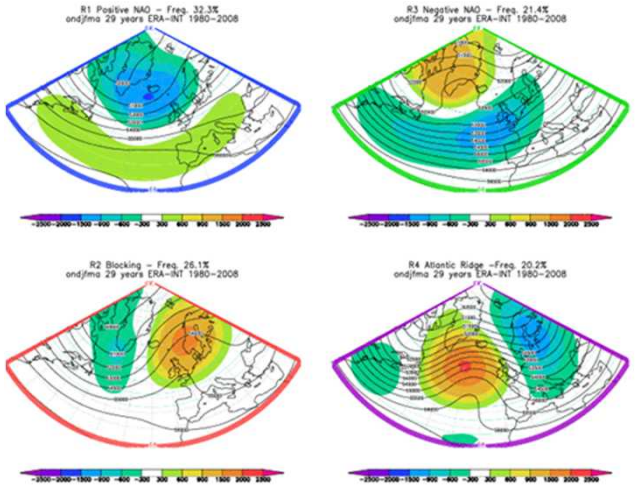
(Ferranti et al. 2014 QJRMS)

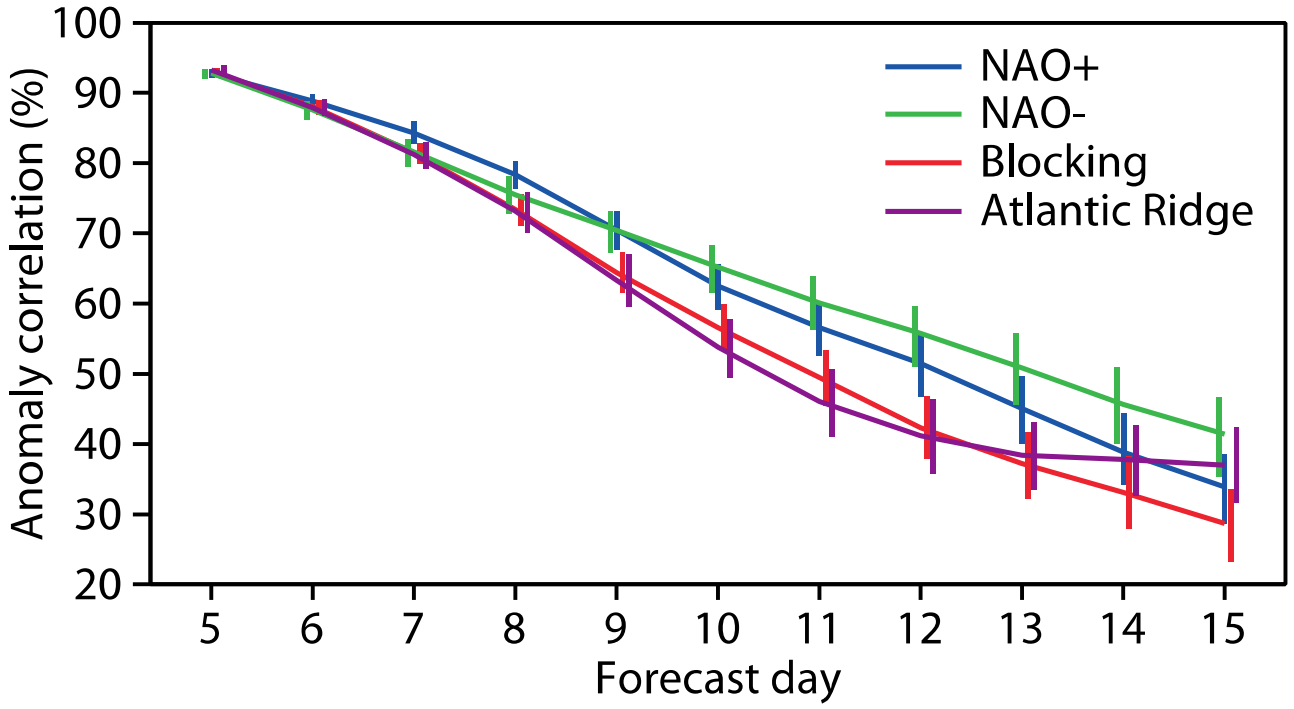
Climatological frequency distribution for the 4 Euro-Atlantic regimes as simulated by the ECMWF ensemble at different forecast ranges



Red indicate the frequency of the BL regime, blue (green) the frequency of the NAO+ (NAO-) and violet the frequency of the AR regime. The observed frequencies are indicated by a circle while the frequencies from the ECMWF operational high resolution and the unperturbed forecasts are indicated by a pointing down and a pointing up triangle respectively.

Which flow regime leads to more skilful predictions?





Anomaly correlation of the ensemble means for the four forecast categories as a function of forecast range. The bars, based on 1000 subsamples generated with the bootstrap method, indicate the 95% confidence intervals.

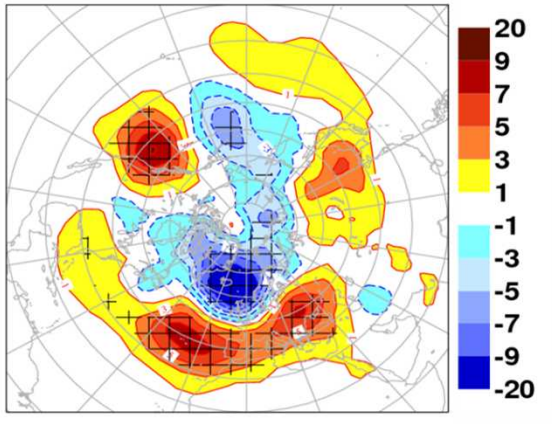
Poor forecasts at day 10

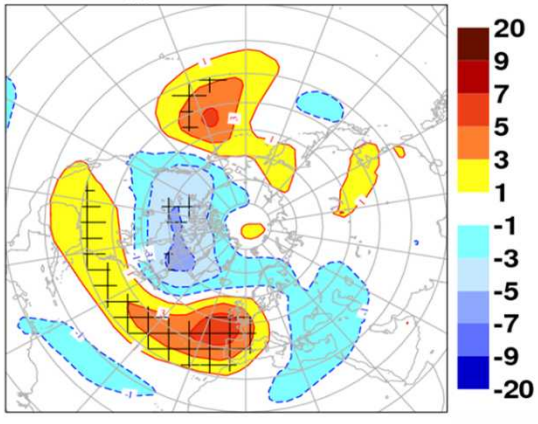
The performance of the Ensemble is assessed by stratifying the cases according to their initial conditions as well as their accuracy at forecast day 10.

Poor (good) forecasts => RMSE of the ensemble mean larger (smaller) than the upper (lower) fifth of the whole RMSE distribution.

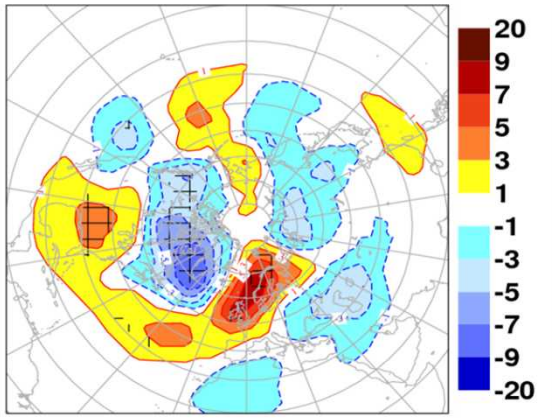
The RMSE is computed over the European domain at day 10. For each group and each category we compute composites maps of z500 anomalies at several time steps.

Forecasting regimes transitions:





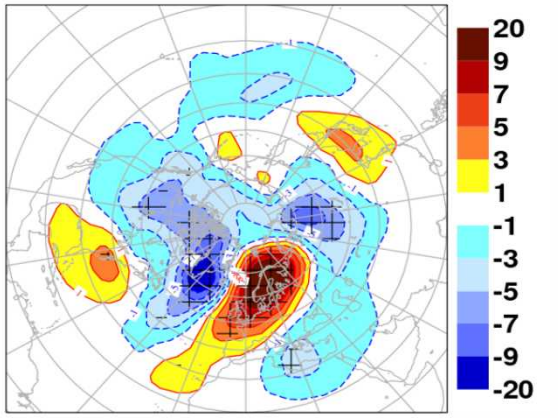
Composites of z 500 anomalies for all the forecasts initiated with flow configuration close to the NAO+ and with a RMSE at day 10 exceeding the upper quintile of the RMSE distribution.

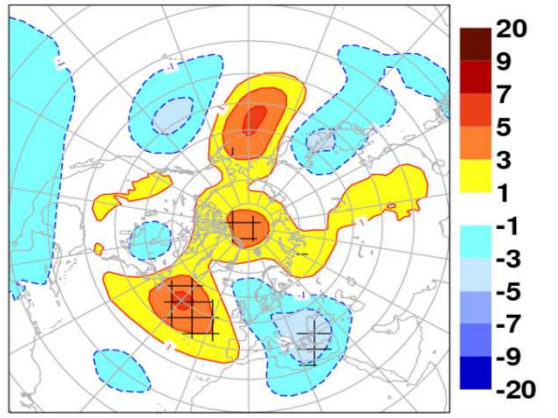


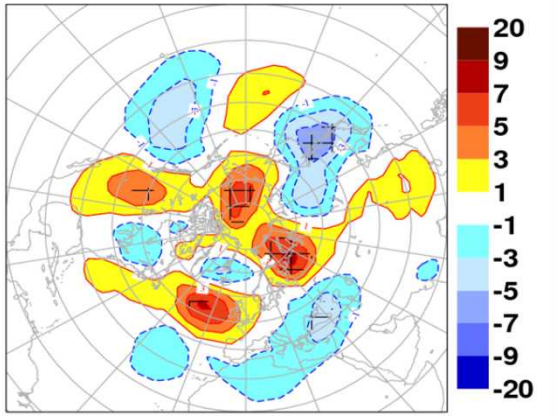
	Day 0	Day 1	Day 5	Day 7	Day 10
Forecasts with large RMSE at day 10					
NAO+	100	81	56, 44	54, 40	37, 21
BL	0	8	28, 40	35, 53	42, 51
NAO-	0	2	0	2	2, 5
AR	0	9	16	9, 5	19, 23

NAO+ (Zonal flow) → BL is underestimated
NAO+ persistence is overestimated

Forecasting regimes transitions:





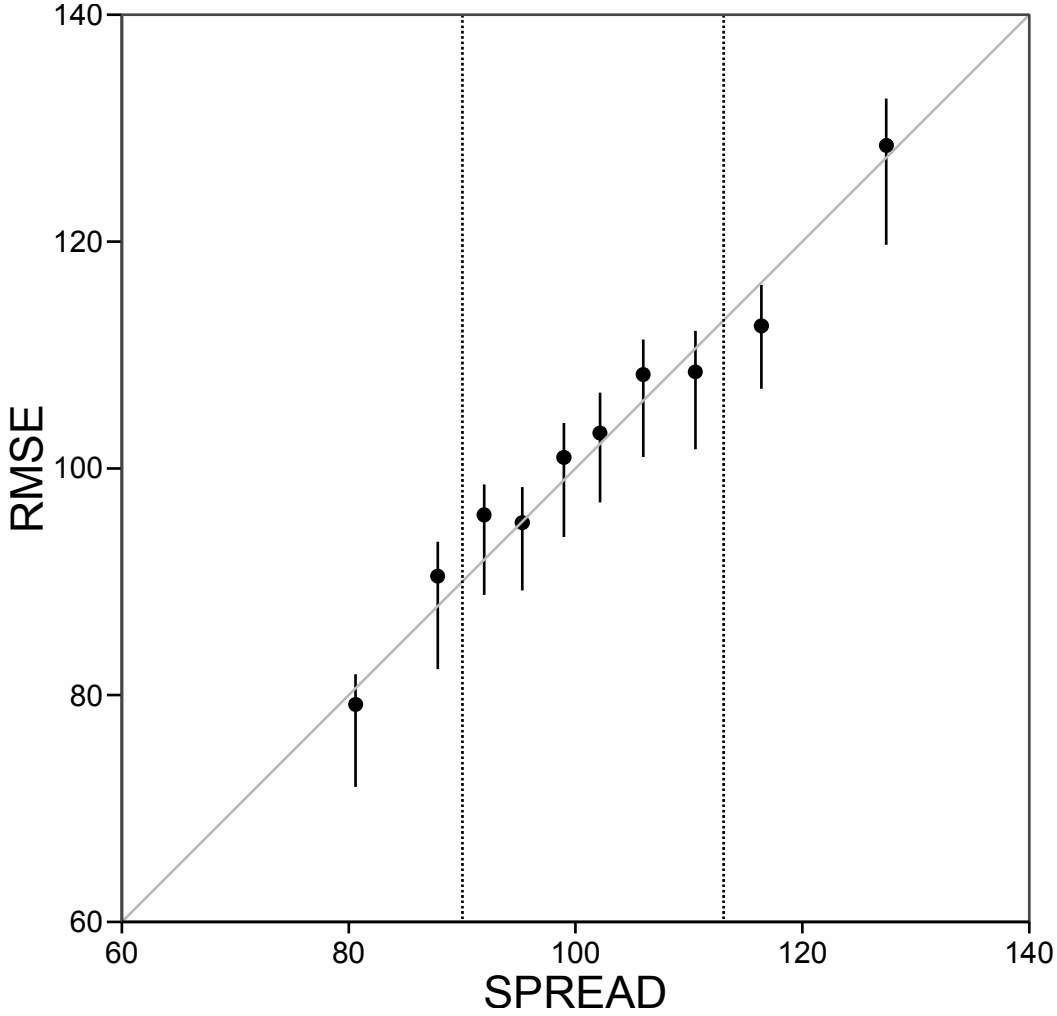


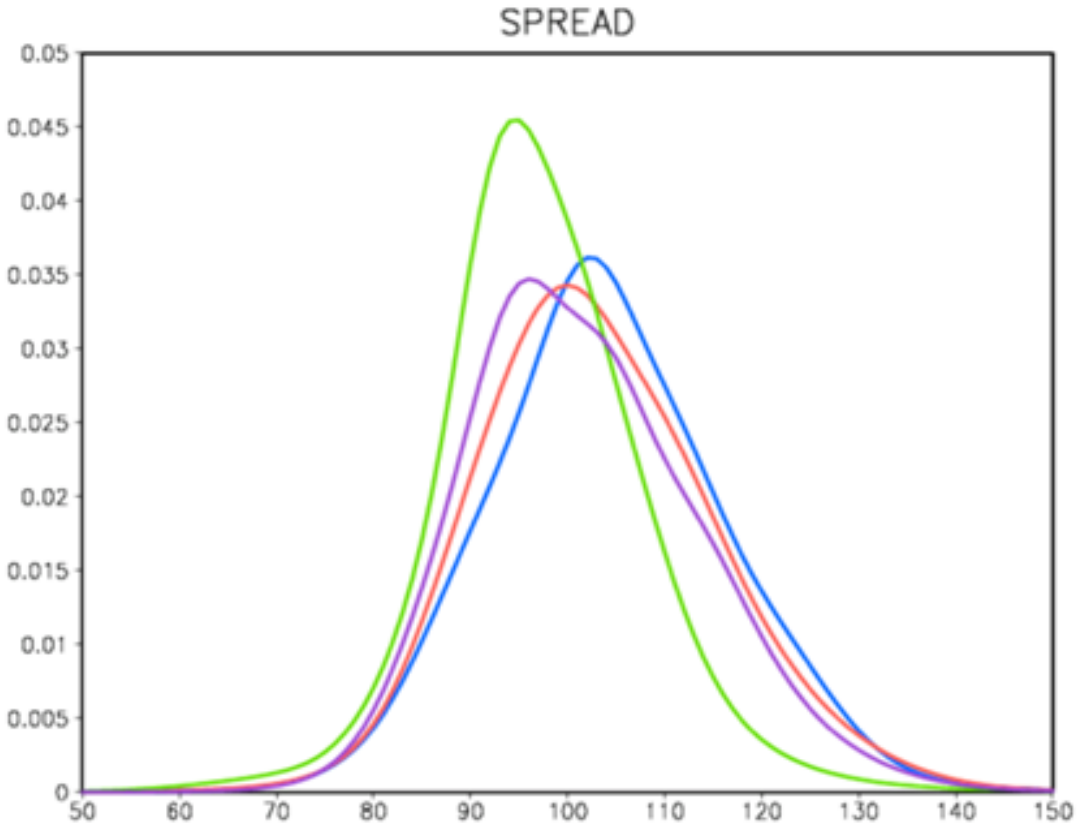
Composites of z 500 anomalies for all the forecasts initiated with flow configuration close to BL and with a RMSE at day 10 exceeding the upper quintile of the RMSE distribution.

	Day 0	Day 1	Day 5	Day 7	Day 10
Forecast with large RMSE at day 10					
NAO+	0	20	25 , 18	28 , 25	28 , 18
BL	100	70	44 , 52	36 , 47	29 , 41
NAO-	0	2	2 , 7	3 , 8	5 , 21
AR	0	8	29 , 23	33 , 20	38 , 20

Blocking persistence is underestimated
BL → NAO- is underestimated

Scatterplot of RMSE versus the spread for day 10 forecasts. The vertical lines in the scatterplot represent the upper and lower fifth values of the ensemble spread distribution





Ensemble spread distribution at day 10 for forecasts initiated in: NAO+ (blue) blocking (red), NAO- (green) and AR (violet) regime

Summary

- The revised cluster product provides the users with a set of weather scenarios that appropriately represent the ensemble distribution
- 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
- The accuracy of the product can be quantified and the use of climatological weather regimes allows flow dependent skill measures
- This clustering tool can be used to create EPS clusters tailored to the users' needs (e.g. different domain, different variables)

Regime predictability:

- There is skill in predicting the Euro-Atlantic weather regimes up to day 15-21 in the monthly forecast and up to month 1 in the seasonal forecast.
- Sys4 skill is improved with respect to Sys3 for all the regimes.

Flow dependent verification:

- Blocking is the regime associated with the least accurate forecasts.
- Poor forecasts underestimate the persistence of blocking while overestimate the maintenance/transitions of/to zonal flow (NAO+)
- The ensemble spread is a useful indicator of the forecast error.
- The spread of the forecasts initiated in NAO- is significantly smaller than for the forecasts initiated in the other regimes. This is consistent with their higher skill.