Operational regime verification used in the ECMWF forecasting

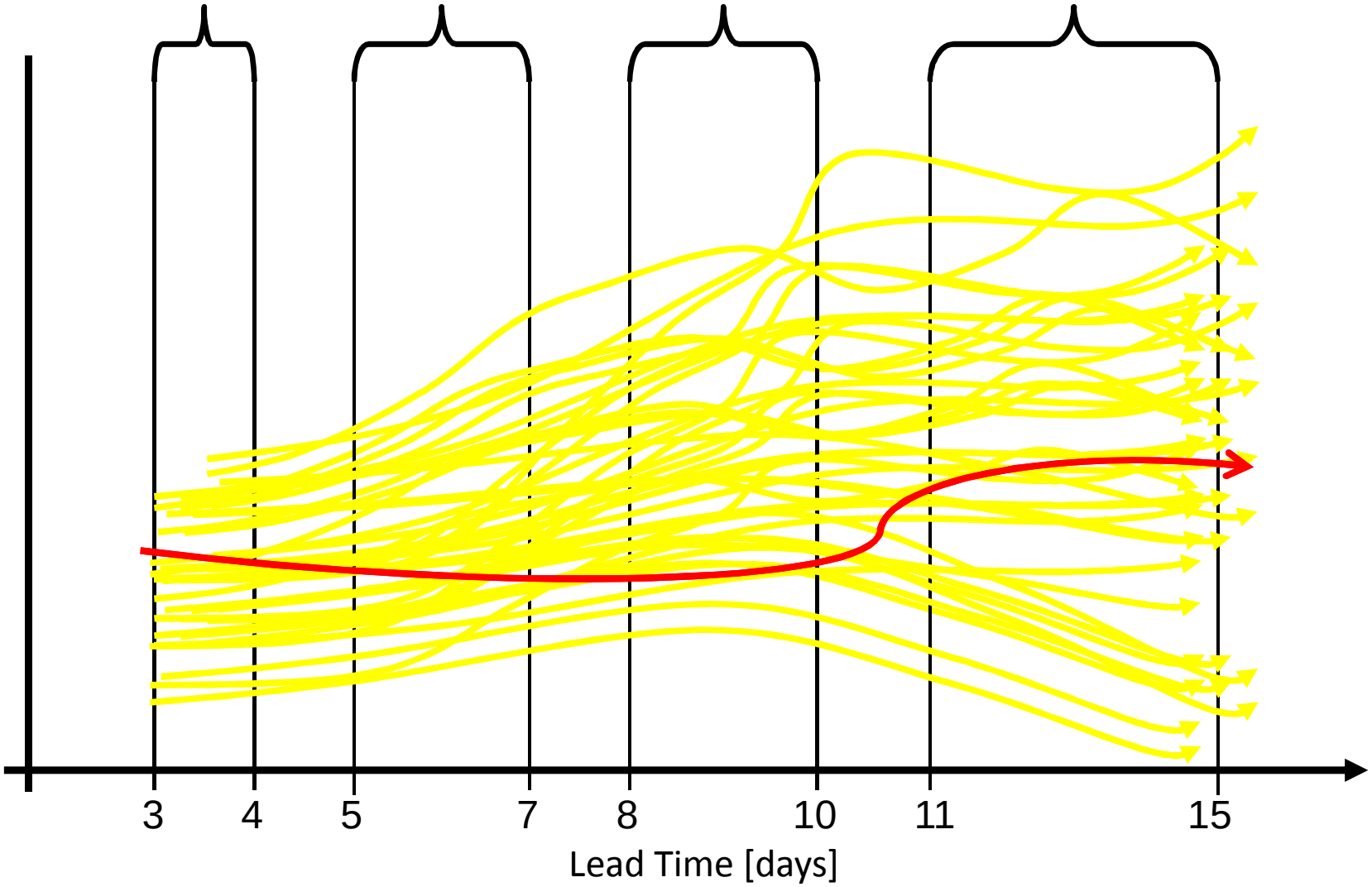
Susanna Corti
Institute of Atmospheric Sciences and Climate (ISAC)
National Research Council (CNR)

Outline

- Cluster product at ECMWF
  - Concept
  - Examples
  - Visualisation on ECMWF web-site
  - Flow dependent predictability of Euro-Atlantic regimes

- Cluster analysis – Generalities
  - Suitable (sub)spaces of states
  - K-means method
  - Significance
  - Number of clusters
  - Examples

ECMWF Ensemble Prediction System: 1 control (+ 50 ensemble members)



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.
- The new 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: 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.

Cluster analysis - Clustering techniques

With the terminology clustering techniques we refer to a set of different methodologies used to identify groups of elements gathered together in a (suitable) (sub)space of states.

Such methodologies can be summarised in essentially three different typologies.

- Hierarchical Clustering Hierarchy of sets of groups, each of which is formed by merging one pair from the collection of previously defined groups. We start with a number of groups equal to the number n of observations/states. The first step is to find the two groups that are closest and combine them into a new group. Once a data vector has been assigned to a group, it is not removed. This process continues until, at the final $(n-k)^{th}$ step all the n observations have been aggregated into k groups.

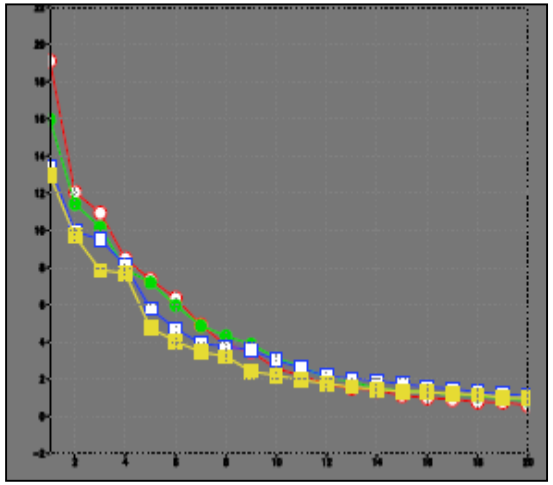
- Nonhierarchical Clustering Methods that allow reassignment of observations as the analysis proceeds.

- Bump-hunting methods Quest of local maxima in the Probability Density Function of states.

Clustering analysis – Suitable (Sub)spaces of states

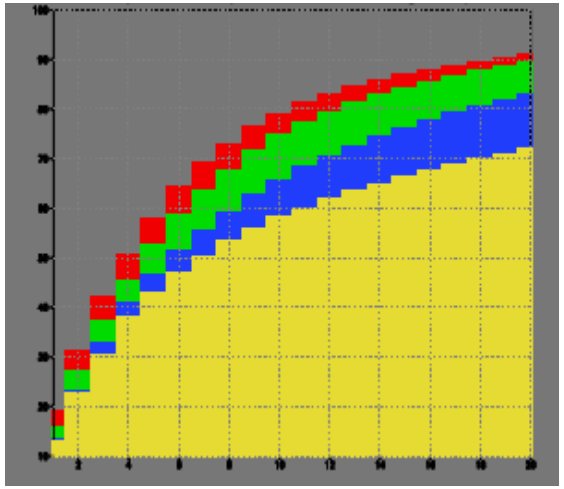
Clustering techniques are effective only if applied in a L-dimensional phase space with L << N (N=number of elements in the data set in question). If the actual space of states is too large (ex: 500 maps with 25x45 grid points) it is advisable to compute the clusters in a suitable sub-space.

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



Analogy between Empirical Orthogonal Functions decompositions and Fourier analysis

Suppose we have a variable expressed in terms of space and time y(x,t). In Fourier analysis we can express this data set in terms of a specified set of functions, normally sines and cosines. If y(x,t) is defined on the interval 0<x<L then we can write:

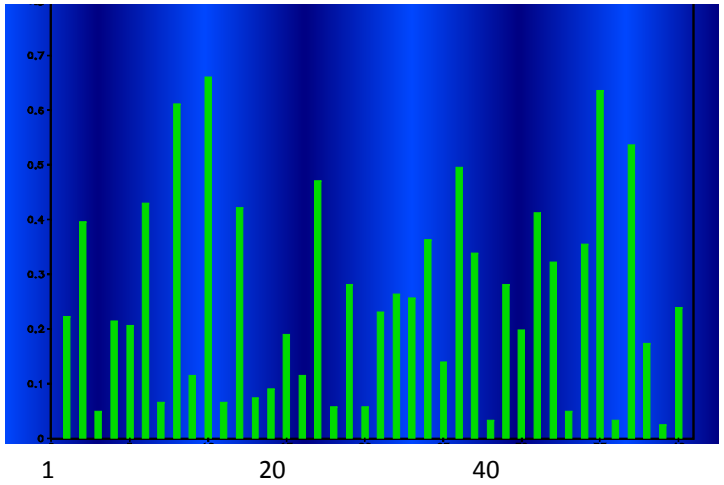
$$y(x,t) = \sum_{n=0}^{N} A_n(t)\cos(2\pi n / L) + B_n(t)\sin(2\pi n / L)$$

Now instead of being expressed as a function of space and time, the spatial dependence is expressed in terms of the functional forms of sines and cosines with different frequencies (basis functions). The time dependence is expressed in the coefficients An(t) end Bn(t) that express the amplitudes of the set of basis functions *cos(2πn/L)* and *sin(2 πn/L)* as functions of time. So now the same information that was contained in y(x,t) is contained in An(t) e Bn(t), but expressed differently.

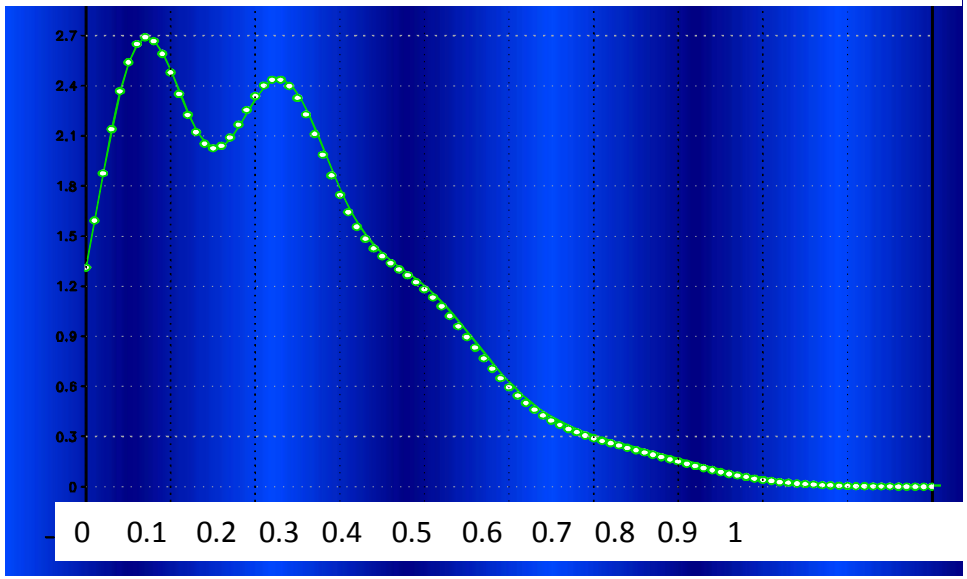
In the Empirical Orthogonal Function (EOF) analysis, we do a very similar thing, except that the spatial structures are not specified beforehand, but rather are determined by the structure of the data itself (e.g. empirical functions)

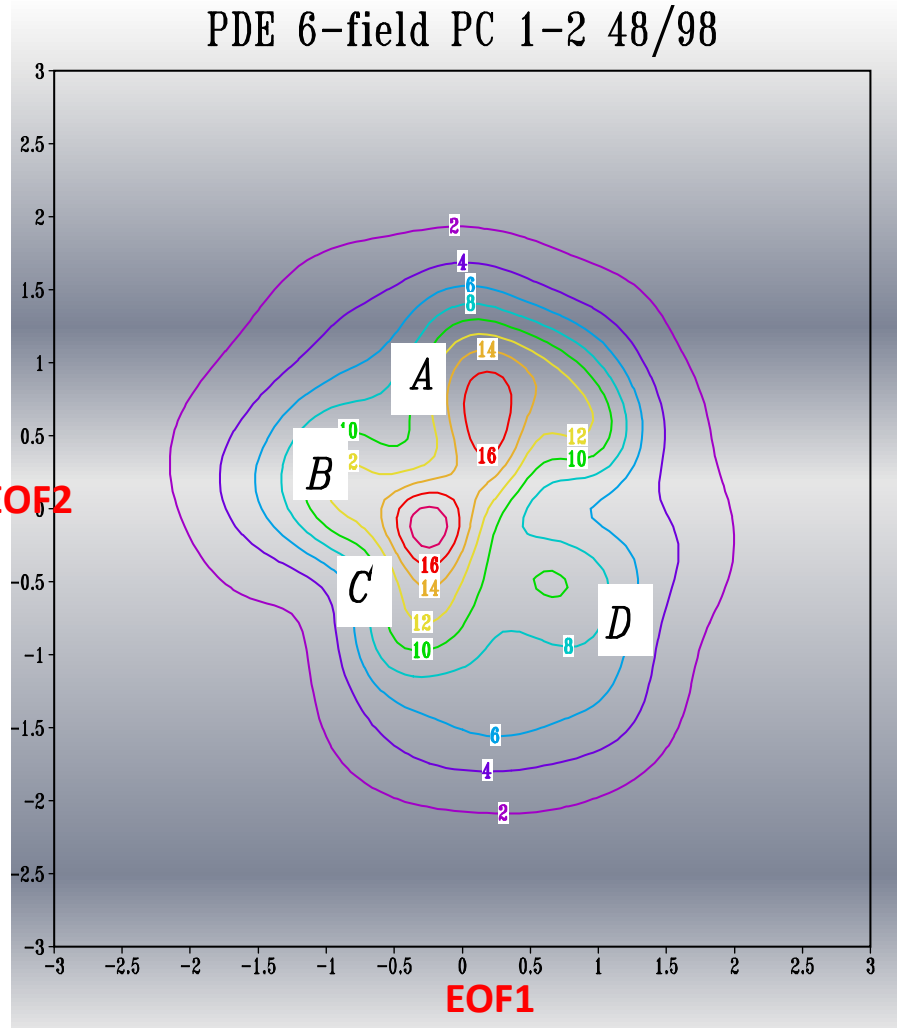
Examples of bump-hunting

S3- DJFM 2009-2010 NAO- frequency



NAO- frequency PDF





After Corti Molteni and Palmer 1999 Nature

Cluster analysis - K-means method

The most widely used non-hierarchical 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).

- 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 pseudorandom 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).

Use of K-means to identify regime structures in observed and simulated datasets

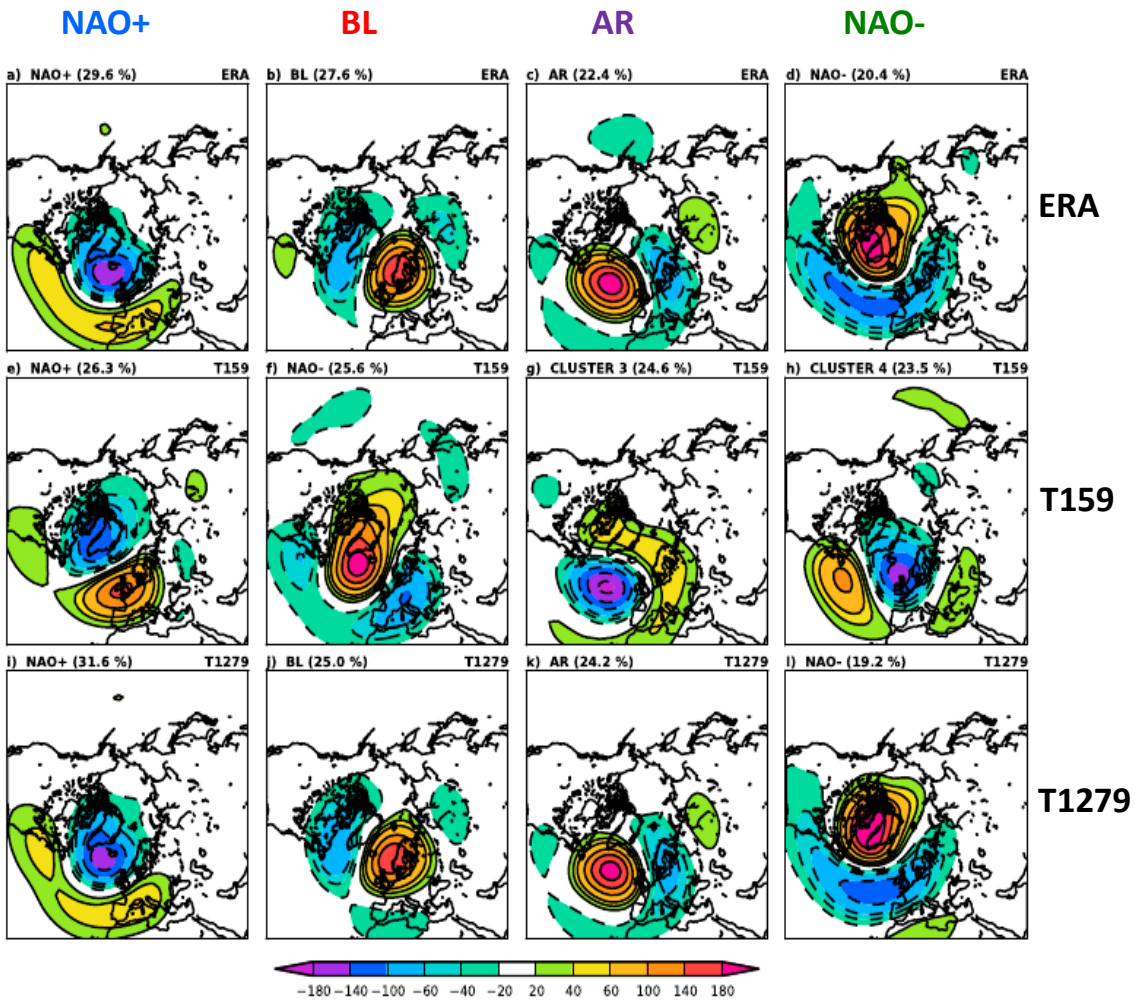
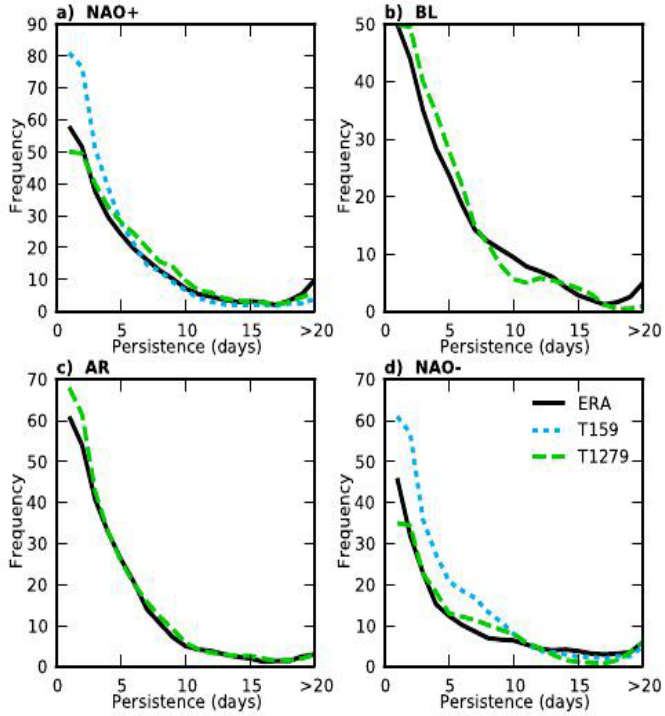


Table 1. Error ϵ Relative to ERA Clusters[a]

	ϵ_{NAO+}	ϵ_{BL}	ϵ_{AR}	ϵ_{NAO-}	$\bar{\epsilon}$
T159	4763.0	–	–	3813.5	3478.4
T1279	125.3	85.7	19.9	134.3	91.3
NCEP	21.8	14.3	8.0	52.0	24.0

[a]The mean error $\bar{\epsilon}$ for the T159 model configuration is the minimum possible error found by matching all permutations of ERA clusters with the T159 clusters.



Simulating regime structures in weather and climate prediction models
A. Dawson, T. N. Palmer, and S. Corti – GRL 2012

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.

Cluster product at ECMWF: large scale climatological regimes

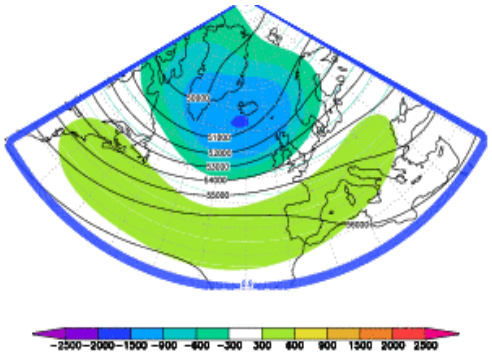
To put the daily clustering in the context of the large-scale flow and to allow the investigation of regime changes, the new ECMWF clustering contains **a second component**. Each cluster is attributed to one of a set of four pre-defined climatological regimes

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

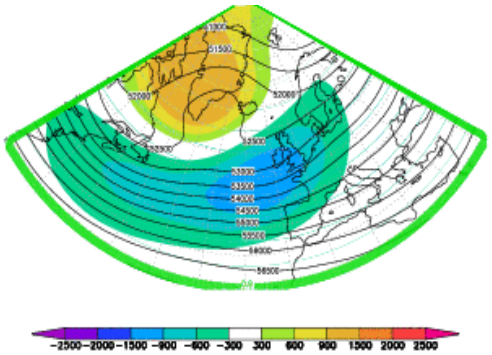
Climatological Regimes in the cold season
Euro-Atlantic Region

500 hPa Geopotential height – 29 years of ERA INTERIM ONDJFM 1980-2008

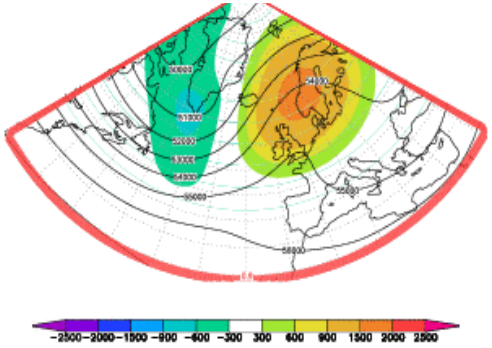
Positive NAO 32.3%



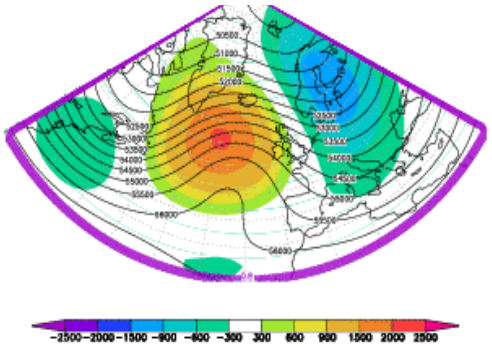
Negative NAO 21.4%



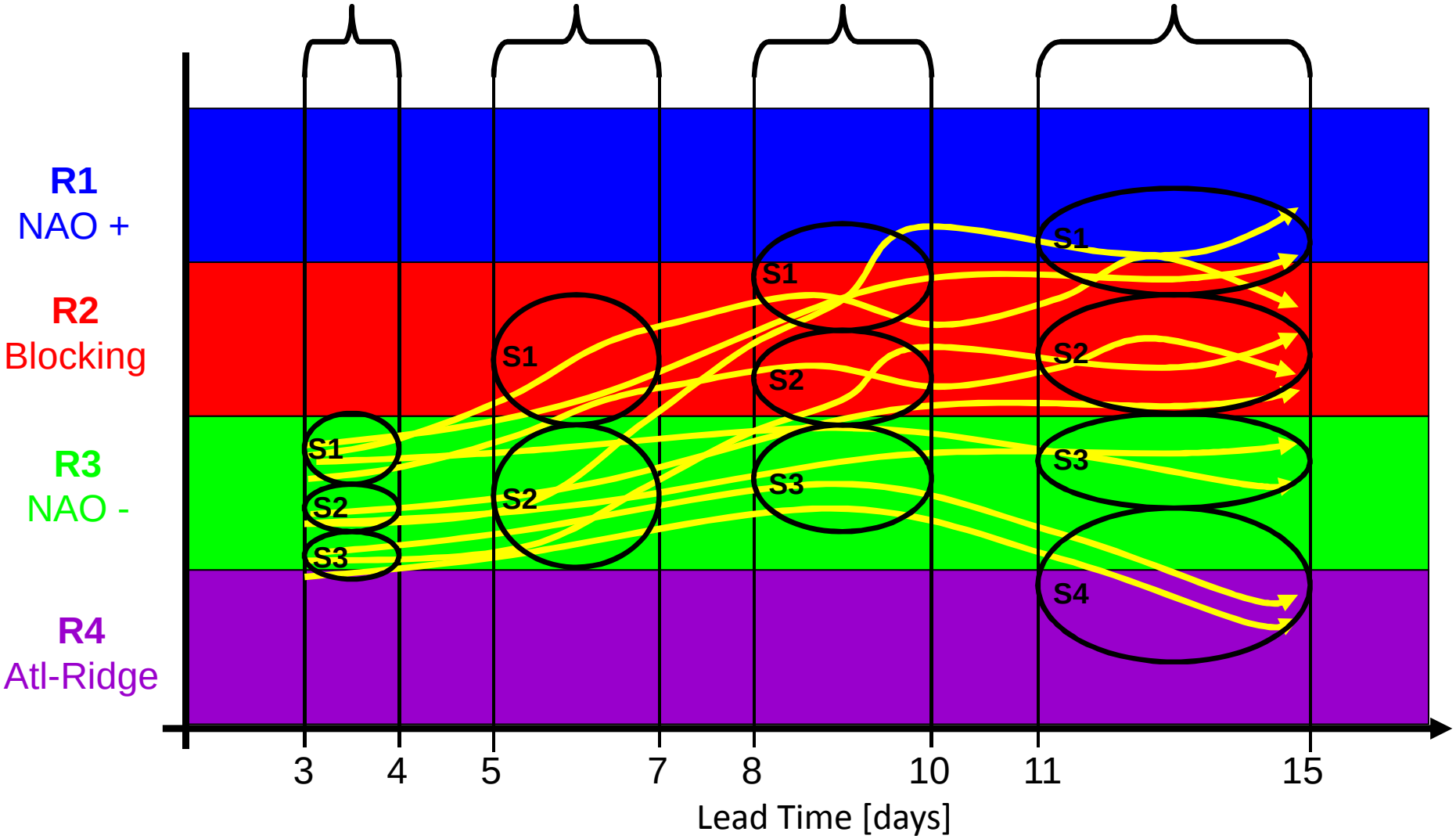
Euro-Atlantic Blocking 26.1%



Atlantic Ridge 20.2%

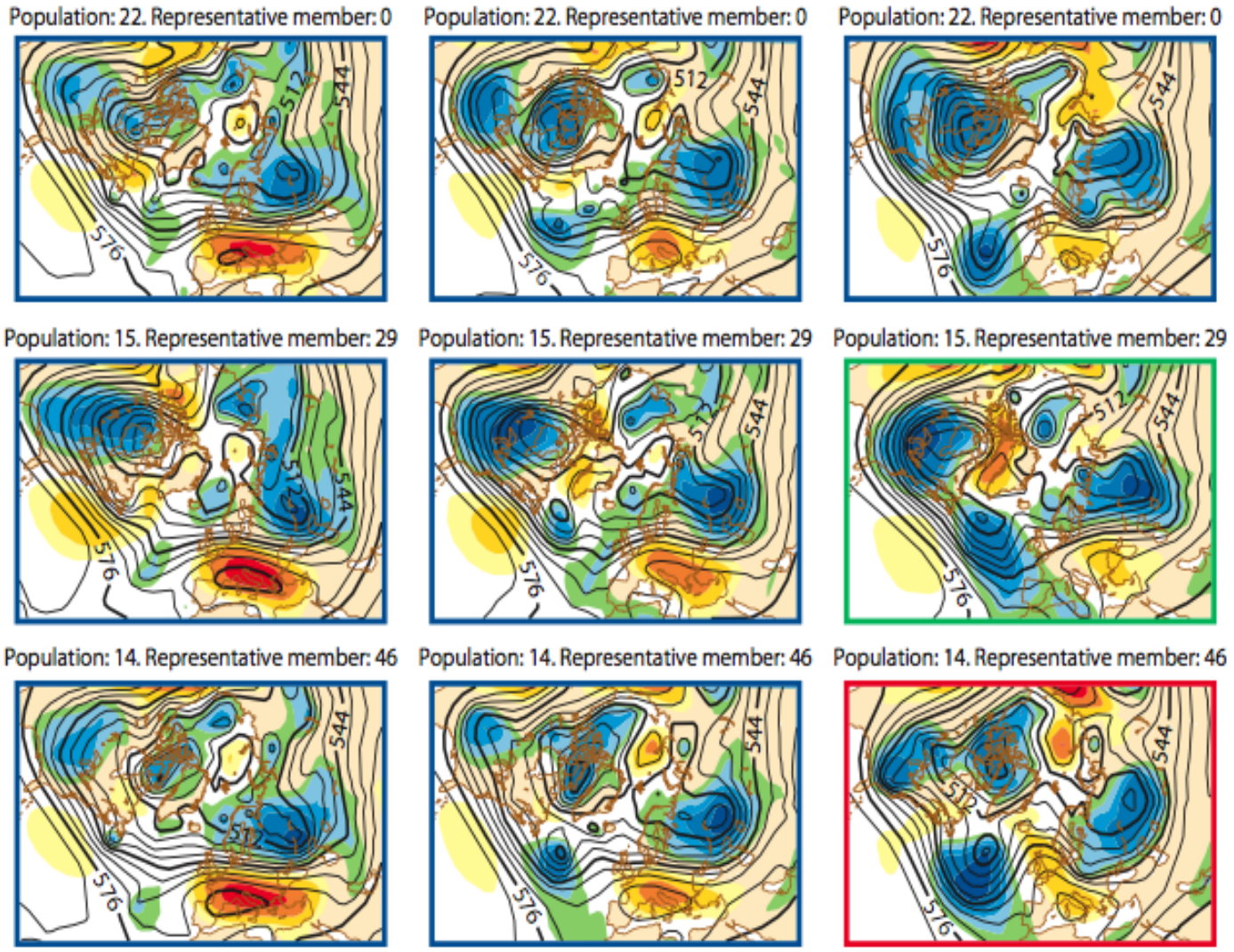


Regimes & Scenarios



Regime transitions within a time window

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



Cluster scenario - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Back Forward Reload Stop Home http://nwmstest.ecmwf.int/products/forecasts/d/charts/med Go Google Search

WebMail Calendar Radio People Yellow Pages Download Customize... Daily Report (Daily Me... Anomaly

Spatial maps | 500 hPa geopotential | 2009 Pacific typhoon ... | Hurricanes Typhoon f... | Cluster scenario

Show guide

Method

Cluster scenario

Parameter

500

1000

Your Room

Add this product

Show overview

Parameter

Cluster

Forecast base time

Download...

PDF (2.7 Mbytes)

Postscript (3.7 Mbytes)

Cluster scenario

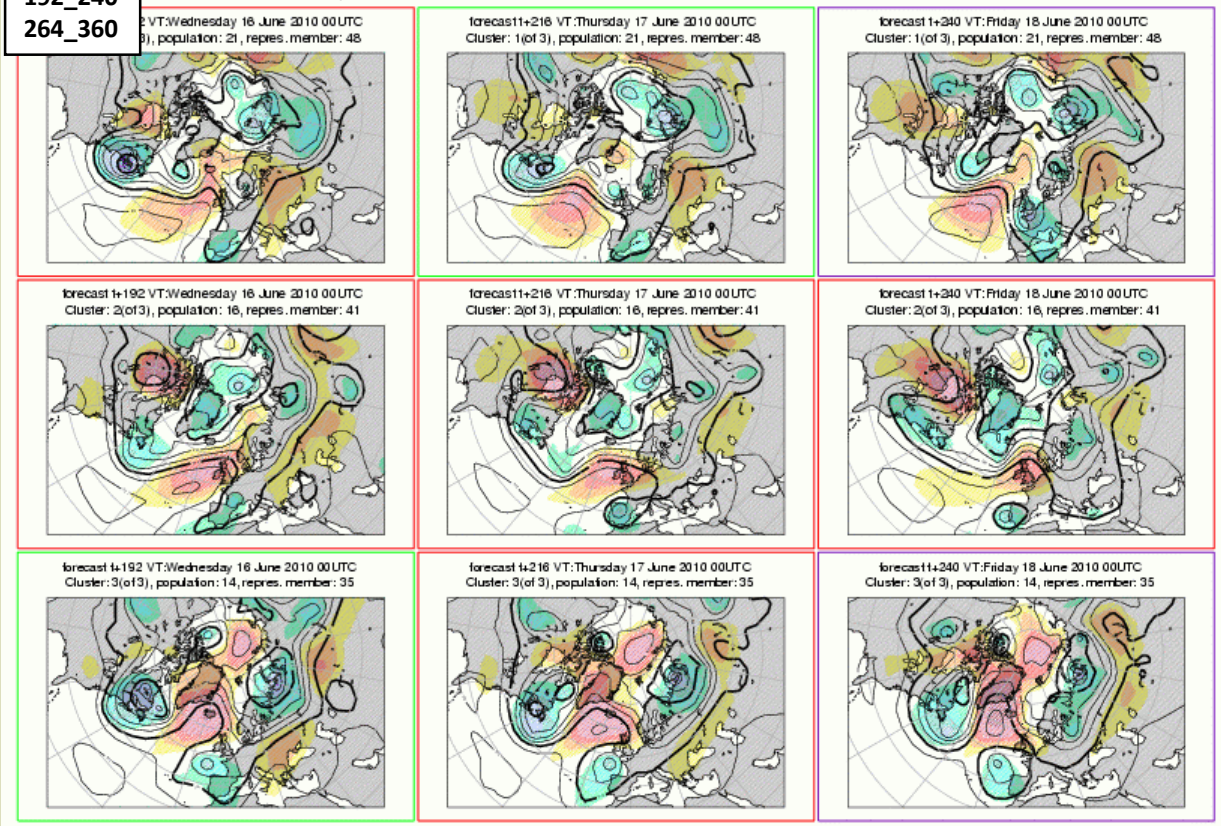
Maps of geopotential height: at 500hPa full field (black contours) and anomalies (colour shading), at 1000hPa full field only. Click on show guide for the full description.

Cluster | Forecast base time

192_240 | Tue 8 Jun 2010 00UTC

72_96
120_168
192_240
264_360

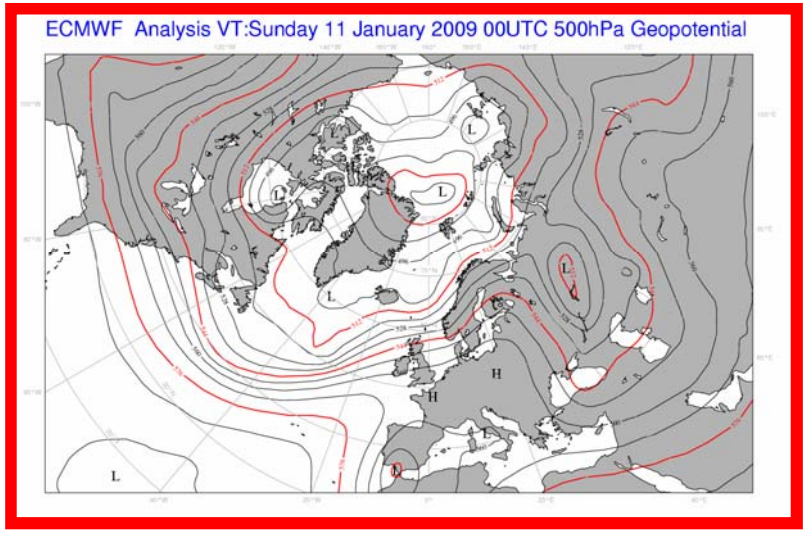
Tuesday 8 June 2010 00UTC ECMWF EPS Cluster scenario - 500 hPa Geopotential
Reference step t+192-240 Domain 75/340/30/40 Cont. in cluster=3 Det. in cluster=3

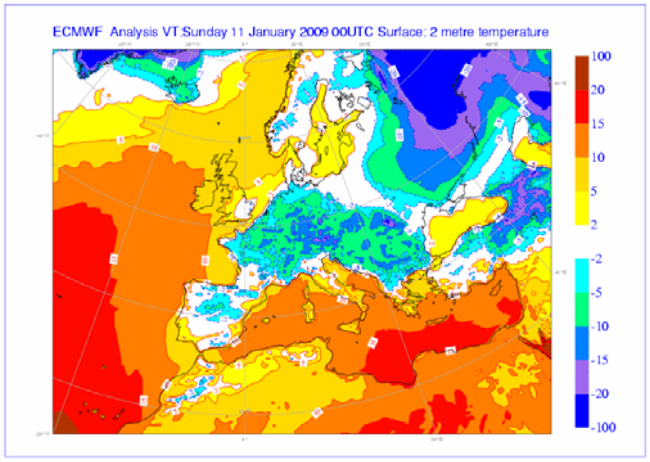


Done

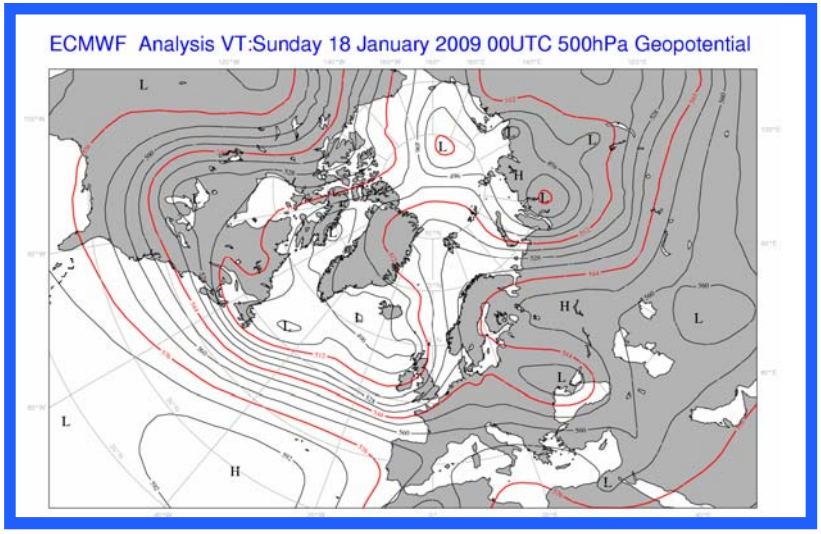
Transition from blocking to westerly flow:

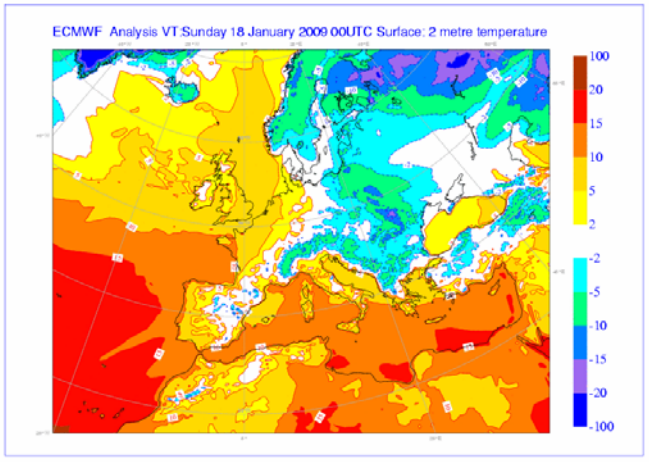
11 January 2009



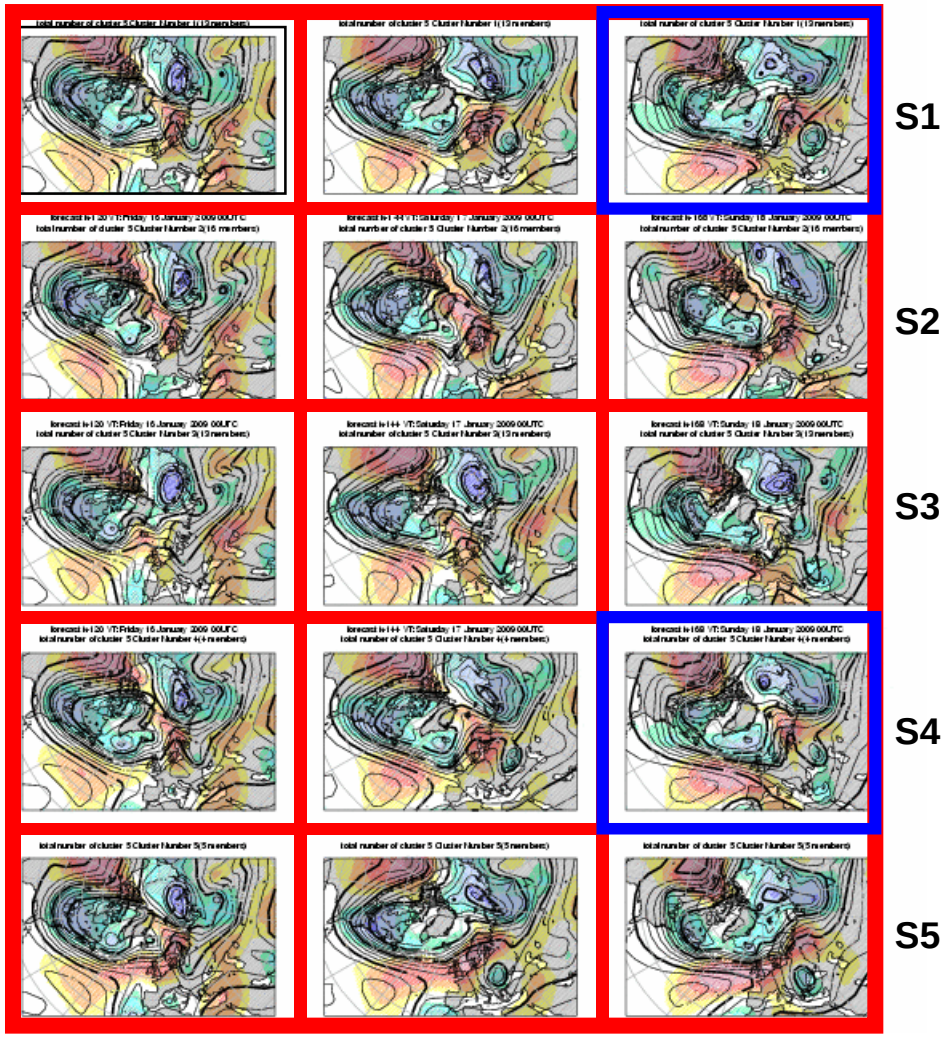


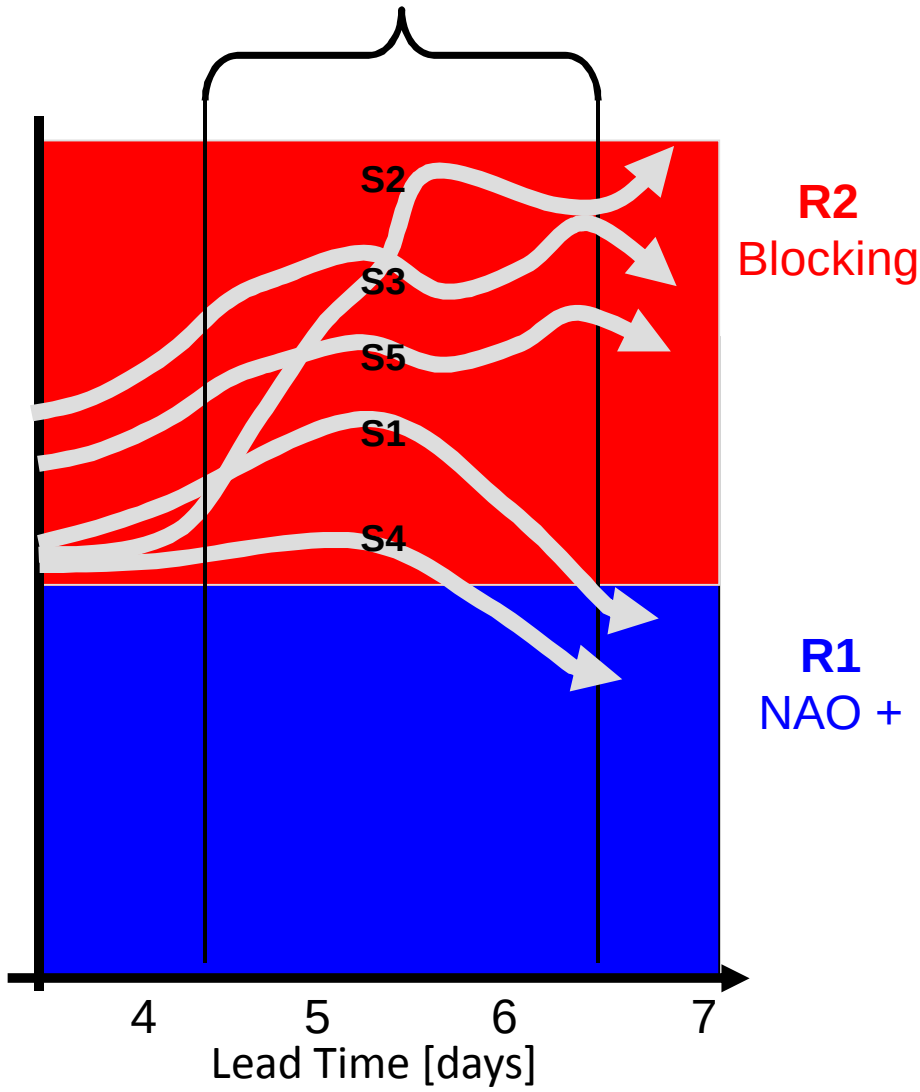
18 January 2009



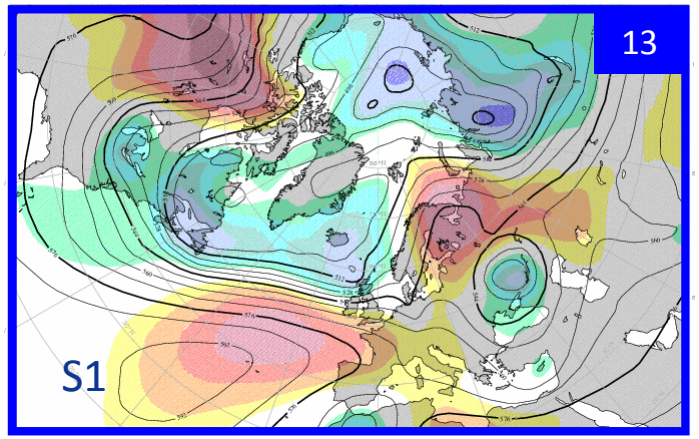


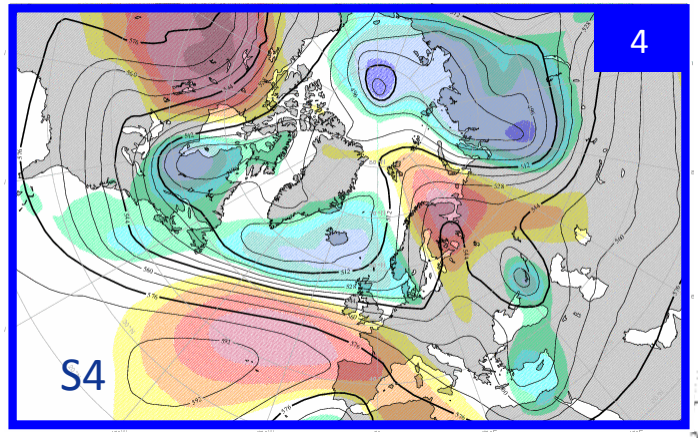
Regime transition within a time window



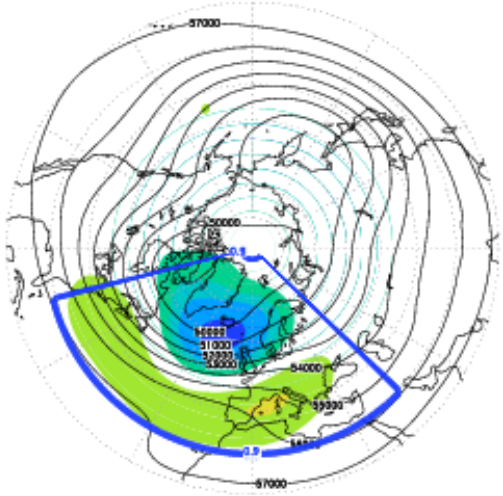


Scenarios fall into two different Regimes at day 7

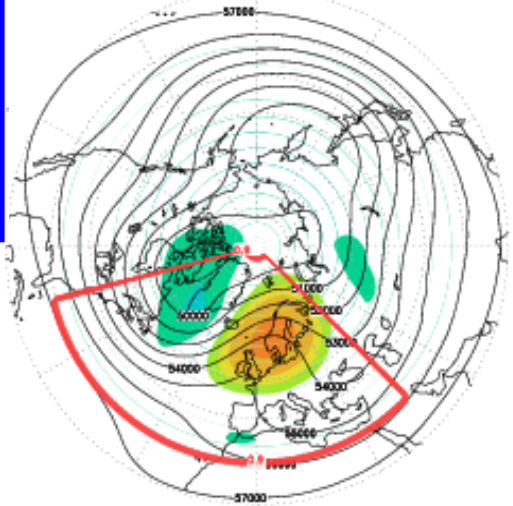


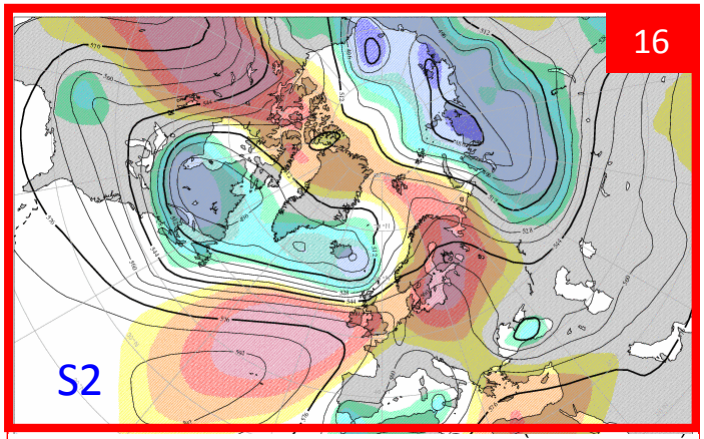


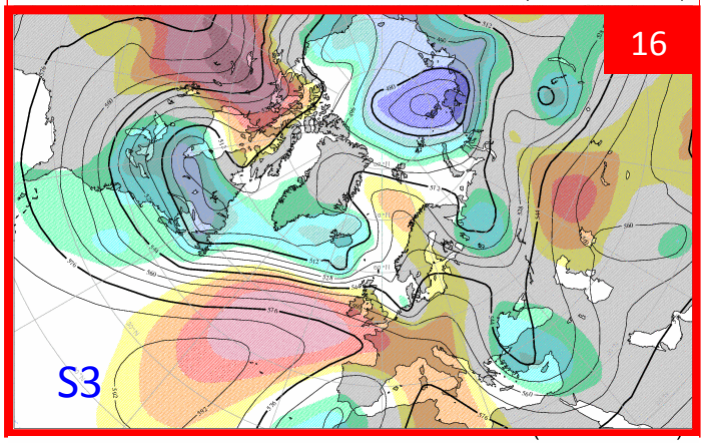
Positive NAO

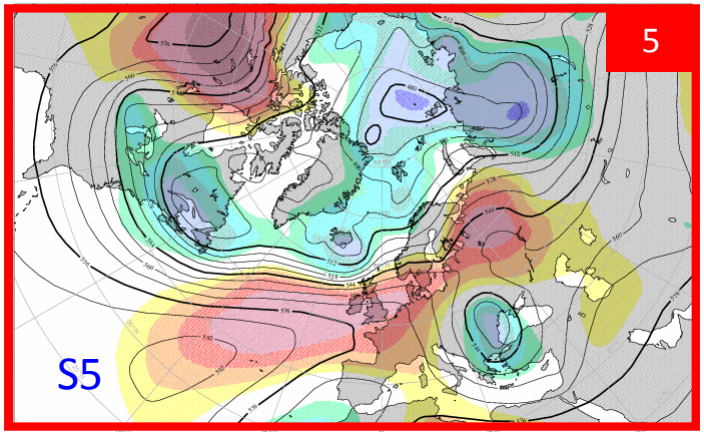


Blocking



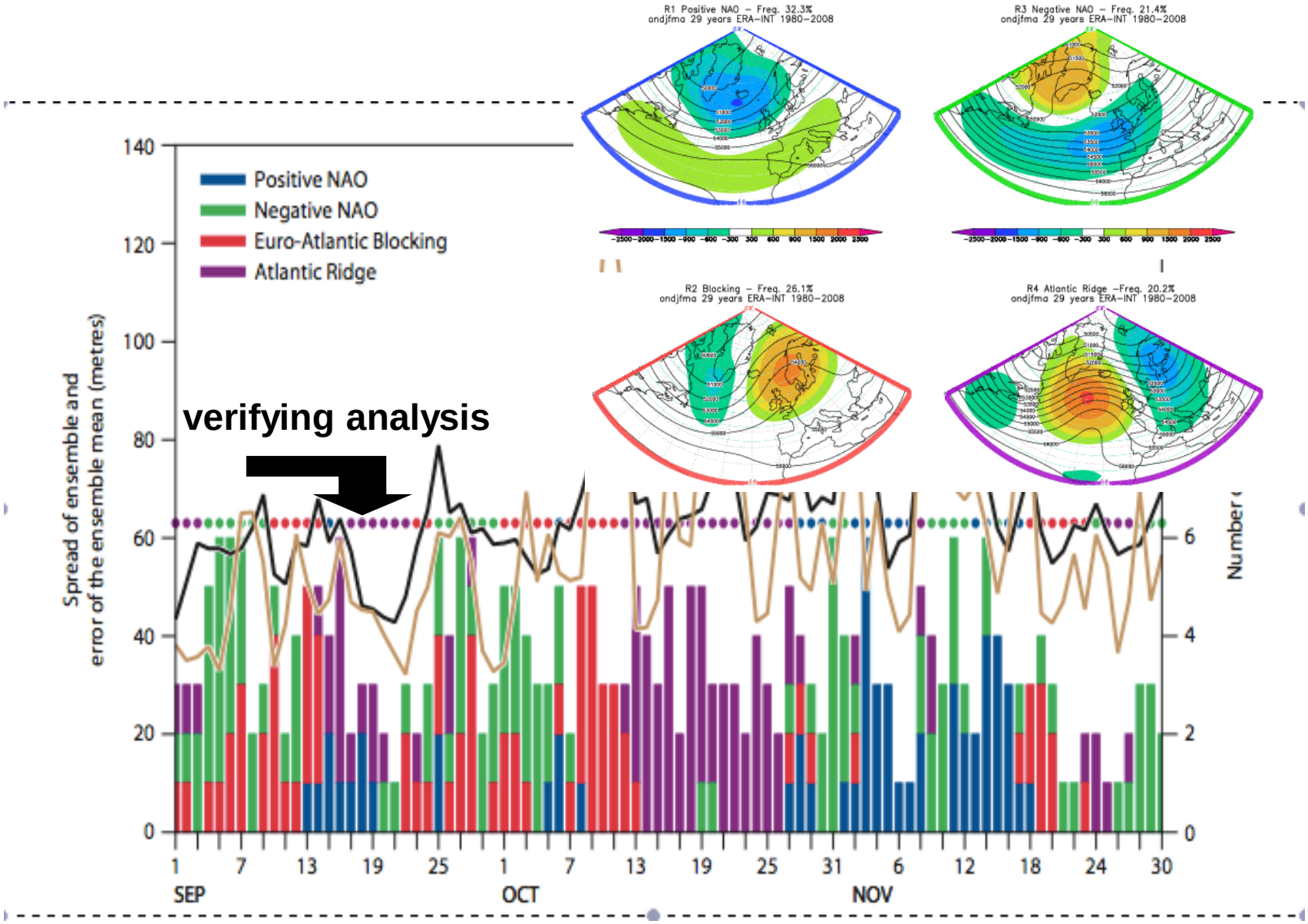






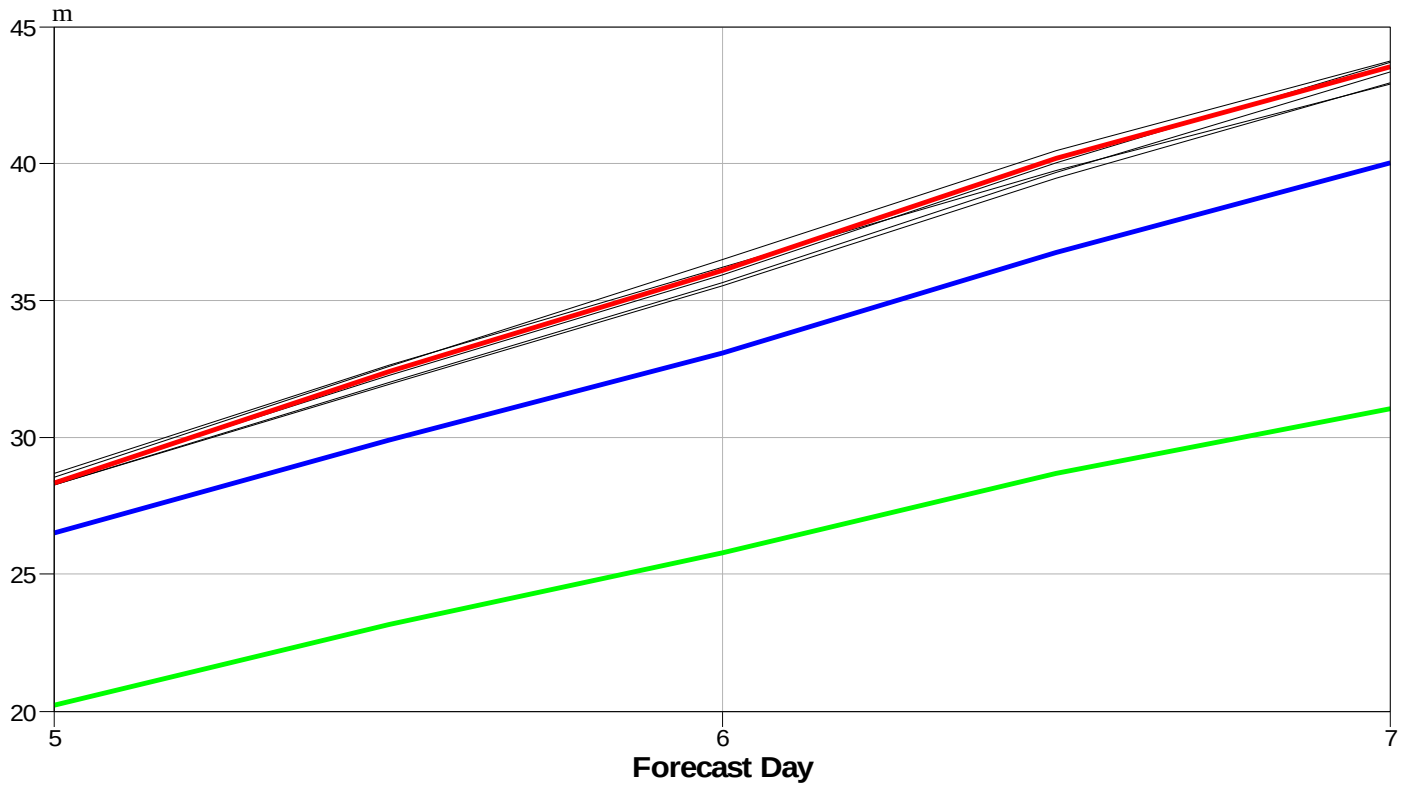
Verification & spread

Large-scale fixed climatological regimes



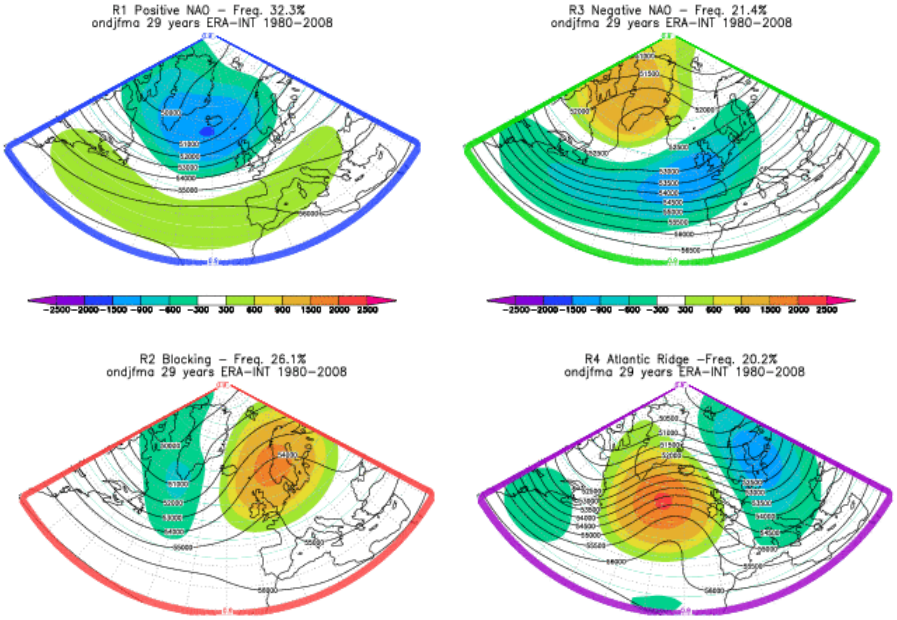
Verification: Continuous Ranked Probability Score (CRPS)

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



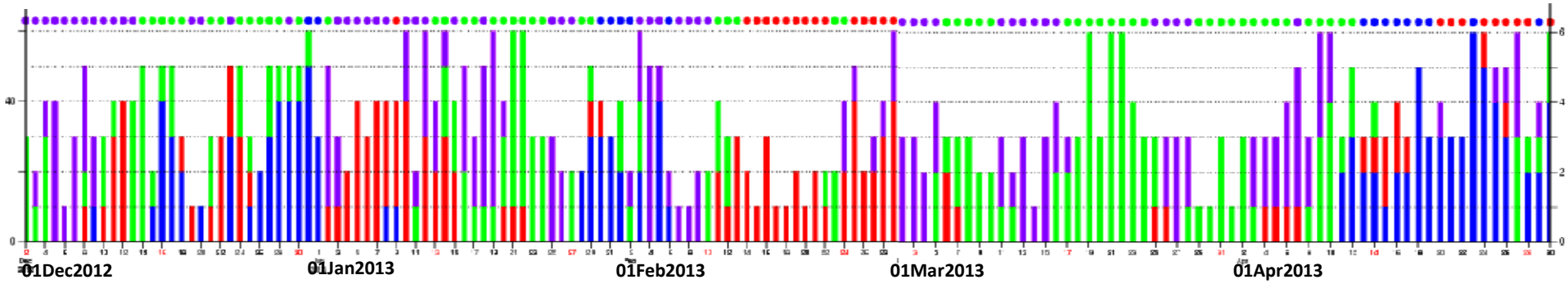
Verification: last season

Large-scale fixed climatological regimes



EPS scenarios z500 t+168 (7 days)

1Dec2012-30Apr2013



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)