Clustering Techniques and their applications at ECMWF

Laura Ferranti

European Centre for Medium-Range Weather Forecasts

Outline

- Cluster analysis - Generalities

- Cluster product at ECMWF

- Flow dependent verification

- Predictability of Euro-Atlantic regimes at different forecast ranges

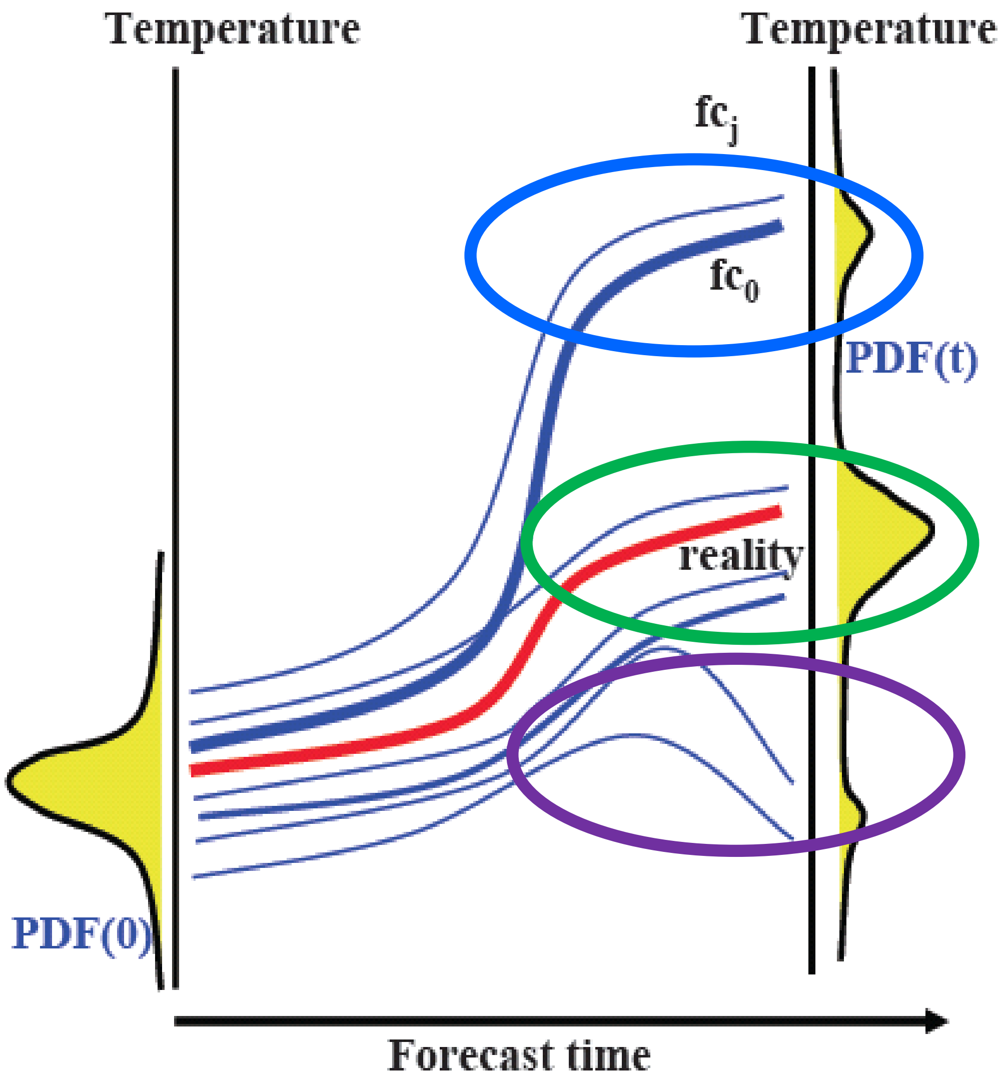
Cluster analysis - Generalities

“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*

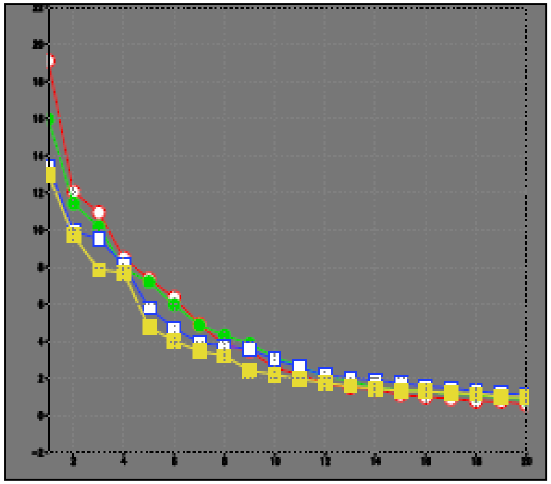
$$d_{i,j} = \left[\sum_{k=1}^{K} w_k (x_{i,k} - x_{j,k})^2 \right]^{1/2}$$

Weighted Euclideian distance between two vectors x_i and x_j

Clustering analysis – Suitable (Sub)spaces of states

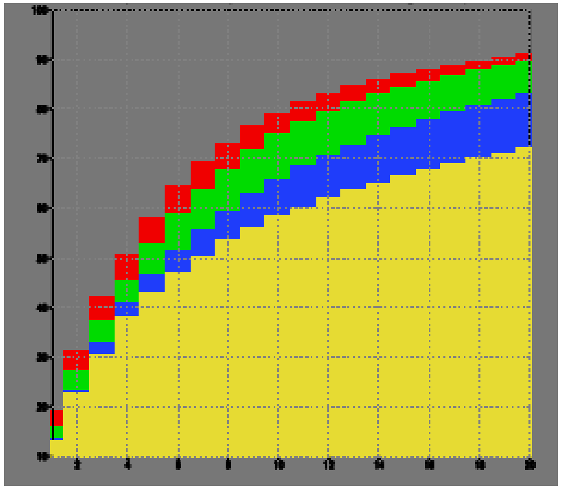
Clustering techniques are effective only if applied in a L-dimensional phase space with L << N (N=number of elements in the data set in question). If the actual space of states is too large (ex: 500 maps with 25x45 grid points) it is advisable to compute the clusters in a suitable sub-space.

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

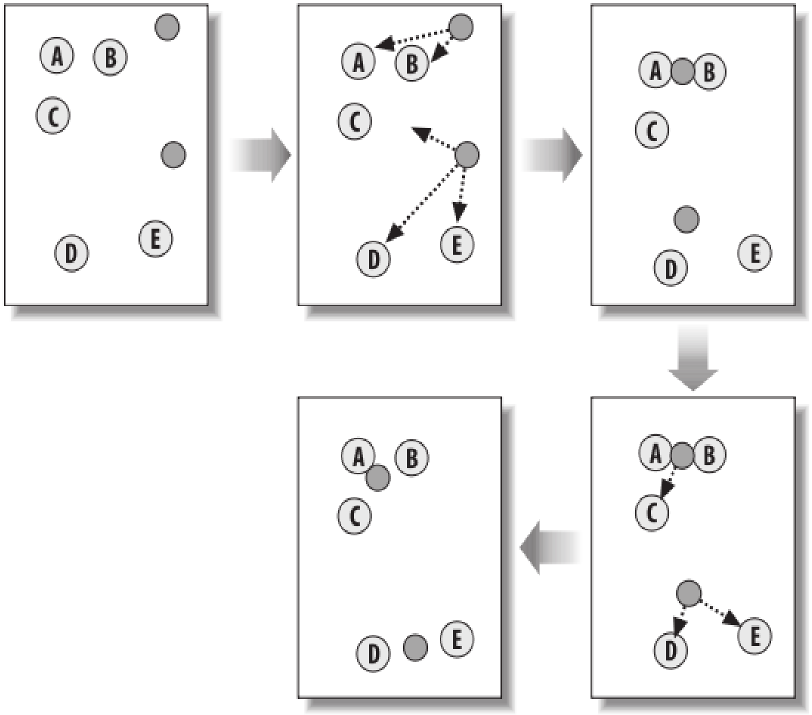
The most widely used exclusive clustering approach is called K-means method. K is the number of clusters into which the data will be grouped (this number must be specified in advance).

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 - 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.

They are archived in MARS and available to forecast users through the operational dissemination of products.

A graphical clustering product is available for registered users on the ECMWF web site: http://www.ecmwf.int/en/forecasts/charts/

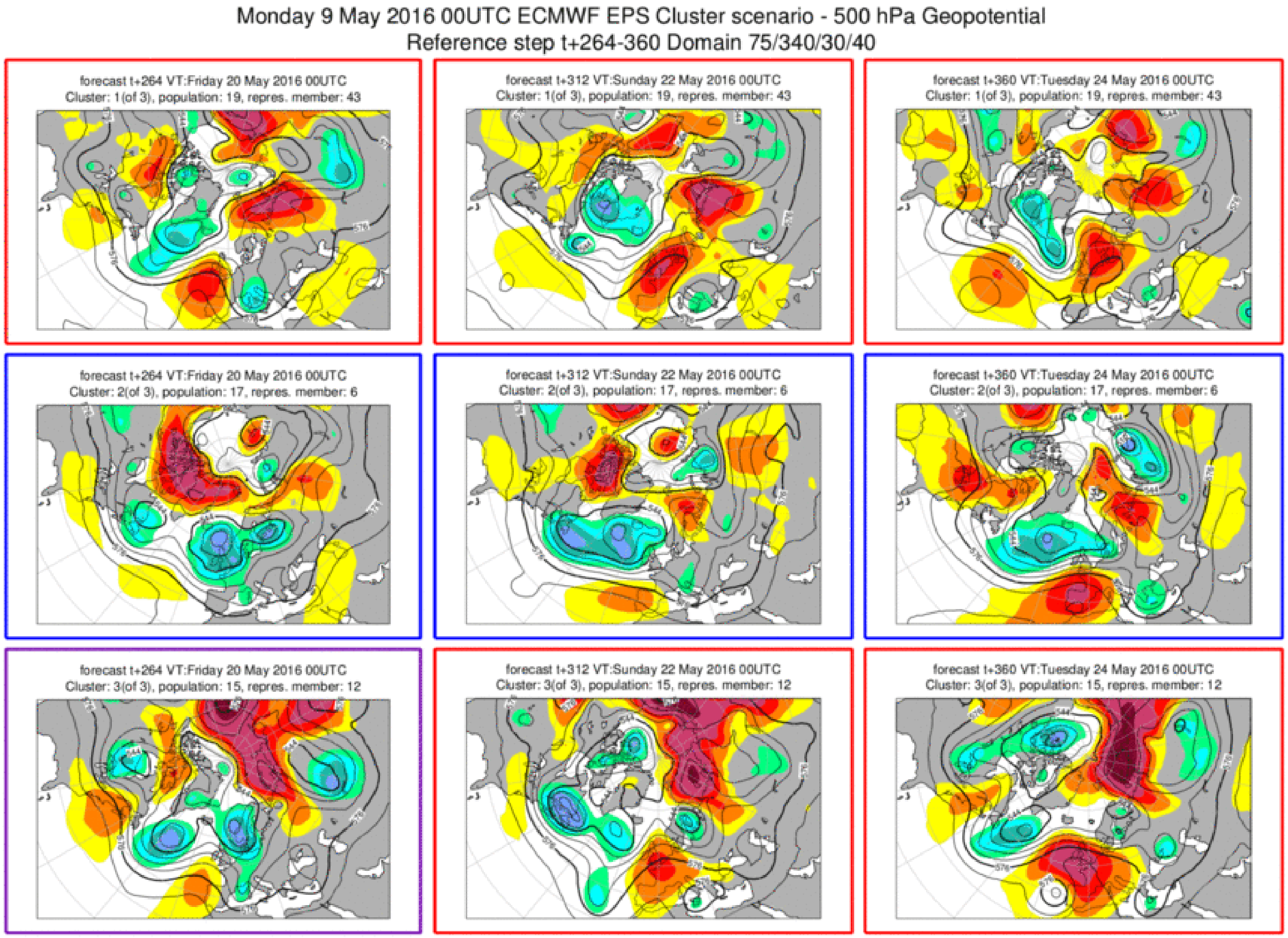
Cluster scenario

Base time

Parameter

Cluster

72-96	> 3-4 days
120-168	> 5-7 days
192-240	> 8-10 days
264-360	**> 11-15 days**



Monday 9 May 2016 00UTC ECMWF EPS Cluster scenario - 500 hPa Geopotential
Reference step t+264-360 Domain 75/340/30/40

forecast t+264 VT:Friday 20 May 2016 00UTC
Cluster: 1(of 3), population: 19, repres. member: 43

forecast t+312 VT:Sunday 22 May 2016 00UTC
Cluster: 1(of 3), population: 19, repres. member: 43

forecast t+360 VT:Tuesday 24 May 2016 00UTC
Cluster: 1(of 3), population: 19, repres. member: 43

forecast t+264 VT:Friday 20 May 2016 00UTC
Cluster: 2(of 3), population: 17, repres. member: 6

forecast t+312 VT:Sunday 22 May 2016 00UTC
Cluster: 2(of 3), population: 17, repres. member: 6

forecast t+360 VT:Tuesday 24 May 2016 00UTC
Cluster: 2(of 3), population: 17, repres. member: 6

forecast t+264 VT:Friday 20 May 2016 00UTC
Cluster: 3(of 3), population: 15, repres. member: 12

forecast t+312 VT:Sunday 22 May 2016 00UTC
Cluster: 3(of 3), population: 15, repres. member: 12

forecast t+360 VT:Tuesday 24 May 2016 00UTC
Cluster: 3(of 3), population: 15, repres. member: 12

+11days

+13 days

+15 days

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).

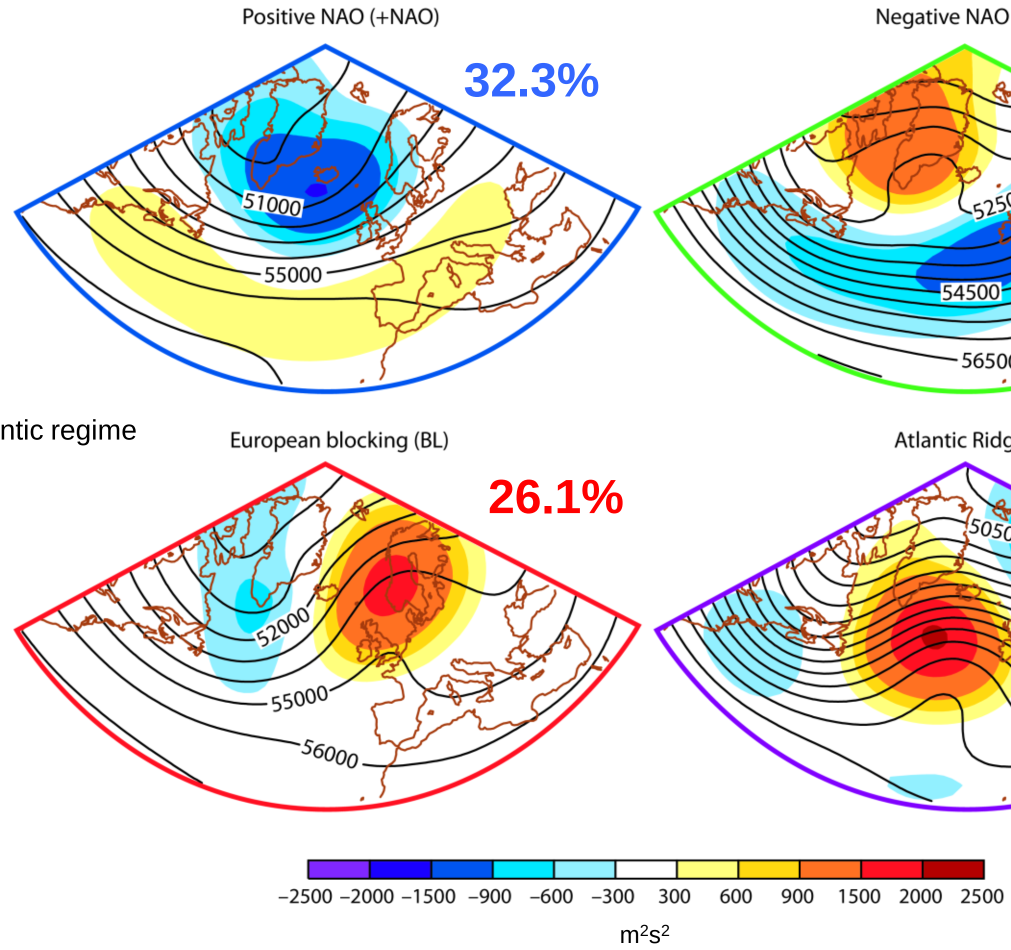
Euro-Atlantic blocking.

Negative phase of the North Atlantic Oscillation (NAO).

Atlantic ridge.

Regimes based on clustering of daily anomalies for 29 cold seasons (October to March1980-2008)

500 hPa geopotential



- Obtain well-known Euro-Atlantic regime patterns

‘k means’ clustering applied to
EOF pre-filtered data
(retaining 80% of variance)

Cluster product at ECMWF: 2-stage process

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

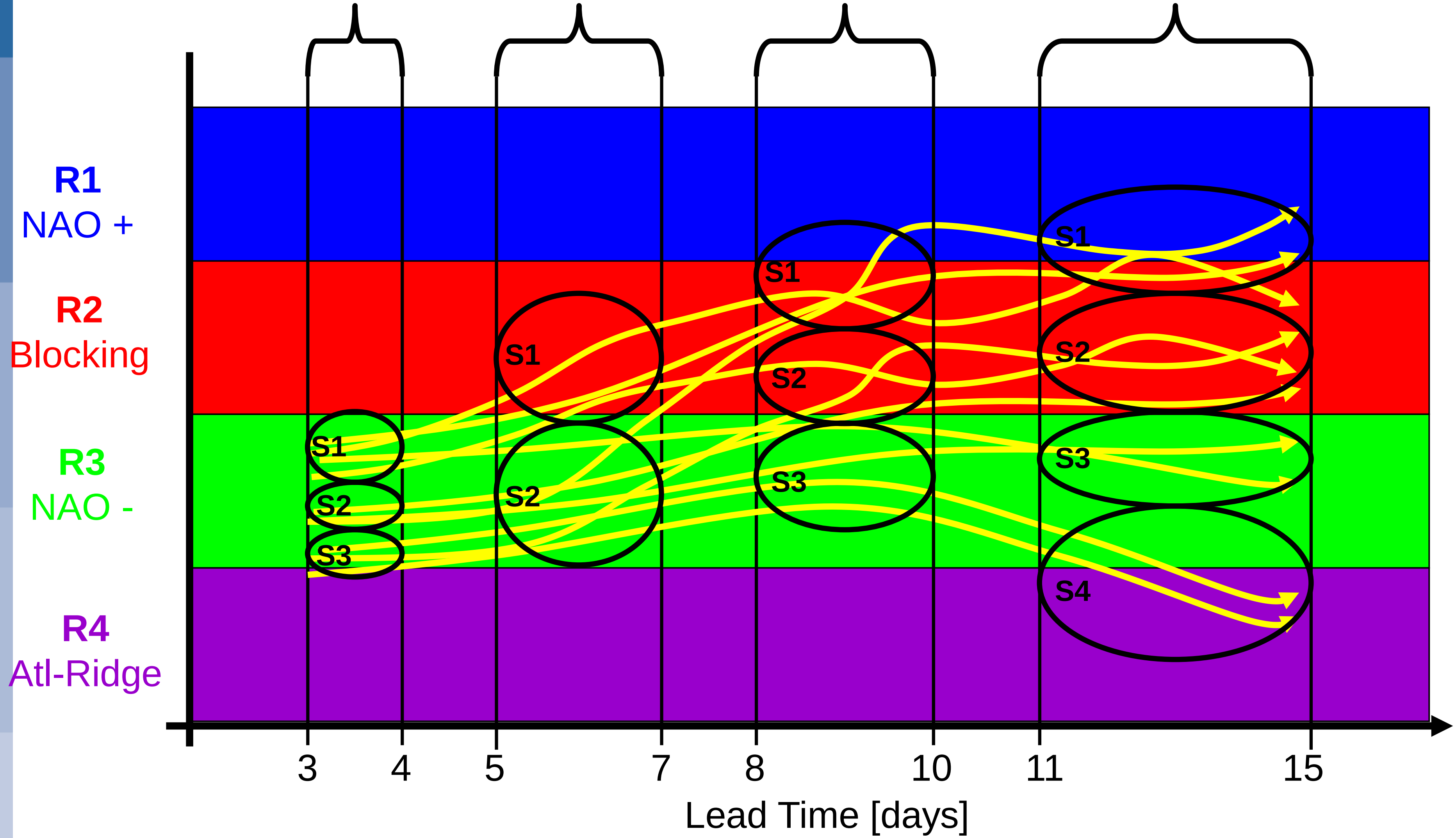
Identification of the climatological weather regimes over selected regions for every season.

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

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

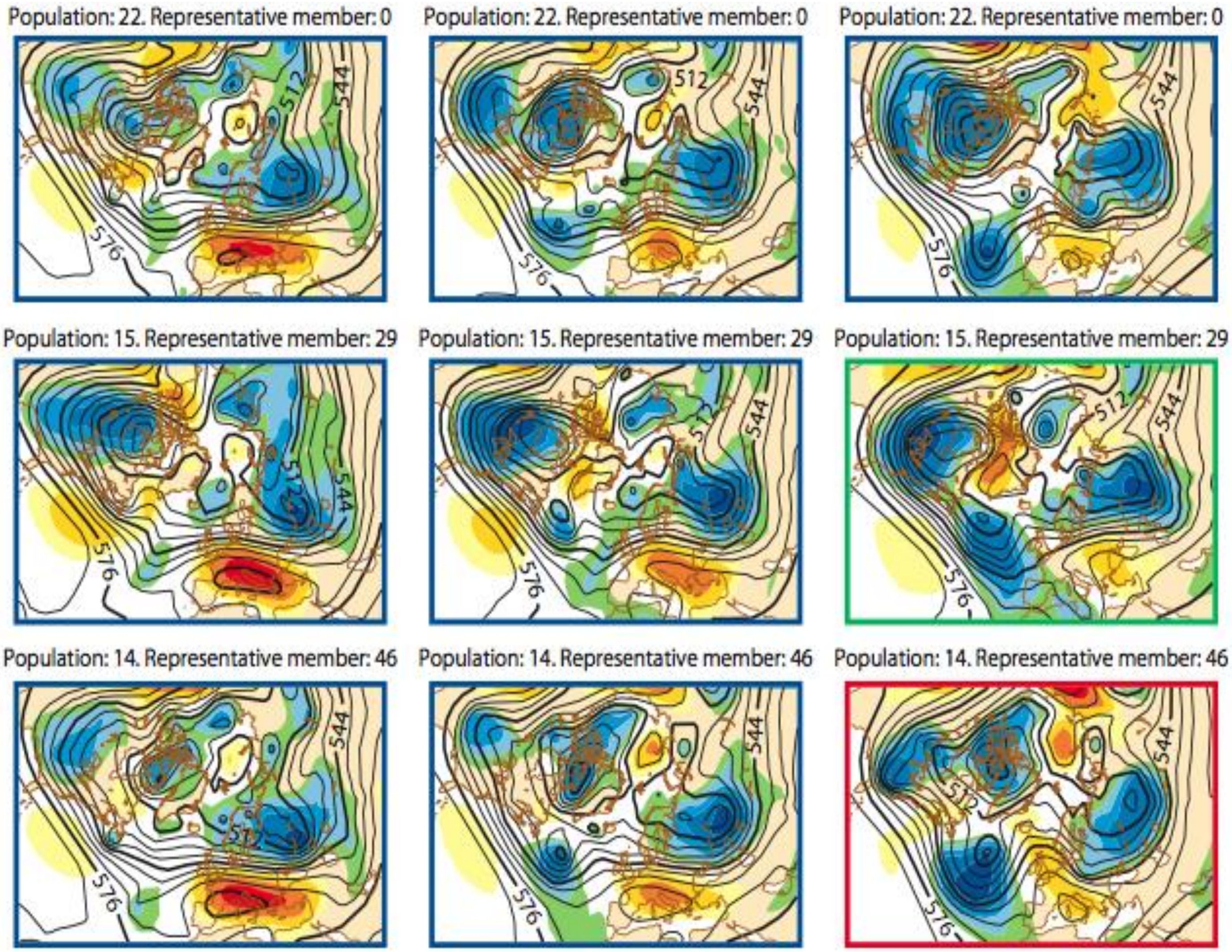
Association of each forecast scenario to the closest climatological weather regime.

Regimes & Scenarios

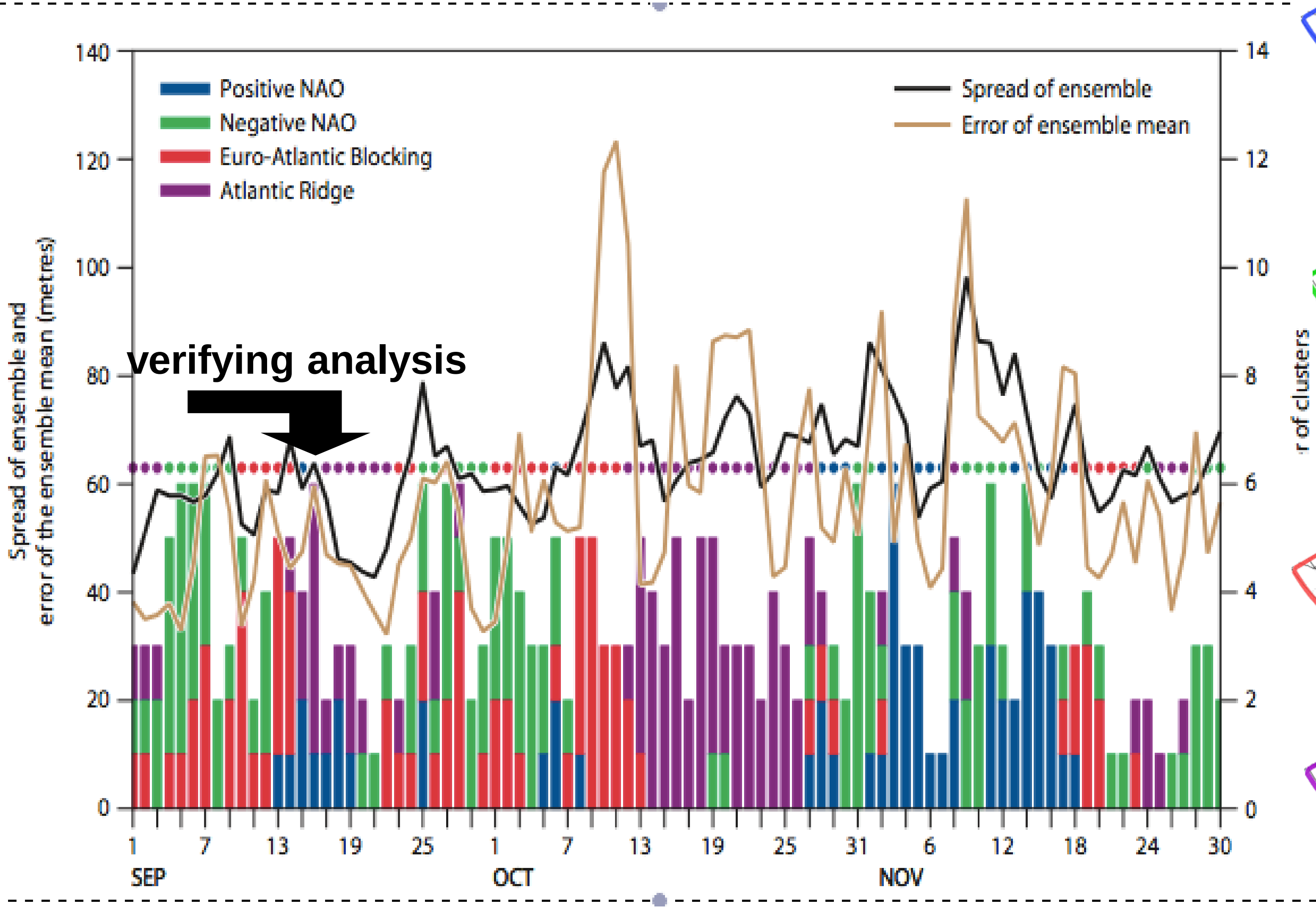


Regime transitions within a time window

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

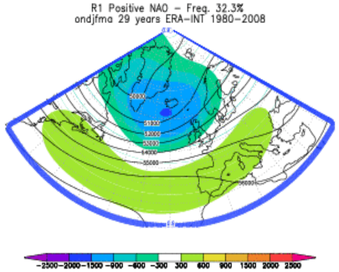

Population: 22. Representative member: 0
Population: 22. Representative member: 0
Population: 22. Representative member: 0
Population: 15. Representative member: 29
Population: 15. Representative member: 29
Population: 15. Representative member: 29
Population: 14. Representative member: 46
Population: 14. Representative member: 46
Population: 14. Representative member: 46
512
544
576

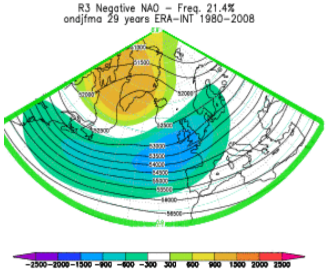
Verification & spread

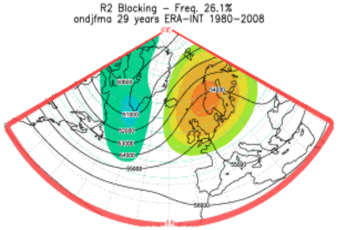


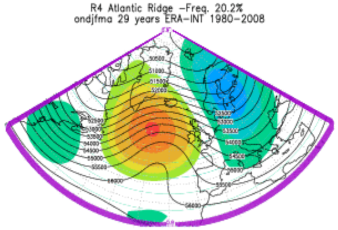
verifying analysis

Climatological regimes

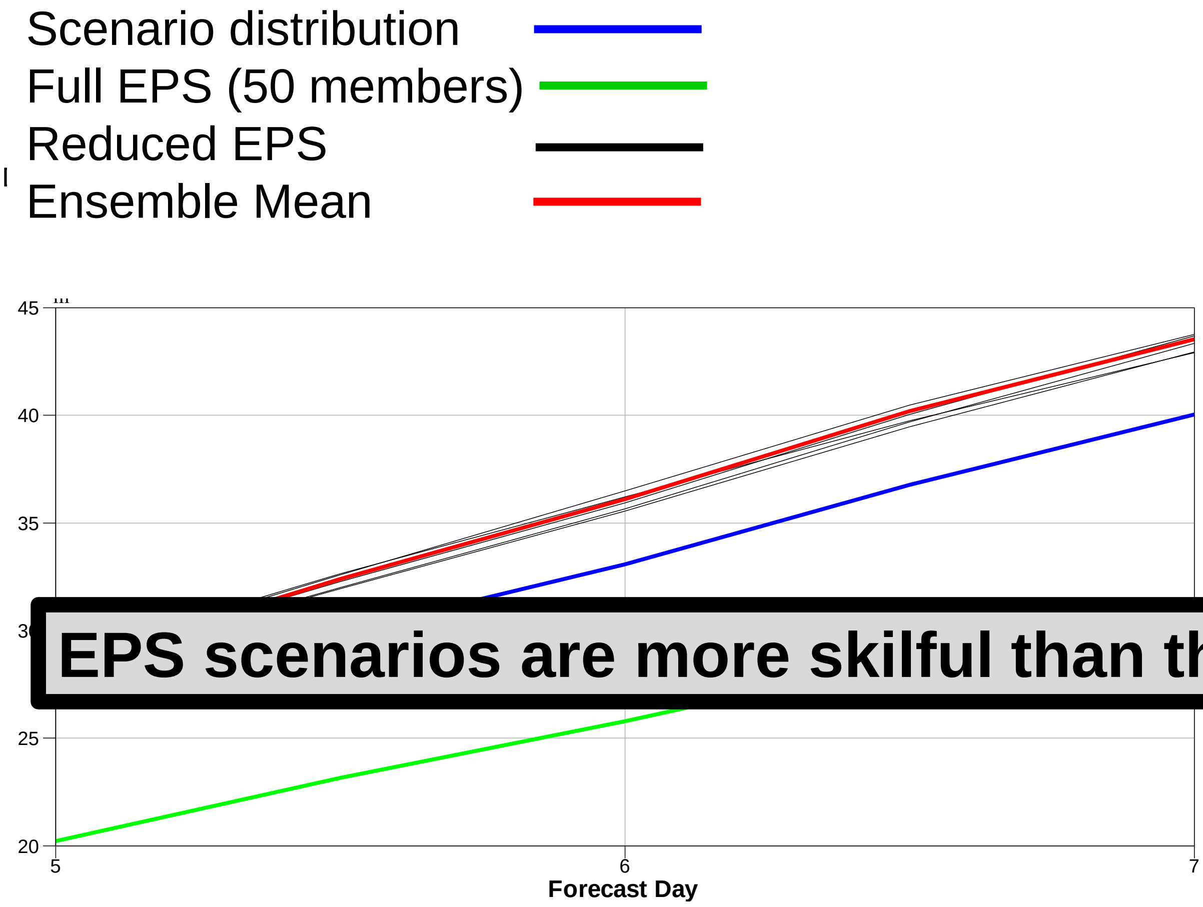

R1 Positive NAO – Freq. 32.3%
ondjfma 29 years ERA–INT 1980–2008
-2500 -2000 -1500 -900 -600 -300 300 600 900 1500 2000 2500


R3 Negative NAO – Freq. 21.4%
ondjfma 29 years ERA–INT 1980–2008
-2500 -2000 -1500 -900 -600 -300 300 600 900 1500 2000 2500




R4 Atlantic Ridge –Freq. 20.2%
ondjfma 29 years ERA–INT 1980–2008

What is the performance of the most probable scenarios?



A randomly chosen 6 member ensemble has a CRPS equivalent to that of the ensemble mean.

Probabilistic scores depend largely on the ensemble size.

A large ensemble provides a more detailed and more reliable estimate of the forecast distribution.

EPS scenarios are more skilful than the ensemble mean

ECMWF Newsletter 127

The 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

This clustering tool can be used to create EPS clusters tailored to the users' needs (e.g. different domain, different variables)

Flow dependent verification over the Atlantic sector

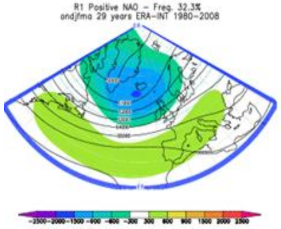
Identifying the flow configurations that lead to a more/less accurate forecast and quantifying the skill changes.

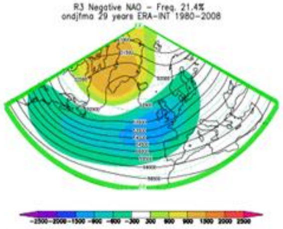
The concept of weather regimes is used to classify different flow configurations.

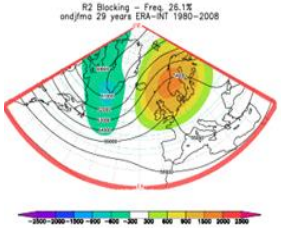
Oper. Forecast data: ENS cold season (Oct to April) 2007-2012 operational analysis

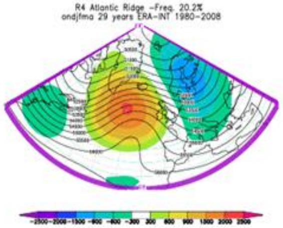
(Ferranti et al. 2014 QJRMS)

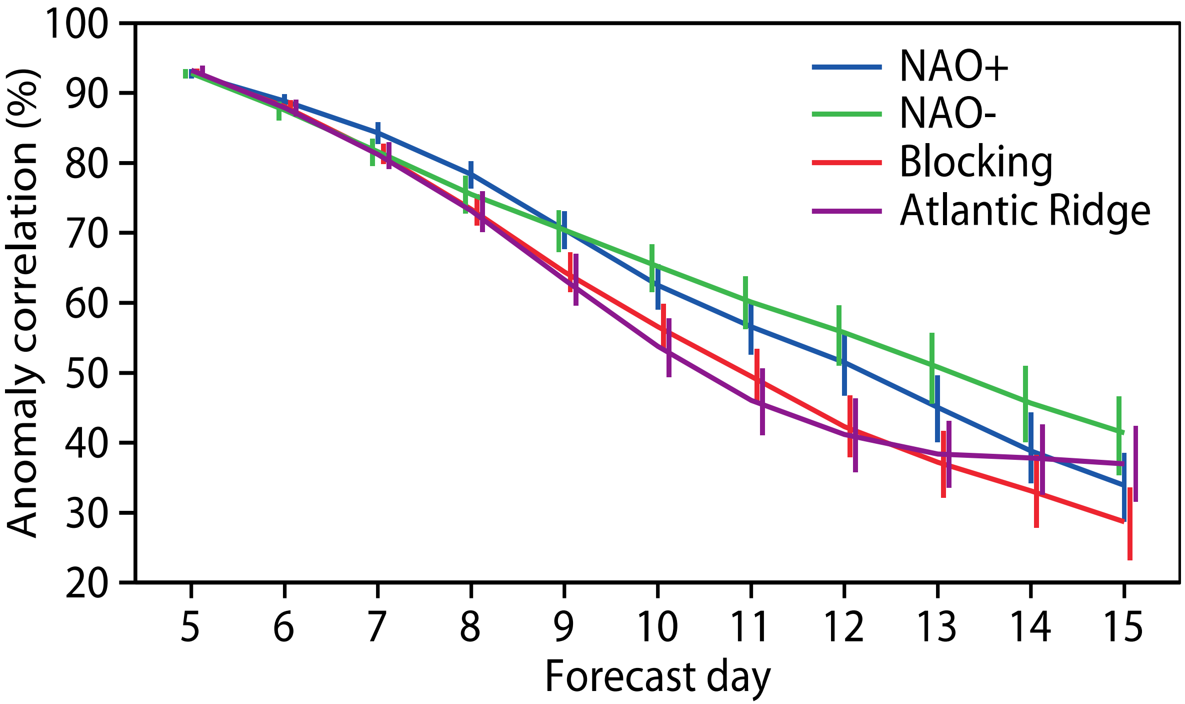
Which flow pattern leads to a more/less accurate forecasts?


R1 Positive NAO – Freq. 32.3%
ondjfma 29 years ERA-INT 1980–2008
-2500 -2000 -1500 -900 -600 -300 300 600 900 1500 2000 2500


R3 Negative NAO – Freq. 21.4%
ondjfma 29 years ERA-INT 1980–2008
-2500 -2000 -1500 -900 -600 -300 300 600 900 1500 2000 2500

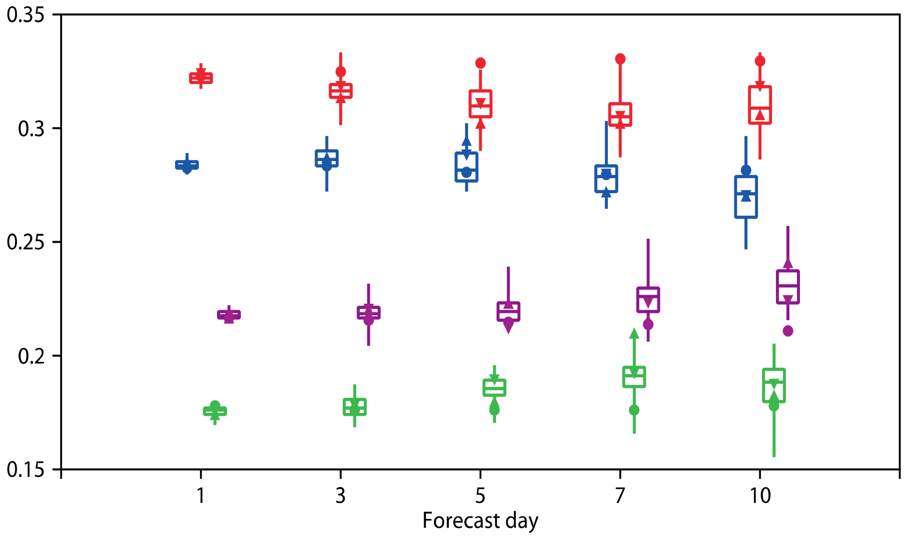



R4 Atlantic Ridge –Freq. 20.2%
ondjfma 29 years ERA-INT 1980–2008
-2500 -2000 -1500 -900 -600 -300 300 600 900 1500 2000 2500



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.

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.

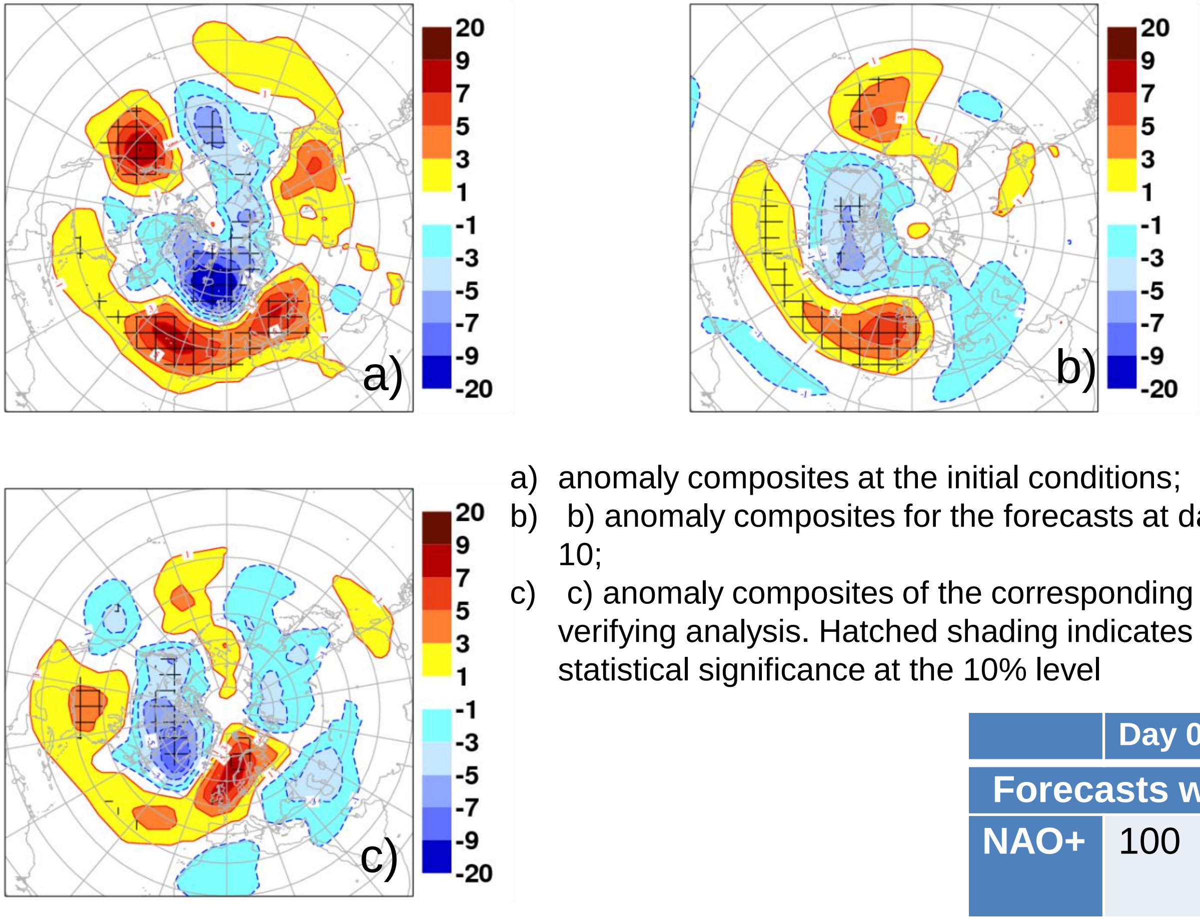
Poor forecasts at day 10

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:



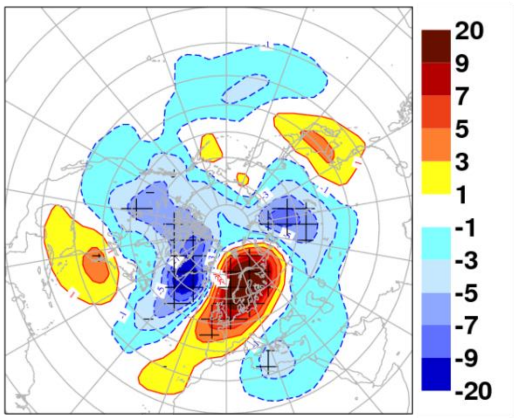
Composites of z 500 anomalies for all the forecasts initiated with flow configuration close to the NAO+ and with a RMSE at day 10 exceeding the upper quintile of the RMSE distribution.

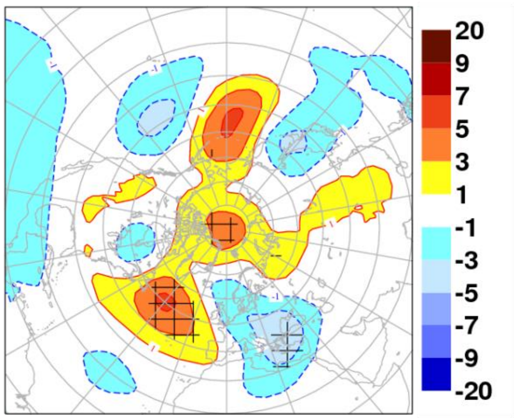
a) anomaly composites at the initial conditions;
b) b) anomaly composites for the forecasts at day 10;
c) c) anomaly composites of the corresponding verifying analysis. Hatched shading indicates statistical significance at the 10% level

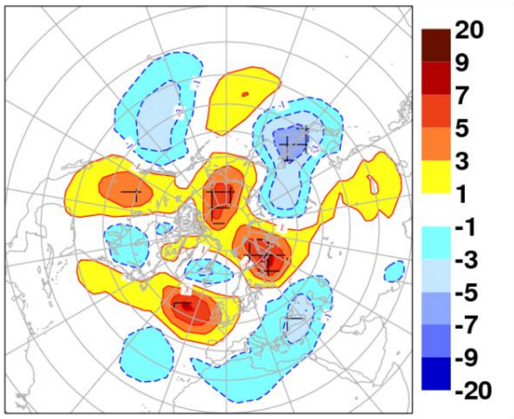
	Day 0	Day 1	Day 5	Day 7	Day 10
Forecasts with large RMSE at day 10					
NAO+	100	81	56, 44	54, 40	37, 21
BL	0	8	28, 40	35, 53	42, 51
NAO-	0	2	0	2	2, 5
AR	0	9	16	9, 5	19, 23

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

Forecasting regimes transitions:



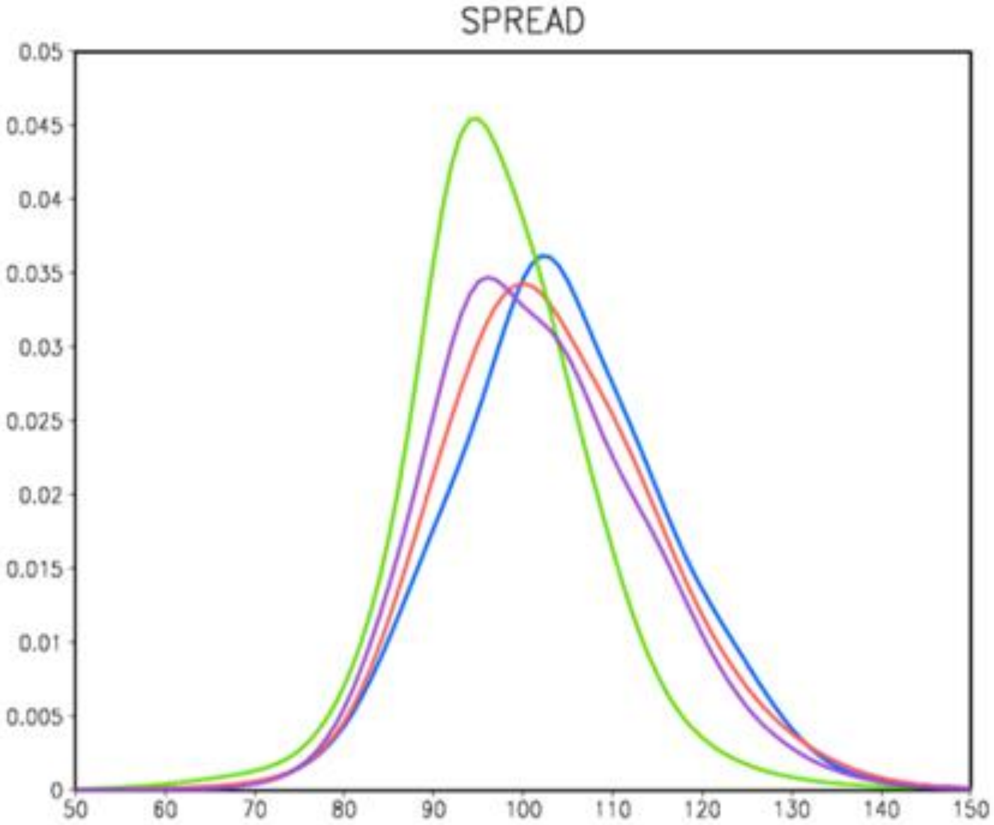




Composites of z 500 anomalies for all the forecasts initiated with flow configuration close to BL and with a RMSE at day 10 exceeding the upper quintile of the RMSE distribution.

	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



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

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.

Predictability of Euro-Atlantic circulation regimes at extended range and its association to extreme events:

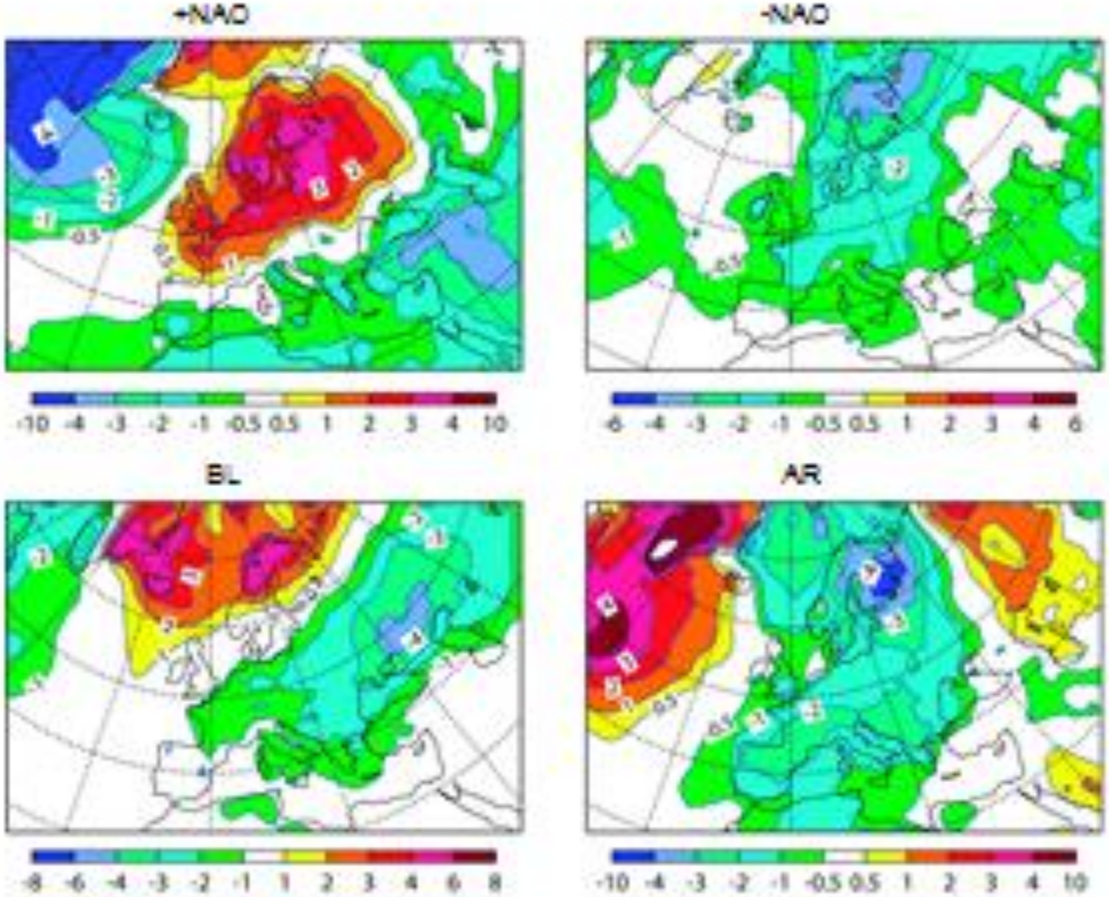
We are evaluating the predictive skill of the EA regimes using the S2S data base (Sub-seasonal to seasonal predictions WWRP/WCRP joint research project)

In particular we are interested in assessing the regime transitions (climatological frequencies, loss of skill, physical processes associated with it)

NAO- and BL are the flow patterns strongly associated with high impact temperature anomalies (heat waves in summer and cold spell in winter).

2m temperature anomalies for persistent regime episodes (> 5days) in winter

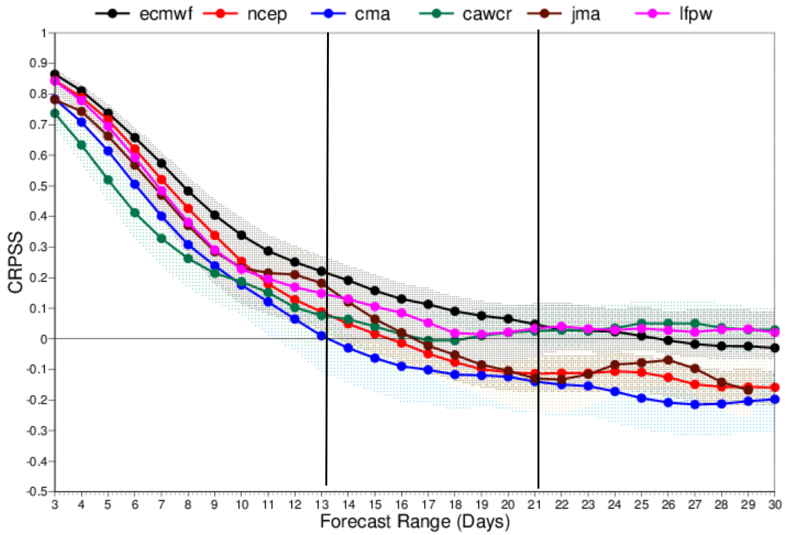
- Europe cold for 3 regimes
- BL and –NAO higher frequency of persistent events



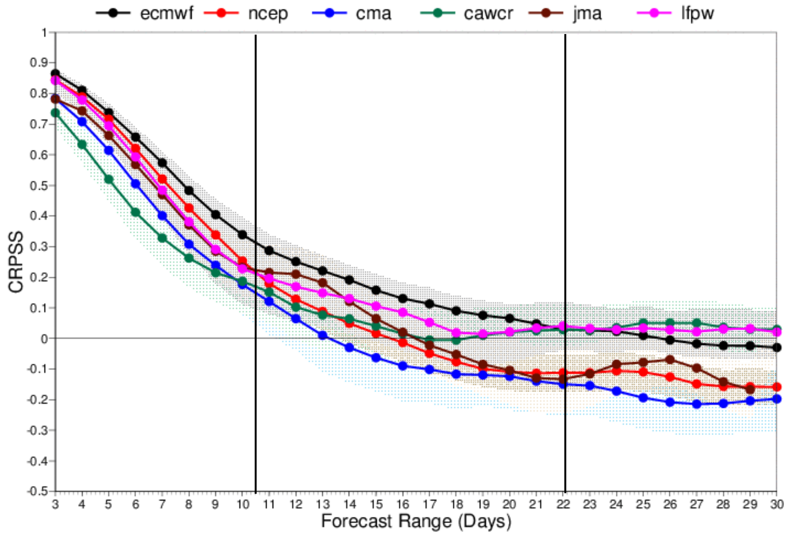
Based on re-forecast data (20 years)

How many days ahead can we predict the Euro-Atlantic regimes?

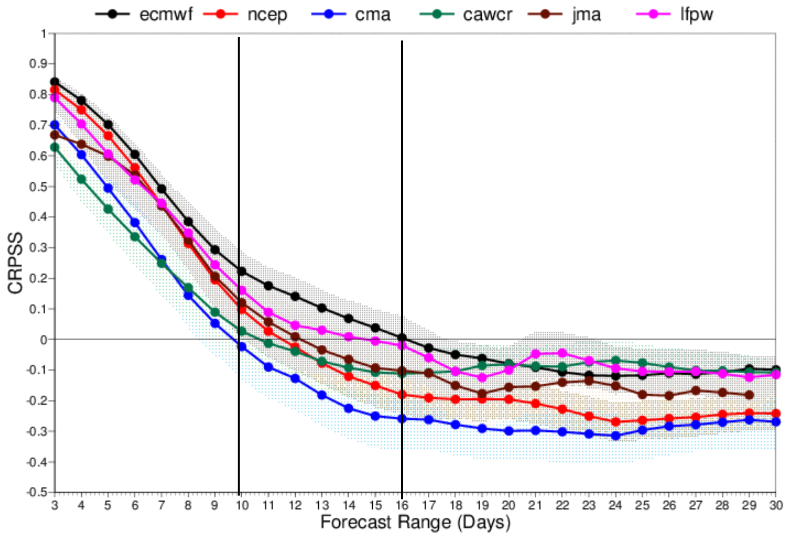
NAO +



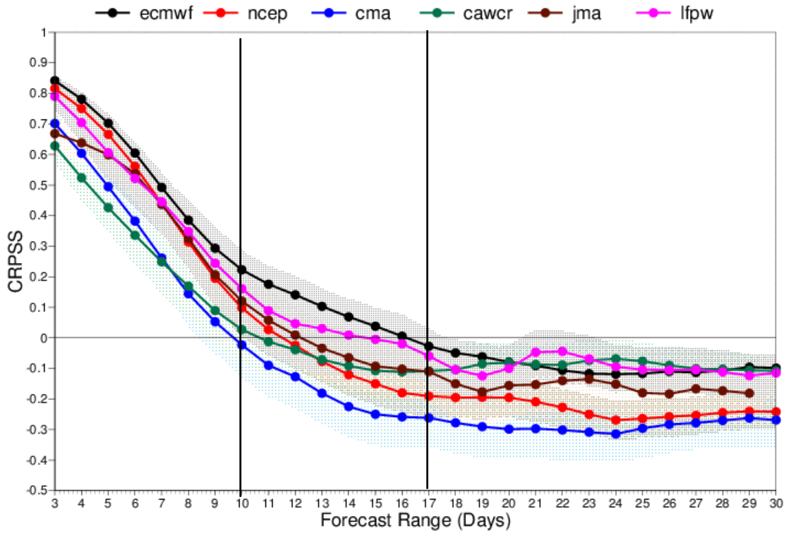
NAO -



Blocking



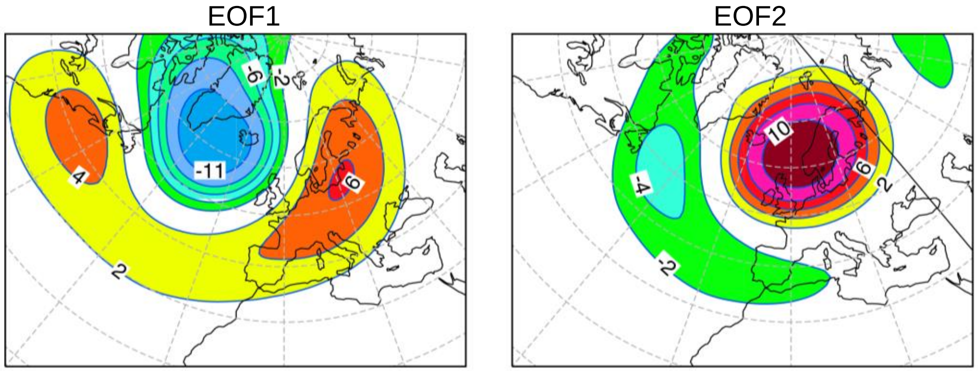
Atlantic Ridge



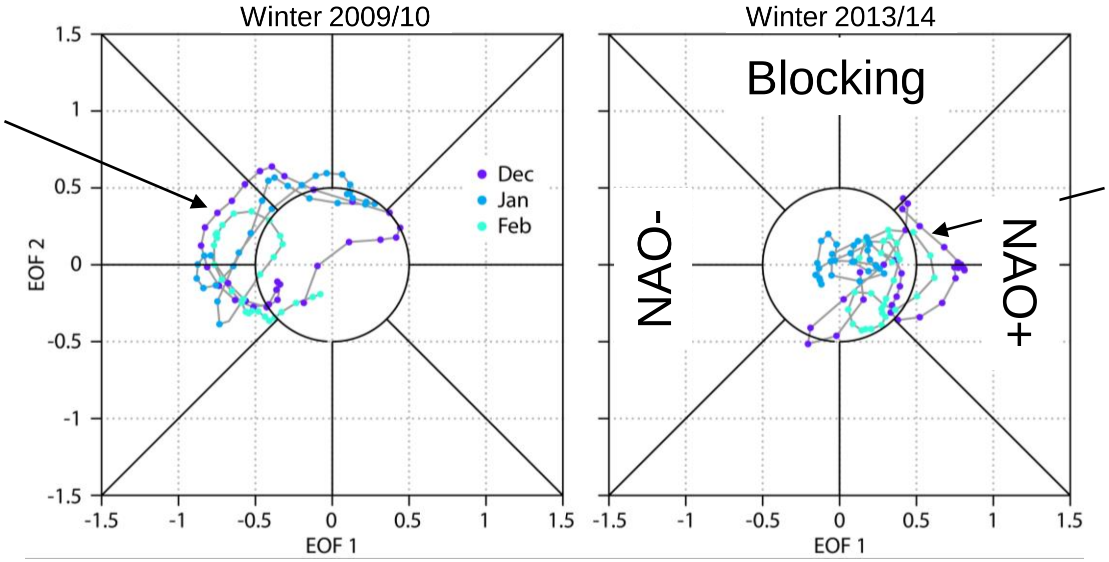
How we can evaluate the model ability in predicting regimes transitions?

Trajectories in phase space (*c.f.* MJO propagation)

- ±EOF1 and +EOF2 represent quite well ±NAO and BL
- Trajectories in phase space summarise regime evolution
- Unlike MJO, no preferred direction



BL: record-breaking cold temperatures over Europe



+NAO: exceptional storminess, but mild temperatures over Europe

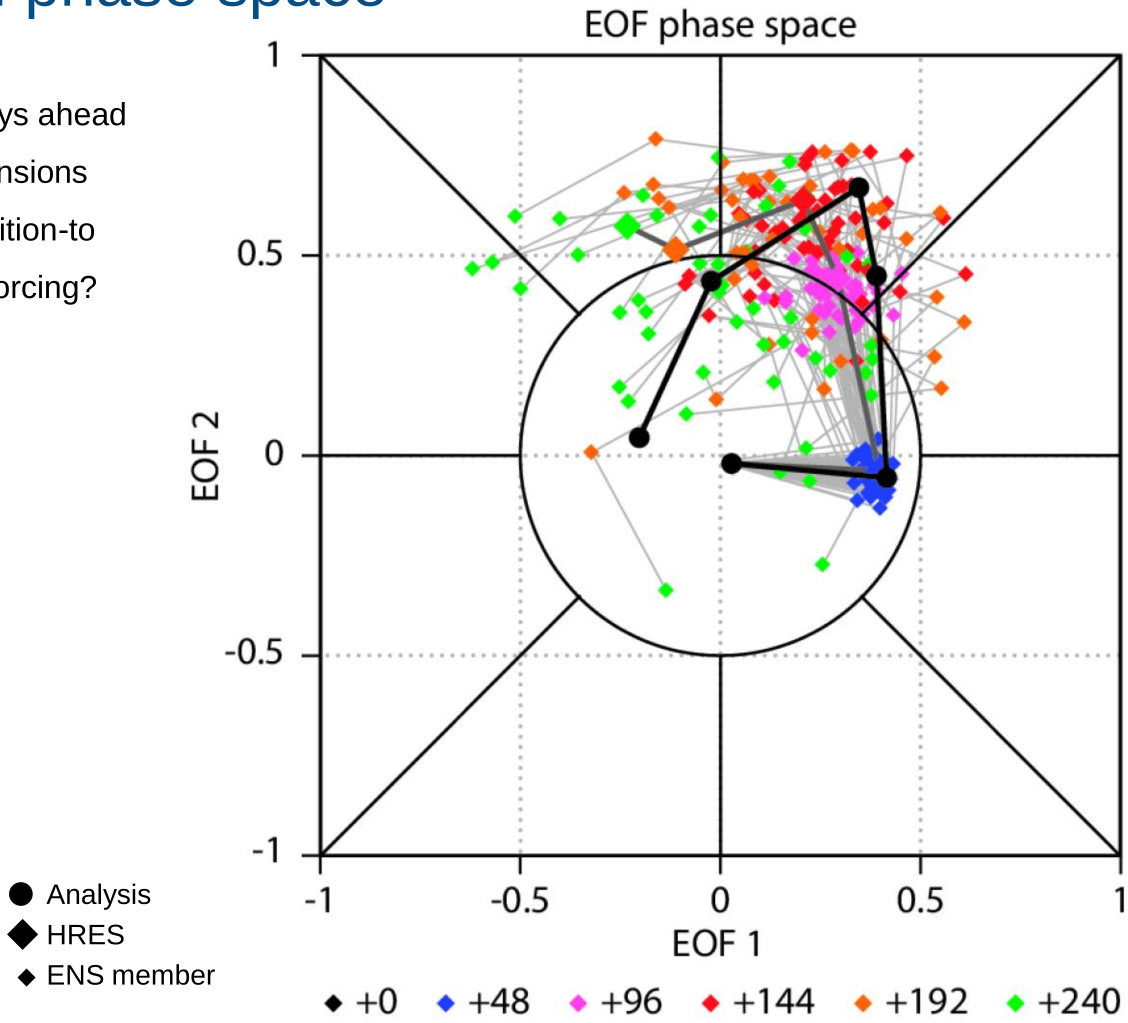
Based on 5-day running means

How we can evaluate the model ability in predicting regimes transitions?

Ensemble evolution in phase space

- Transition to blocking well-predicted 4 days ahead
- Nice way to summarise ENS in two dimensions
- Future: What processes involved in transition-to and maintenance-off blocking? Tropical forcing?

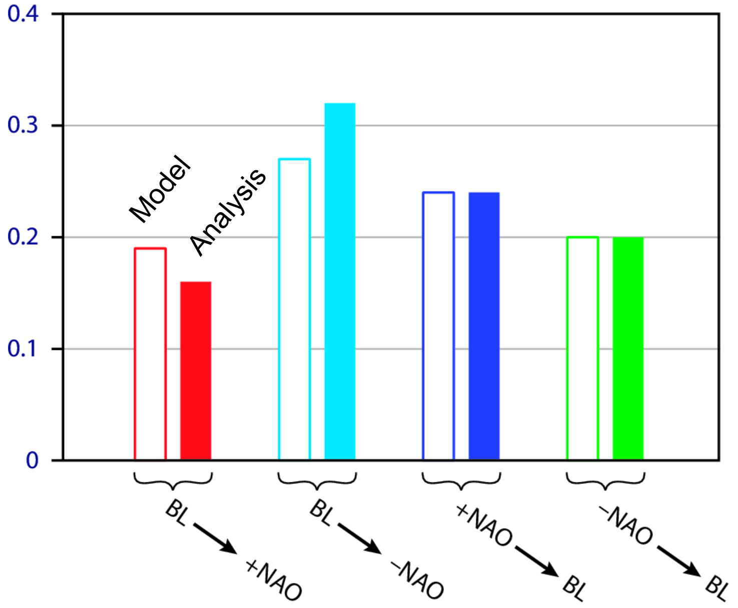
Initial date: 22 September 2015 0UTC



How we can evaluate the model ability in predicting regimes transitions?

Regime transition-frequencies and predictability

Frequencies of transitions between persistent regimes (>5 days)



- Transition frequencies good. Slight over-preference for BL → +NAO
- ECMWF has 1-2 days better skill than NCEP
- PC1 is ~2 days better than PC2 (due to high persistence of –NAO?)

5-day running mean applied prior to correlation calculation

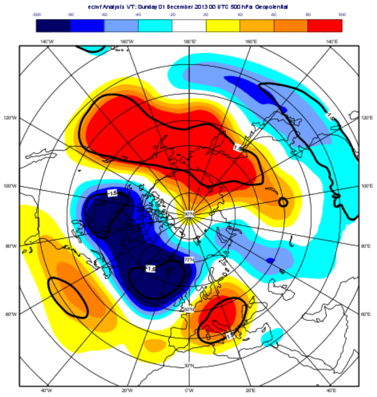
Questions?

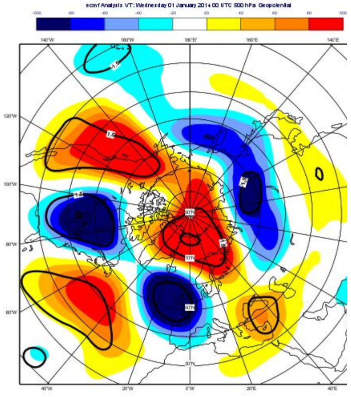
Danube, Feb 2012

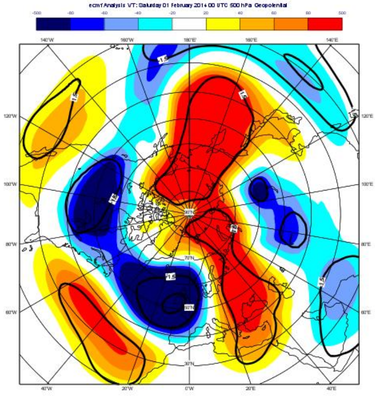
Regime analysis for DJF 2013-14 :
Monthly means

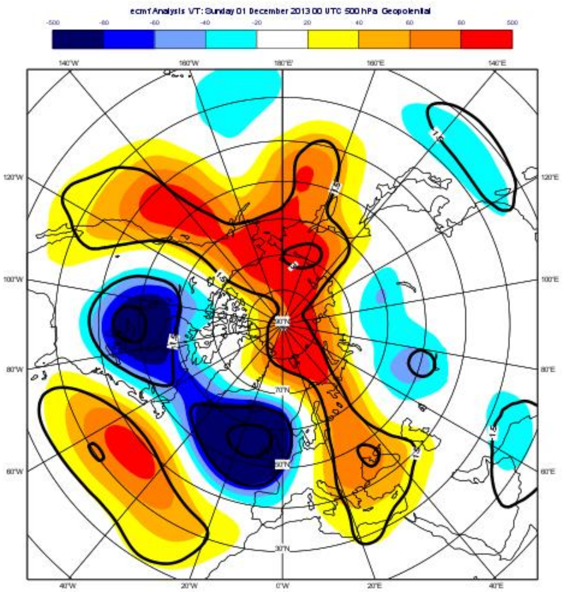
DJF 213-14 anomalies

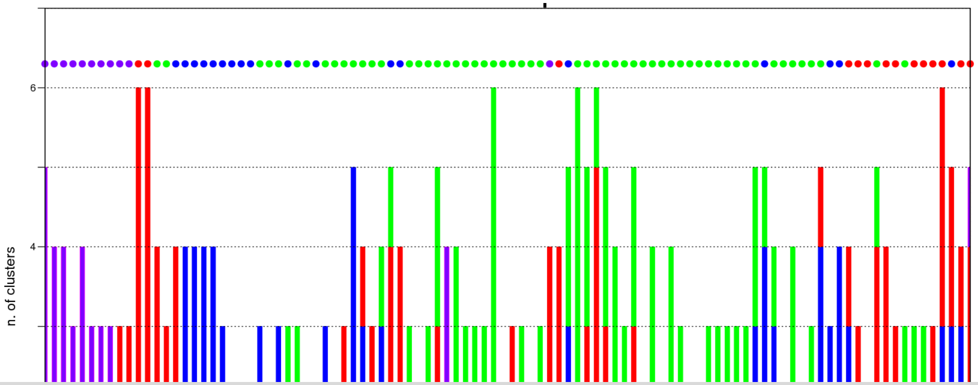
December 2013 January 2014 February 2014

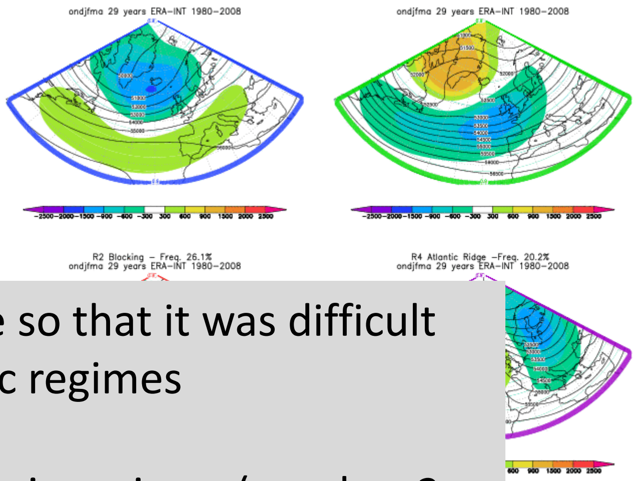

ecmf Analysis VT: Sunday 01 December 2013 00 UTC 500 hPa Geopotential


ecmf Analysis VT: Wednesday 01 January 2014 00 UTC 500 hPa Geopotential


ecmf Analysis VT: Saturday 01 February 2014 00 UTC 500 hPa Geopotential


ecmf Analysis VT: Sunday 01 December 2013 00 UTC 500 hPa Geopotential

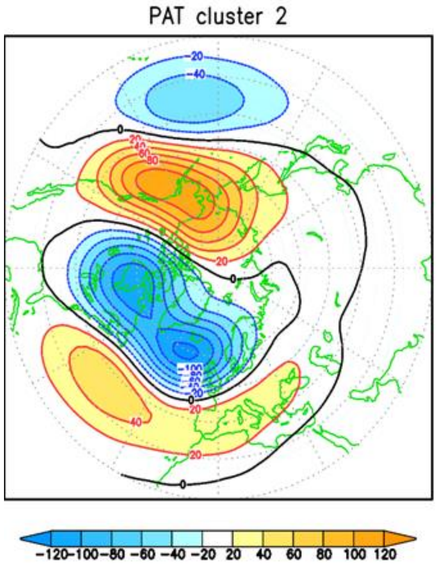


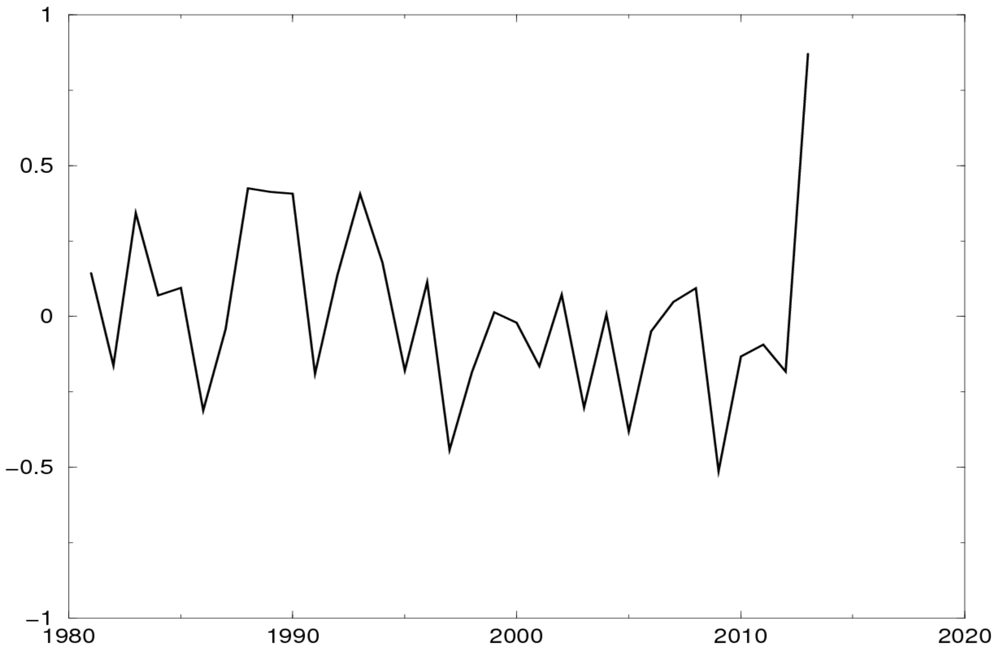

ondjfma 29 years ERA-INT 1980-2008
ondjfma 29 years ERA-INT 1980-2008
R2 Blocking – Freq. 26.1%
ondjfma 29 years ERA-INT 1980-2008
R4 Atlantic Ridge –Freq. 20.2%
ondjfma 29 years ERA-INT 1980-2008

- 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

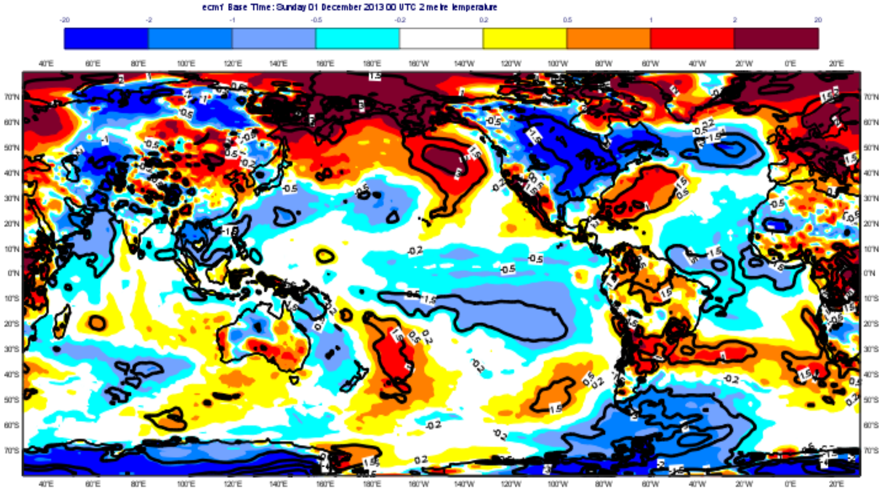


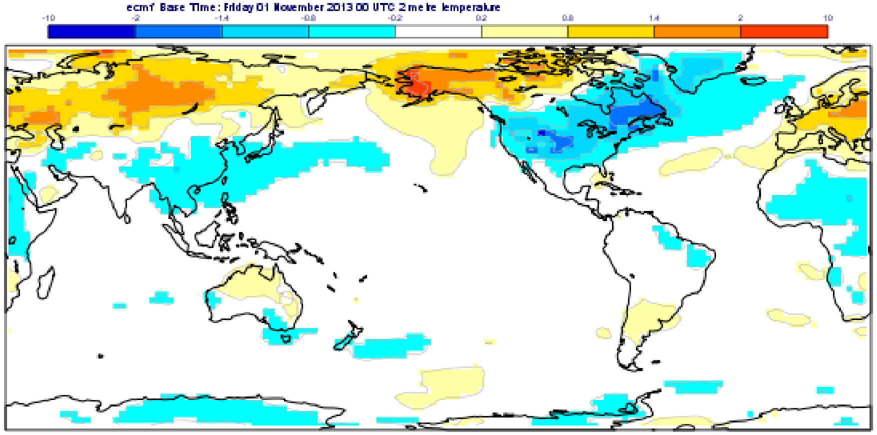


DJF 2014 anomalies:

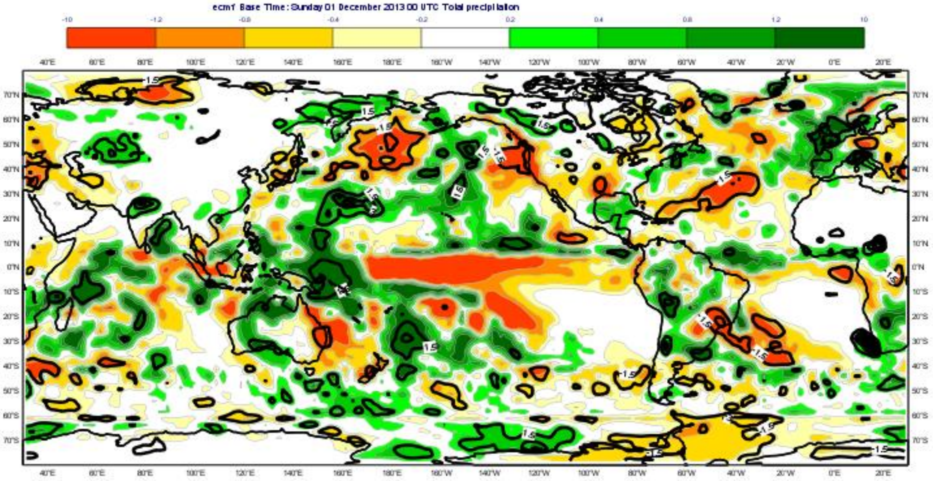
verifying analysis: Composites for proj. onto NH Cl2

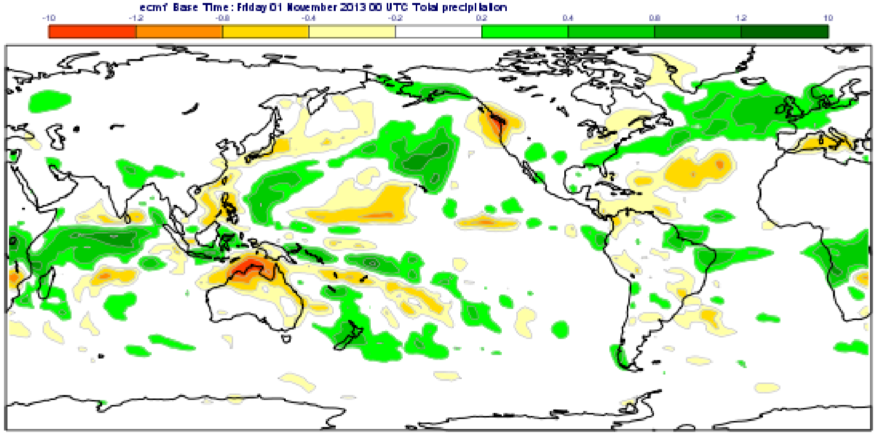
2m temp.




ecmf Base Time: Friday 01 November 2013 00 UTC 2 metre temperature
-10
-2
-1.4
-0.8
-0.2
0.2
0.8
1.4
2
10

GPCP precip




ecmf Base Time: Friday 01 November 2013 00 UTC Total precipitation
-10
-1.2
-0.8
-0.4
-0.2
0.2
0.4
0.8
1.2
10

S2S reforecasts data used for the skill assessment:

model	Bom	Cma	Ecmwf	Ncep
Rfc. lenght	0-60 days	0-60days	0-46 days	0-44 days
Resol.	T47L17	T106L40	T639/319 L91	T126L64
Rfc. size	33	4	11	4
Rfc. period	1981-2013	1994-2014	1994-2014	1999-2010
Rfc. Freq.	6/months	daily	2/weekly	daily

In order to increase the Cma and Ncep ensemble size, we have combined 3 ensemble forecasts (initiated on consecutive days) into a single 12-member ensemble. (We define the initial date to be that of the central sub-ensemble; this has little effect on results at extended leadtimes).