Barcelona Supercomputing Center
Centro Nacional de Supercomputación

WEATHER REGIMES: ECMWF SEASONAL FORECASTS VERIFICATION

BSC-ESS-2016-010

Nicola Cortesi, Nube González-Reviriego, Albert Soret and Francisco J. Doblas-Reyes

Earth Sciences Department
Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS)

1 January 2016

TECHNICAL REPORT

Series: Earth Sciences (ES) Technical Report

A full list of ES Publications can be found on our website under:

https://earth.bsc.es/wiki/doku.php?id=library:external:technical_memoranda

Barcelona Supercomputing Center–Centro Nacional de Supercomputación (BSC-CN)

C/Jordi Girona, 31 | 08034 Barcelona (Spain)

Summary

This document focuses on the assessment of the goodness of the ECMWF seasonal forecasting system S4 in simulating the observed Euro-Atlantic weather regimes from the ERA-Interim reference reanalysis and their monthly frequencies of occurrence, persistence and transition probability, as well as their impact on 10-metre wind speed and 2-metre temperature, to understand which regimes are more predictable and might lead to enhanced forecasts. Global results show that S4 is highly skilled in reproducing, at all lead times (up to six months in advance), the spatial structure of the observed WRs anomalies, their average interannual frequencies, persistences, transition probabilities and impact on wind speed and temperatures, except during autumn months, while spring and summer months perform almost as well as the winter season. However, S4 shows almost no skill in reproducing the interannual monthly variability of the frequencies of the WRs.

Contents

1. Introduction 3
2. Data and methodology 5
2.1. Data and pre-processing 5
2.2. Methodology 6
3. Results 9
3.1. Observed monthly regimes 9
3.2. Pattern correlations 11
3.3. Frequency correlations 13
3.4. Average frequency bias 16
3.5. Persistence bias 17
3.6. Transition probability bias 18
3.7. Impact of WRs on wind speed 19
4. Conclusions 23
5. Appendix 24
6. Acknowledgements 27
7. References 28

Index of figures

Figure 1. Flow chart of the observed and simulated WRs classification and of the impact analysis. 8

Figure 2. ERA-Interim winter (DJF) SLP regime anomalies. 9

Figure 3. Taylor diagrams of the regime patterns. 10

Figure 4. Simulated regime anomalies for January. 12

Figure 5. Pattern correlation 13

Figure 6. Frequency correlation. 14

Figure 7. Simulated interannual frequencies of occurrence for January. 15

Figure 8. Bias of the average frequency 16

Figure 9. Persistence bias.....18

Figure 10. Transition probability bias.....19

Figure 11. WRs impact on 10-m wind speed for target month of January.....21

Figure 12. Wind Speed's impact correlation. 22

Figure 13. Taylor diagrams of the monthly regime patterns for the two clustering methodologies.....25

1. Introduction

Large-scale circulation is often studied by classifying the high number of all possible atmospheric configurations in a small number of recurrent and quasi-stationary states called 'weather regimes' (WRs). The transitions between WRs determine a large part of the intra-seasonal circulation variability (Michelangeli et al., 1995). It has been demonstrated that the response of the atmosphere to external forcings, such as the enhanced greenhouse gas concentration, is partly determined by a change in the frequency of the occurrence of the WRs (Corti et al. 1999). WRs are usually defined by performing clustering algorithms on an observed circulation variable such as the geopotential height or the mean sea level pressure (Dawson et al., 2012).

Most studies on WRs in the northern hemisphere focus on the winter season (Fil & Dubus, 2005, Fereday et al., 2008, Vrac et al., 2014, Ferranti & Corti, 2011) or the whole cold season from October to April (Ferranti et al., 2015) because in these months WRs are more stable in time (more persistent) and have a stronger influence on local climate. The number of WRs selected depends on the spatial domain and the type of study. For the North Atlantic-European region, the more commonly used approach is based on the Bayesian information criteria (BIC) which identifies four WRs (Yiou & Nogaj, 2004). This choice also maximize the representativeness of each regime, since in the seasonal time range, the high ensemble spread of the forecasts and its high associated uncertainty make it difficult to correctly discriminate between different WRs, if the classification relies on a high number of regimes.

In winter, two of the four WRs obtained in the Euro-Atlantic region are consistent with the spatial patterns of the two opposite phases of the North Atlantic Oscillation (Hurrell & Deser 2009) and are therefore called "NAO+" and "NAO-" regimes. A third WR is called "Blocking" (BL), since it has a positive anomaly centered over Scandinavia and a negative one centred over the Atlantic Ocean, similar to the atmospheric flow during BL events in Europe. The remaining winter Euro-Atlantic WR is called "Atlantic Ridge" (AR) and it is represented by a positive anomaly over the Atlantic Ocean and a negative one over Scandinavia.

Recently, the 'Laboratoire des Sciences du Climate et de l'Environnement' (LSCE) released an online tool to classify Euro-Atlantic WRs seasonally, to the benefit of the scientific community (https://a2c2.lsce.ipsl.fr/index.php/deliverables/102-continuous-time-weather-regimes). WR classifications can be established also in summer, albeit with different patterns and employing the geopotential height (Cassou et al., 2006), while the high temporal variability of the Euro-Atlantic atmospheric circulation in the intermediate spring and autumn seasons makes more difficult to classify a set of stable WRs during these periods.

Because of the chaotic dynamic of the atmospheric circulation, certain WRs are more stable in time than others and hence more predictable (Ferranti et al., 2015). As a consequence, the skill of a forecast system is state-dependent. It is important to understand which WRs are predictable and which are unpredictable, to know beforehand which ones lead to a more predictable configuration than others, thus enhancing the forecast skill without having to modify the forecast system itself (Neal et al., 2016).

This study focuses on the assessment of the skill of the European Centre for Medium-Range Weather Forecasts (ECMWF) seasonal forecasting System 4 (S4) in simulating the observed WRs in the Euro-Atlantic region. To the best of our knowledge, this publication is the first documented attempt of assessing the WRs obtained from a seasonal prediction system along all the months of the year, taking into account different start dates and all possible lead times. Up to now only for medium-range weather forecasts (first two weeks) there already exists two publicly available WRs classifications aimed at validate the forecast skill using atmospheric flow classifications (Neal et al., 2016 and Ferranti et al., 2015). Both publications agree that the NAO+ and NAO- regimes are the ones with the highest forecast skill. NAO+ regime has the highest annual skill, while NAO- regime the highest wintertime skill during the study period. Skill is also higher in winter than in summer. Ferranti et al. (2015) detected a systematic overestimation of the mean frequencies of occurrence simulated by AR regime, compared to the ERA-Interim reanalysis. Moreover, BL regime leads to less skilful predictions, consistent with previous studies (S.Tibaldi & F.Molteni, 1989; Pelly & Hoskins, 2003) which highlighted that transitions to BL regime are very difficult to predict for medium-range forecasts.

This study is organized as follows: Section 2 introduces the reader to both forecast and verification datasets employed, describing the methodology to define the observed and simulated WRs. In Section 3, we present the results, first in term of the observed WRs, then by means of the comparison between observed and simulated WRs: which of them lead to more accurate predictions in terms of spatial and temporal correlations, frequency of occurrence, average frequency, persistence and transition probabilities. Their simulated impact on wind speed and temperature, the two variables selected to fulfil the objectives of task 1d1 of RESILIENCE Project, is also compared with the reference impact. Finally, Section 4 is dedicated to the discussion and the general conclusions.

2. Data and methodology

2.1. Data and pre-processing

Observed daily-means data of sea level pressure (SLP), 10-m wind speed and 2-m temperature from ERA-Interim reanalysis dataset (Dee et al. 2011) were employed, with a native spatial resolution of 0.75°. Forecast daily-means data of SLP, 10-m wind speed and 2-m temperature proceeds from the ECMWF-S4 with a spatial resolution of about 80 km and 15 ensemble members during the hindcast period 1981-2015 (Molteni et al. 2011). S4 data was re-gridded to the same resolution of ERA-Interim (0.75°) with a bilinear interpolation, and daily-means data were computed as average of 6-hourly raw data (0000, 0600, 1200 and 1800 UTC). The daily means of the SLP were considered in this project instead of the geopotential height (which provides better results in summer, see Cassou et al. 2006) because its time series doesn't show any temporal trend (Hafez & Almazroui 2014), so it was preferred to the geopotential.

The ECMWF-S4 is an operational seasonal prediction system based on a fully coupled global climate model which seasonal forecasts are probabilistic in nature as several independent model simulations (ensemble members) are provided. Each simulation has different perturbed initial conditions and/or different physical parameters (Doblas-Reyes et al., 2009). Current version of S4 stores forecasts up to six months in advance, for a total of seven different lead times. Start dates correspond to the month of the year where the prediction is started, while lead times indicate the number of months between the start date and the target month to be predicted (ranging from 0, when the target month is the same of the start date, up to six months in advance). The presence of several ensemble members allows the estimation of the forecast uncertainty. This way, there will be predictions for some days with a small spread between the different ensemble members, corresponding to a quite predictable atmospheric state, or days with large spread, indicating a more unpredictable state.

All the above-mentioned variables for both datasets were extracted only for the Euro-Atlantic region (27°N–81°N, 85.5°W–45°E) and the period 1981-2015, the common period between the reanalysis data and the forecast data. Before applying the methodology, we obtained the daily anomalies for each individual month along the year and for both reanalysis and forecasts datasets separately. The anomalies were referred to the daily climatology of the respective dataset for the period 1981-2015 previously filtered with a LOESS polynomial regression (degree of smoothing α = 0.35) to remove the influence of both the annual cycle and the short-term variability (Mahlstein et al. 2015). While daily SLP anomalies were employed for WRs classification and to obtain their corresponding anomalies maps and frequencies, daily wind speed and temperature anomalies were used to obtain the impact of the WRs on these two variables. A diagram with all the steps followed is shown in Figure 1.

2.2. Methodology

Several variants of the WRs classification exist in the scientific literature. Most of them perform a principal component analysis (PCA or EOFs) before classifying the WRs. However, in this study the PCA filtering was not applied to take into account also the more extreme SLP values, which are usually filtered out by the PCA process. WRs were classified with the k-means cluster algorithm of Hartingan and Wong (1979), with 4 clusters, 30 random starts and 100 maximum iterations, enough to provide stable WRs. In the Euro-Atlantic region, the four clusters correspond to the NAO+, NAO-, BL and AR WRs (Vrac et al., 2014).

The monthly WRs classification was applied to ERA-Interim daily SLP anomalies covering the period 1981-2015, as a baseline with which to compare the predicted WRs. As sampling daily mean SLP over monthly periods might not be sufficient to adequately represent the clustering space, as exposed by Dawson et al. (2012), a 3-months running k-means cluster, centred in the target month, was applied (Figure 1). Its output is a 3-months daily time series for each year with four possible values that corresponds to the four clusters. Nevertheless, in order to obtain the WRs anomaly pattern for a specific month, only days inside the central month of each year were taken into account. For example, for January, the k-means clustering were applied to December-January-February data but only values belonging to January were selected from the output. This way, the SLP anomalies of days belonging to the central month and the same cluster were averaged over the period 1981-2015, obtaining the SLP pattern anomalies associated with that regime (hereafter regime anomalies). A full description of the reasons to introduce a 3-months running cluster is presented in the Appendix.

In a second step, a monthly WRs classification was obtained from predicted S4 daily SLP anomalies, applying the clustering to each of the 12 individual start dates and each of the 7 lead time separately. In case of S4, the running cluster technique was not applied because the ensemble members in the forecast provides a large number of simulated SLP data (15 times that of ERA-Interim), so the S4 clustering is more robust than the ERA-Interim clustering (see Appendix). As a result of the clustering of S4, fifteen daily time series were obtained, each one with four possible clusters, one series for each ensemble member. Each cluster was associated to the ERA-Interim regime that more closely resembled, in term of the spatial pattern correlations for the whole Euro-Atlantic region. The simulated regime anomalies were obtained as an average over the period 1981-2015 of the S4 SLP anomalies of all days associated to a given WR, start date and lead time.

Once the regime anomalies were identified for both reanalysis and forecasts, the monthly frequencies of occurrence of each WR from 1981 to 2015 were measured; in case of ERA-Interim, frequencies were measured simply by counting the number of days a certain WR is observed in a month and normalizing for the number of days of the month; in case of S4, the monthly frequencies were obtained first by counting, for each ensemble member, the number of days a certain WR is predicted by that member, then summing this value for all the ensemble members, and finally normalizing for the number of days of the month and the number of members (15). A Mann-Kendall test was introduced to identify significant trends in the frequency time series.

As illustrated in Figure 1, the impact of a WR on wind speed or temperature was measured averaging 10-metre wind speed (or 2-metre temperature) anomalies of all days associated to a given WR and month (or WR, start date and lead time for S4). Impact maps were calculated only for the European region, because the main interest in this project is to measure the impact over land and the surrounding sea shore, and not in the middle of the ocean. A t-test was performed to assess the level of significance of the anomalies of the impact maps (with respect to the wind/temperature anomalies).

Finally, seven metrics were selected to validate, for each start date and lead time, S4 skill in reproducing the WRs defined with ERA-Interim. First metric was “Pattern Correlation”, which is the spatial correlation obtained from correlating the predicted and observed regime anomalies patterns. Second metric was “Frequency Correlation”, defined as temporal correlation between the simulated and observed monthly frequencies time series of the WRs. Then, three metrics focused on different bias were introduced to measure the difference between the simulated and observed values of a particular index: the “Average Frequency Bias” measures the difference between the simulated and observed average monthly frequency of occurrence of a WR during 1981-2015; the “Persistence Bias” is defined as the mean number of days before a WR is replaced by a new one (typically 3-5 days for Euro-Atlantic WRs); the “Transition probability bias” measures the frequency by which a WR is replaced by a different one. Finally, the last two metrics focused on the comparison between the simulated and observed impact of the WRs on 10-metre wind speed and 2-metre temperature, by means of the spatial correlations of their impact patterns.

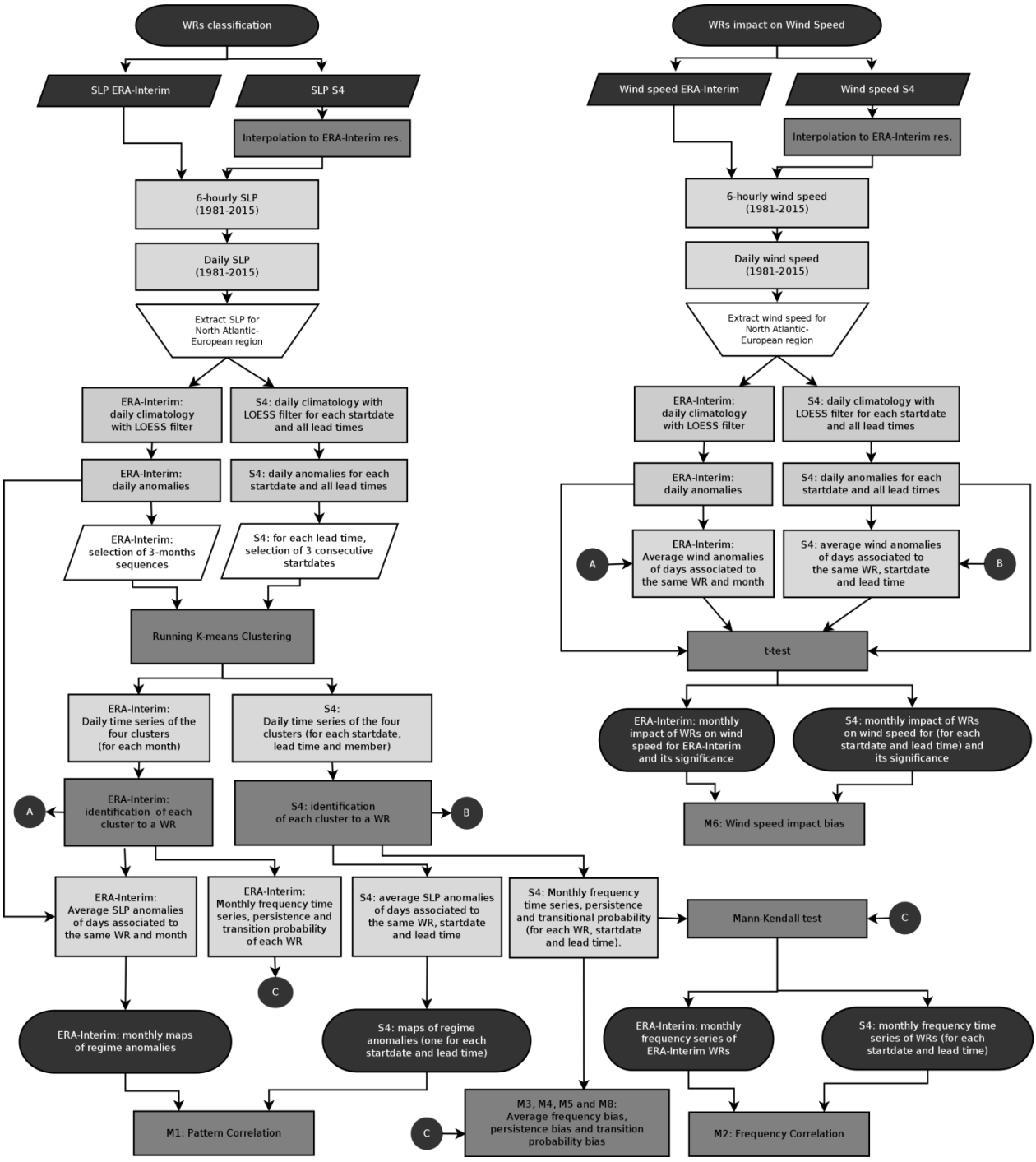

***Figure 1**. Flow chart of the observed and simulated WRs classification and of the impact analysis.*

Flow chart with the sequence of the steps followed to obtain the classification of the WRs and their impact on wind speed, for both reanalysis and forecast data, starting from raw data up to the extraction of the regime anomalies and frequencies. A similar sequence was adopted for the impact of the WRs on temperature.

3. Results

3.1. Observed monthly regimes

Usually, European WRs are classified during winter (DJF) season (see Figure 2) or during the extended winter season (NDJFM). To extend the WR classification to each month of the year, it must be able to reproduce individual monthly spatial regime patterns similar to those observed in winter season (DJF), at least during the three winter months. Moreover, monthly regime patterns should not show strong differences from one month to the next, i.e: each WR should have an high intermonthly pattern continuity.

To evaluate if the methodology based on the 3-months running clustering satisfies these two proprieties, a Taylor diagram with the comparison between the regime patterns for each individual month and the winter (DJF) regime patterns is shown in Figure 3. January is the winter month which more closely resemble the regime patterns of the winter class ($r > 0.99$), followed by February and December ($r > 0.95$). Even outside winter months, there are months such as March and November which show a good correspondence ($r > 0.9$). Other months have correlations between 0.6-0.9, meaning that their spatial regime patterns are not too much different from the winter (DJF) ones. The only exceptions are the months of December and May, in case of the AR: both of them show correlations close to 0 in Figure 3, so they are not comparable with the winter regime patterns. In these two cases, their regime structures are also very different from the regime patterns of the neighbouring months (not shown).

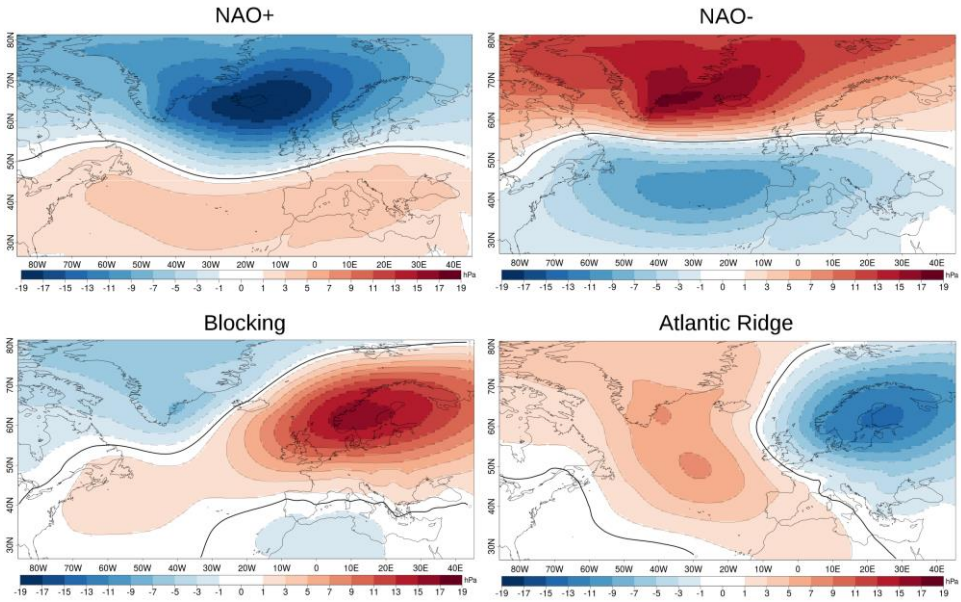

***Figure 2**. ERA-Interim winter (DJF) SLP regime anomalies.*

For each of the four WRs (NAO+, NAO-, BL and AR), the average of its SLP anomalies (in hPa) during winter (1979-2013) in the North Atlantic-European region reveals a spatial pattern which is specific of each regime.

For these reasons, December and May AR regimes are not properly defined with this classification. Thus, AR cannot be properly validated using ECMWF-S4 during December and May. Notwithstanding, in this work, results of the validation are presented for all winter months, including AR's December and May regimes, to be able to evaluate the errors and bias associated to these two ambiguous regimes, even though they are not directly comparable with the ECMWF-S4 simulated regimes.

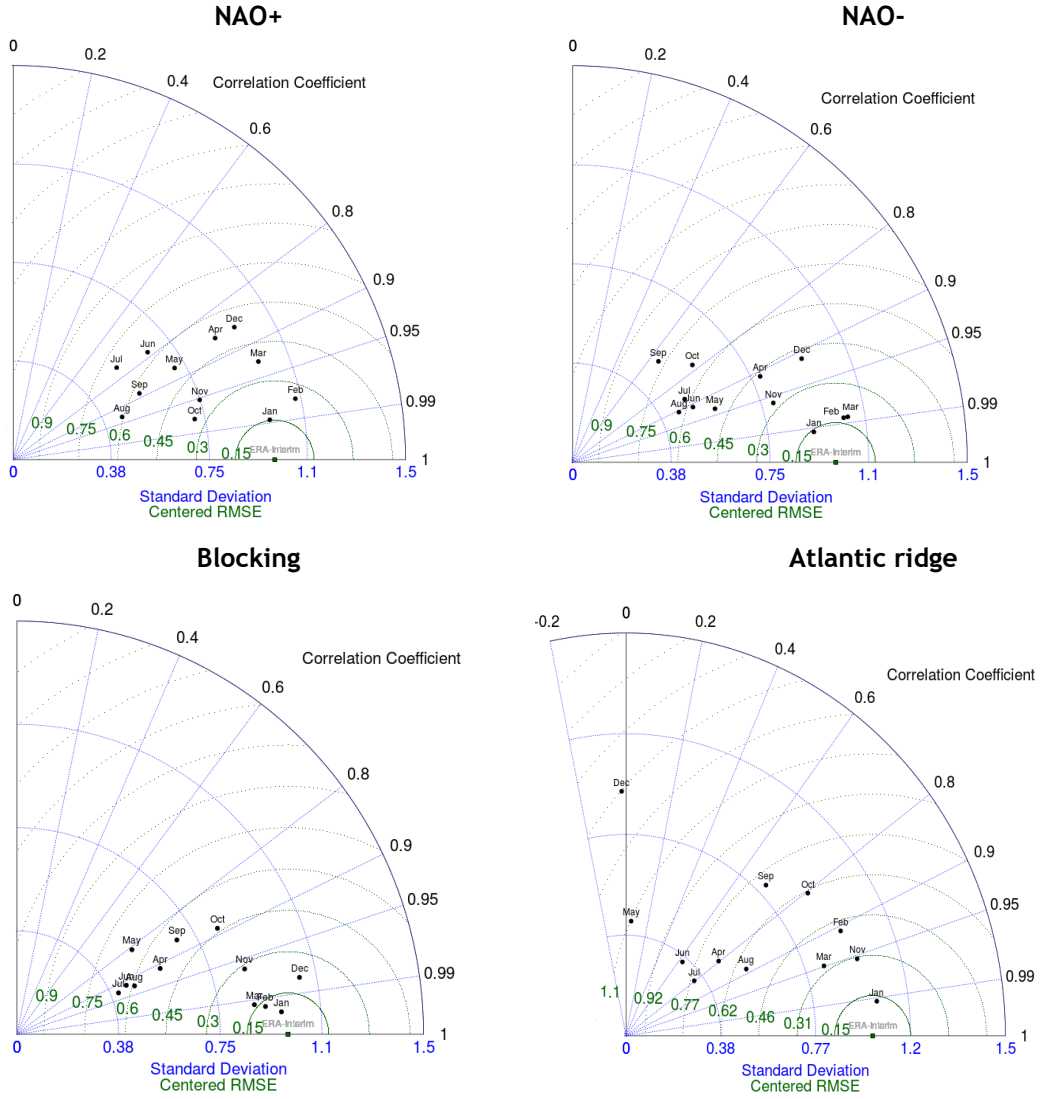

***Figure 3**. Taylor diagrams of the regime patterns.*

Taylor diagrams of the comparison between ERA-Interim monthly regime anomalies classified with the 3-months running clustering. Each point represents a different month, each diagram a different regime. Points closer to the black point on the bottom axis labelled "ERA-Interim" represents regime anomalies that more closely resemble the winter (DJF) ones, in term of spatial correlation, standard deviation ratio and Root Mean Square Error (RMSE).

3.2. Pattern correlations

For each month of the year, four Euro-Atlantic WRs predicted with ECMWF-S4 were obtained from the seven start dates previous to the target month and confronted with the corresponding ones obtained with ERA-Interim reanalysis. For example, if we choose January as the target month, we can assess the goodness in simulating the corresponding WRs along lead times, since the previous seven start dates (from July to January) were used (Figure 4). The WR patterns for the remaining months through the year were obtained and they are available at the Earth System Services-Barcelona Supercomputing Centre (ESS-BSC) catalogue[1]. From Figure 4 it is evident that the simulated WRs' patterns, for the target month of January, closely resemble the monthly patterns from reanalysis (last column to the right), for all lead times. Such a high degree of similarity translates to a high spatial correlation coefficient, as shown in Section 3.2.

To summarize the similarity of the simulated regime anomalies maps from all the possible combinations of start dates and lead times beyond the example above, spatial correlations between simulated and observed regime anomalies (also called "pattern correlations") were measured (Figure 5). For the majority of target months, correlation values in Figure 5 are above 0.6 (the threshold between the red and the blue colours) even for late lead times, except a few cases: September and October regimes, when low correlation values are found for all the WRs and lead times, AR regime in January for lead time 0 and AR's May and December regimes. Last two cases are due to the fact that the AR patterns found in ERA-Interim data in May and December are defined in an ambiguous way, as explained in the previous page. Other regimes show high pattern correlations, most of them above 0.9, demonstrating that when the ERA-Interim regimes are well defined, S4 regimes quite often resemble the observed ones.

[1] http://www.bsc.es/ESS/catalogue

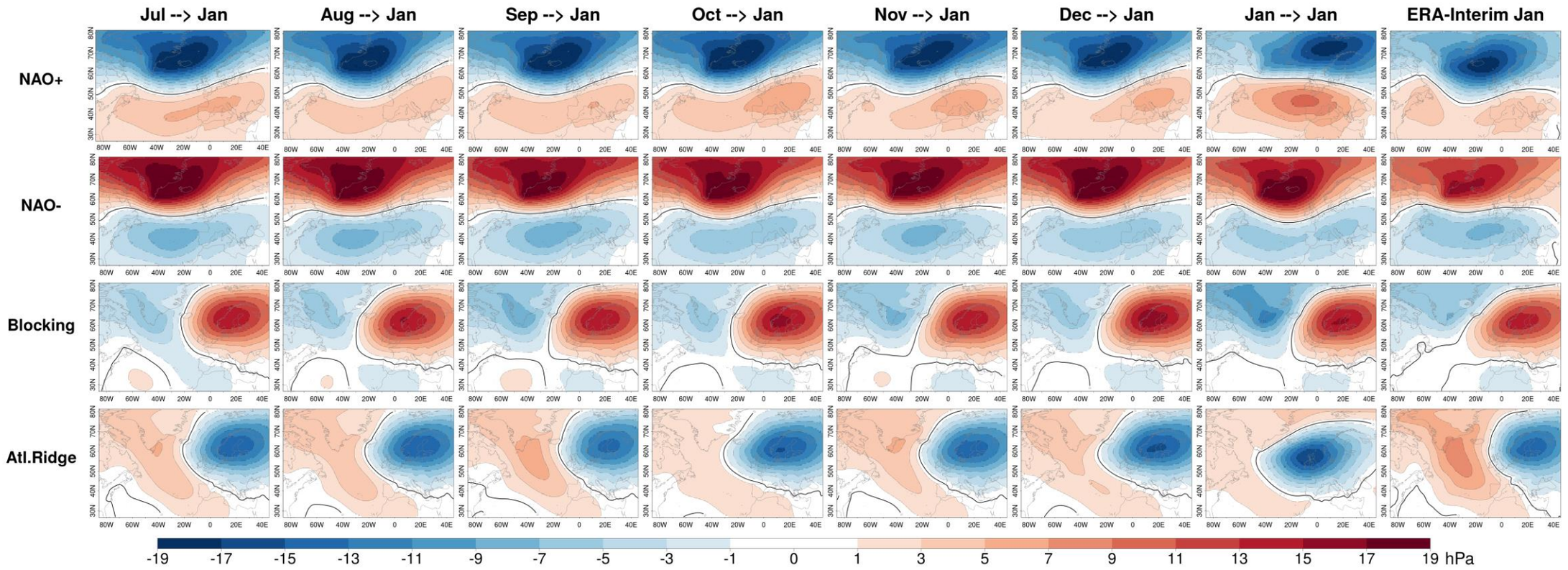

***Figure 4**. Simulated regime anomalies for January.*

ECMWF-S4 simulated regimes anomalies (in hPa) for the target month of January (1981-2015). From top to bottom: NAO- , NAO- , BL and AR regime anomalies. From left to right: Start date varies from July (left column) to January, so the lead time decreases from 6 (left column) to 0. The right column shows the ERA-Interim observed regimes anomalies for January, for comparison with the simulated ones. The bold black lines show the null anomalies.

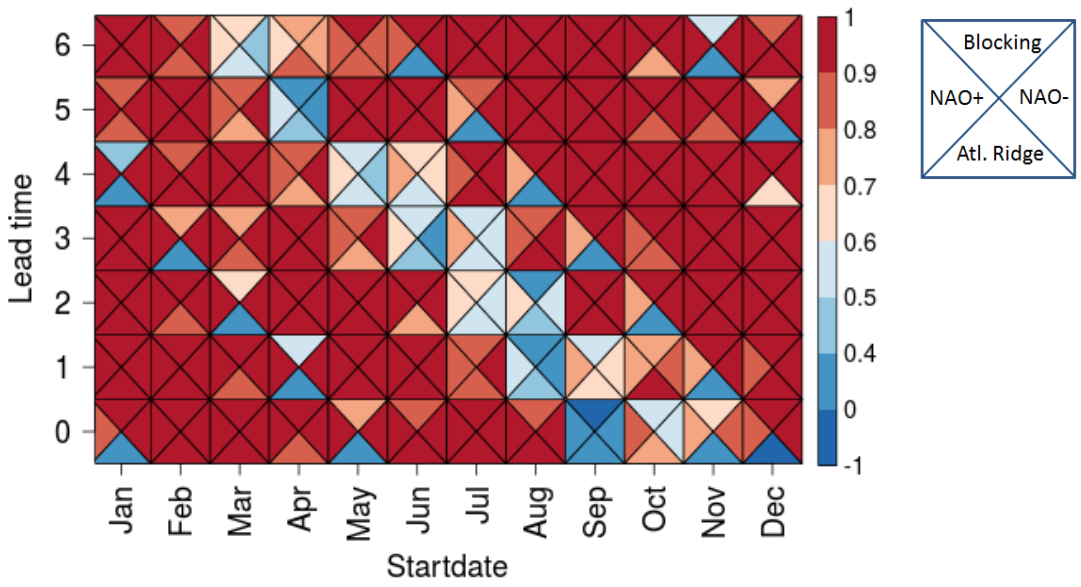

***Figure 5**. Pattern correlation*

Spatial correlation coefficients between the simulated and the observed monthly regime anomalies (1981-2015). Each triangle represents a different WR, as shown in the legend to the right. X-axis shows the start date and y-axis the lead time (in month); lead time 0 refers to a forecast for the same month of its start date. Each diagonal line represents all the forecasts for the same target month but with different lead times.

3.3. Frequency correlations

To evaluate the efficacy of S4 in reproducing the observed inter-annual monthly frequencies of WRs, the correlation coefficients between the simulated and observed monthly frequency of occurrence of each WR was obtained for seven different lead times and all individual months. As an illustration of the S4 and ERA-Interim frequency time series for a specific target month, we are showing them for the month of January having into account all possible lead times (Figure 7). Results obtained from the rest of target months across the year are shown in the ESS web catalogue.

From Figure 7, we can conclude that simulated frequencies with S4 (red and blue bars in columns from 1 to 7 on Figure 7) showed a reduced variability with respect to those from the ERA-Interim (red and blue bars on the right column of Figure 7). This was due to the effect of calculating the ensemble mean of the members, which caused a reduction of noise. However, the variability of the simulated frequencies of individual members was comparable to that of the ERA-Interim as shown by the grey bars, which represent the spread (maximum and minimum values) of the members across the ensemble for each year. There is also a good concordance between the simulated and observed average monthly frequencies over the period 1981-2015 (grey horizontal line on Figure 7). Similar results were found for the rest of the months across the year.

For the target month of January, correlation values between simulated and observed inter-annual frequencies are quite low for all lead times except for the zero lead time, which values are moderately high in the case of the NAO+ and NAO-, with r = 0.55 and 0.62, respectively. The similarity of the inter-annual monthly frequencies for all the months and lead times can be shown in Figure 6. In general, it temporal correlation values are above 0.5 only for lead time zero (last row in Figure 6), and quickly drop below 0.5 at higher lead times. Hence, there is no skill in predicting the monthly frequency of any WR beyond lead time zero. The two NAO regimes are the ones with usually have the highest correlations for lead time zero.

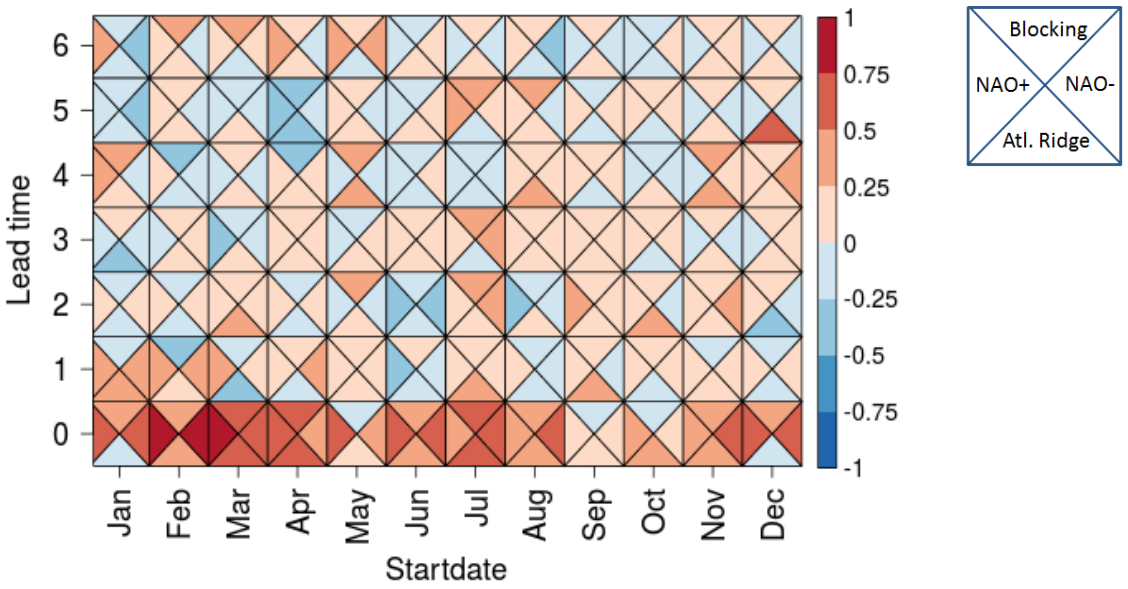

***Figure 6**. Frequency correlation.*

As in Figure 5, but for the correlation values between the simulated and observed monthly frequencies of occurrence of each regime (1981-2015).

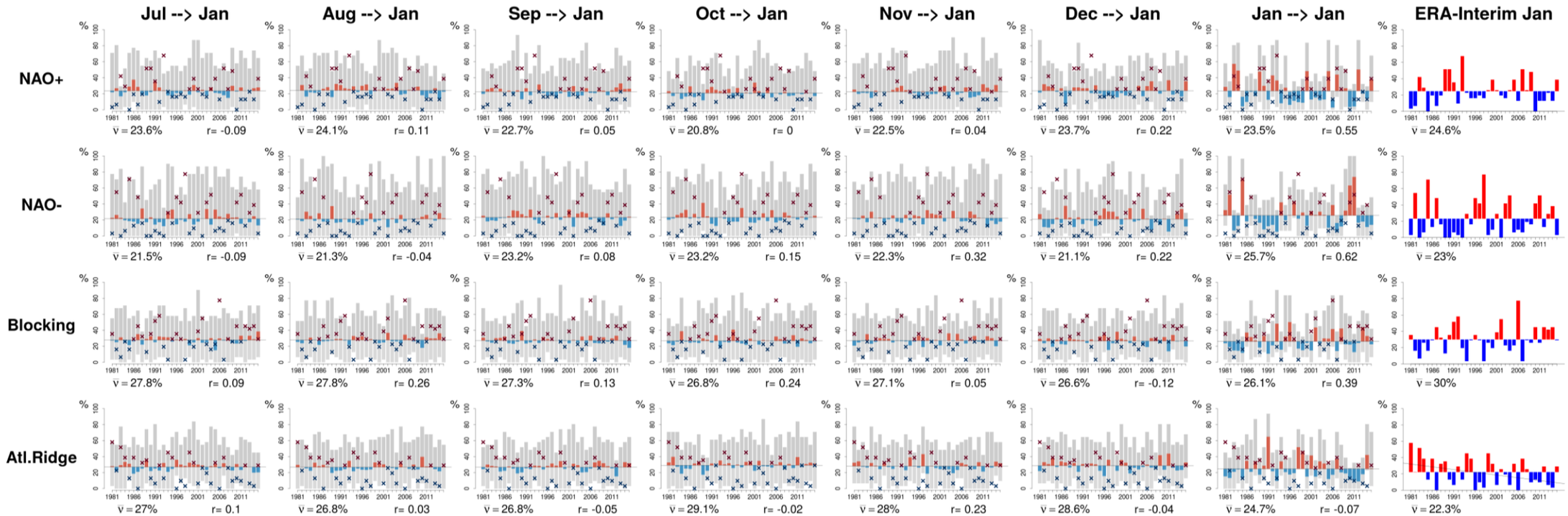

***Figure 7**. Simulated interannual frequencies of occurrence for January.*

S4 simulated time series (1981-2015) of the inter-annual regime frequencies (in %) for the target month of January and different lead times (from left to right, lead times varies from 6 to 0 months) versus ERA-Interim observed frequency series (last column to the right). Red and blue bars indicate the monthly frequency (in case of S4, of the 15-members ensemble mean) compared to the average frequency 1981-2015 (grey line). Grey bars show the maximum and minimum monthly frequency across the individual 15 members for each year; red and blue crosses show the observed frequency (also shown by the red/blue bars in the last column to the right). Bottom numbers show the average frequency in % and correlation value (r) between the simulated and observed inter-annual frequencies.

3.4. Average frequency bias

Even if S4 doesn't adequately reproduce the inter-annual monthly frequencies of the WRs (see previous Section), it might be still able to reproduce the average (1981-2015) monthly frequencies correctly. If S4 rarely predicts the interannual monthly frequencies of a certain WR correctly, yet it simulates their observed average frequency quite well, it is likely that the problem lies in the inherent unpredictability of the frequency of occurrence of the WR (Ferranti et al. 2015). On the contrary, if S4 average frequency of a WR is significantly different from the observed one, such a low skill it is likely due to the misrepresentation of some physical process in the model, and not to a model fault.

The average simulated monthly frequencies of each WR during 1981-2015 vary between 18% and 34% of the total days of the month (not shown). BL regime has the highest average simulated frequency for most of the start dates and lead time combinations, being higher than 26% in most of the cases. It is followed by the AR regime, while the two NAO phase's regimes have the lowest average simulated frequencies, usually below 26%.

For each WR, the difference between its simulated and observed average monthly frequency for 1981-2015 was plotted in Figure 8. Almost all possible combinations of start dates and lead times have a bias limited to ±10%, meaning that S4 is quite able of reproducing the average monthly frequency of the WRs. Positive bias values (red triangles) are mostly found for BL and AR regimes, while negative ones (blue triangles) are more related to the two NAO regimes. Even if the average frequency bias is not large, we can say that S4 forecasts overestimate the frequency of the BL and AR regimes except for winter months (December to February) in the case of BL and spring months (March to May) and June for AR where there is an underestimation.

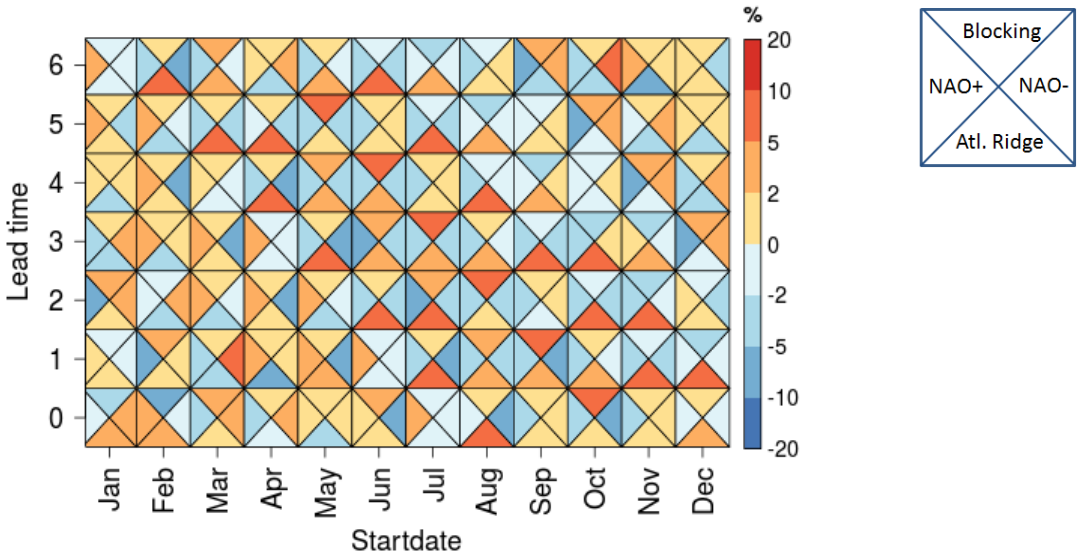

***Figure 8**. Bias of the average frequency*

Difference between S4 simulated and observed average (1981-2015) monthly frequency (in %).

For NAO+ and NAO- regimes, there is an underestimation in all season except for early summer (May to July) in the case of the NAO+ and the spring months (March to May) in the case of NAO-, when an overestimation is evident. This is consistent with Ferranti et al. (2015), who also found a similar behaviour for the AR regime simulated with the ECMWF medium-range forecast system. The frequency bias doesn't show any clear seasonality or any increase with the lead time.

3.5. Persistence bias

Atmosphere tends to persist in the same WR for a few days before evolving to a different one (Fil & Dubus 2005). Persistence indicates the mean number of consecutive days before a WR is replaced by a new one; it is typically equal to 3-5 days for Euro-Atlantic WRs. The NAO- regime shows the highest observed persistence (close to 5 days), except in autumn (see image in the catalogue)); the other three WRs are shorter than NAO-, particularly BL and AR, which last less than 4 days. The simulated persistence can vary a little with the increase of the lead time but, as for the average frequencies, it does not show any strong monotonic increase or decrease that would indicate the presence of a systematic error in the forecast model.

The persistence bias is defined as the difference between the simulated and the observed persistence (in days) and it is shown in Figure 9. S4 shows high skill in reproducing the observed persistence, since most of the persistence bias values are limited within ±1 days/month. Negative bias values (blue triangles) are more numerous than positive bias values (red triangle), meaning that S4 forecasts slightly underestimate observations for the two NAO regimes (more negative bias); such bias is similar to that detected for the frequency bias (Figure 8). BL and AR bias is even smaller, and as it is limited to ±0.5 days/month, except for AR during December and May, the two months without a clear observed regime structure (see Section 3.1). Finally, the bias doesn't show any clear seasonality or any dependence on the lead time.

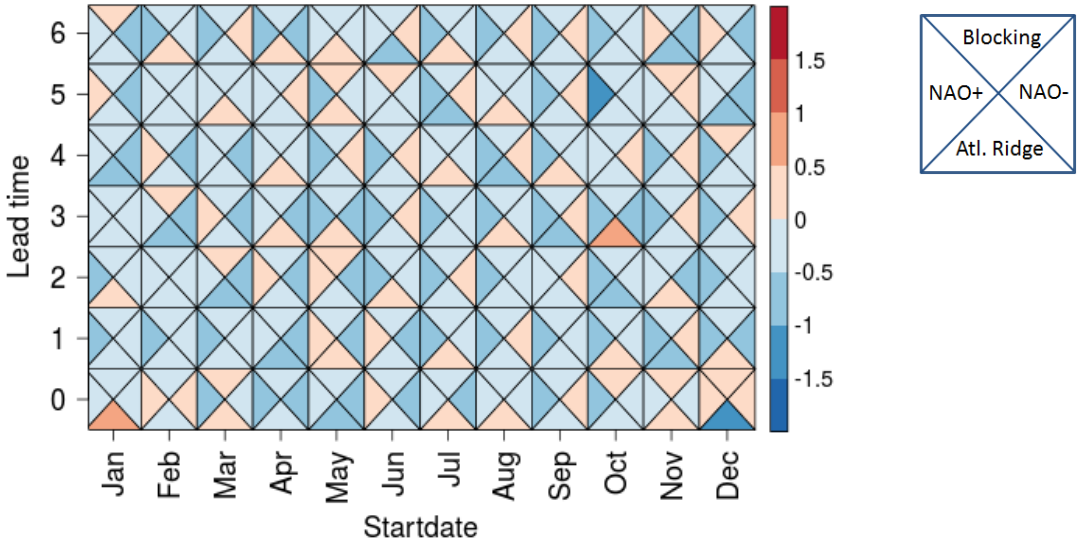

***Figure 9**. Persistence bias.*

Difference between S4 simulated and observed monthly persistence (in days/month).

3.6. Transition probability bias

Transition probabilities measure the frequency by witch a WR is replaced by another one. Thus, they allow identifying the preferred transitions between WRs. Observed transition probabilities displayed in the ESS-BSC web catalogue[2], show a low probability for NAO+ regime to be followed by a NAO- regime and vice versa. NAO+ regime shows the highest transition probability to the AR regime and NAO- to the BL regime. In the case of BL and AR regimes, there is no a transition on preferred WR, except from August to January where the BL tend to have a slightly high transition probabilities to the NAO+ regime.

Regarding the simulated transition probabilities, the S4 is quite good agreement with observations in terms of matching red and blue colours in similar target months and lead times; nevertheless, there are differences regarding the intensity. To evaluate this bias, the difference between the simulated and observed transition probability (in %) is shown in Figure 10. For example, S4 transition probabilities from NAO+ to NAO- (and vice versa) are only of a few points percent, while the corresponding observed probabilities are higher. Therefore S4 underestimates the transition probability for all target months, being up to -20% in months from August to December.

Overall, the simulated preferred transition of NAO+ regime is the AR regime while for NAO- regime it is the BL regime, as shown in the ESS-BSC web catalogue. However, S4 tends to overestimate, in general terms, the transition probability from NAO- (NAO+) to BL (AR) regime. In the former, the maximum value of the overestimation is found during September, October, March and April, reaching a value of 20%; in the latter, the maximum values are up to 20%, except from April to July where transition probabilities are underestimated.

[2] http://www.bsc.es/ESS/catalogue

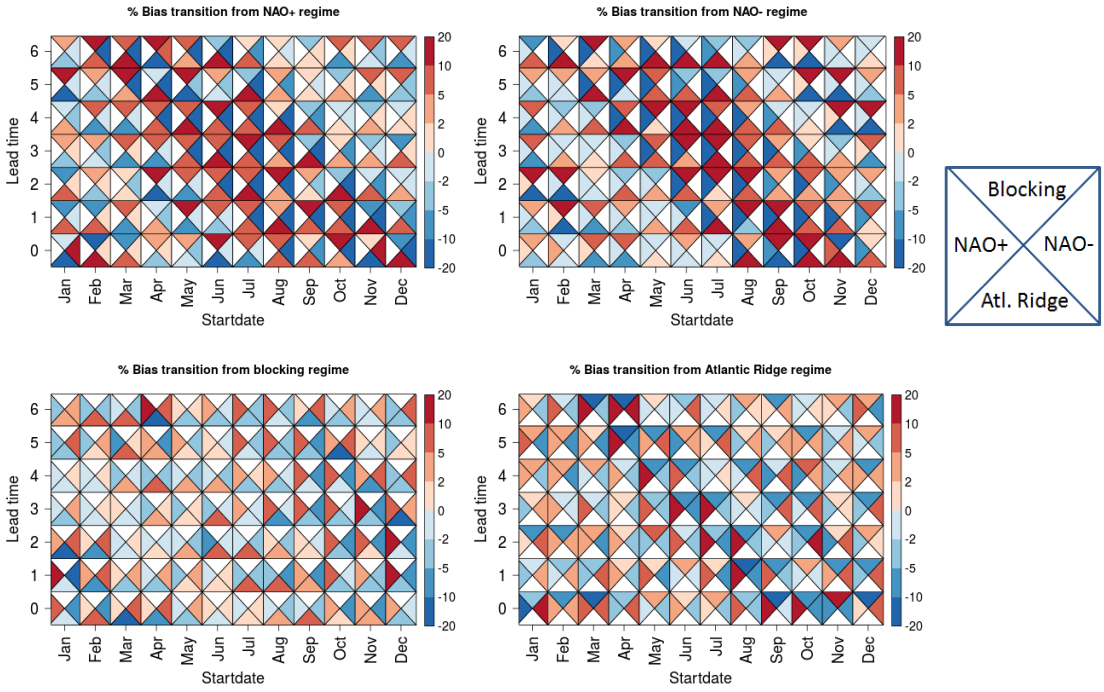

***Figure 10**. Transition probability bias.*

As in Figure 5, but for the bias of the transition probability (difference between the simulated and observed transition probability in %). White triangles correspond to transitions from a WR to the same one.

The simulated preferred transition of BL regime is the NAO+, in agreement with observations, but there is an overestimation of the transition probability from September to February. In the case of AR regime there is no simulated preferred transition, but the highest bias is found in the transition to the NAO+ with values around 20% are exhibited for September and October target months.

3.7. Impact of WRs on wind speed

Monthly simulated and observed impact maps of the four WRs on 10-m wind speed where compared for different target months and all possible lead times. In general, the impacts are ranged within a range from -3m/s to 3 m/s (or ±10 km/h). In this document only the results obtained for the target month of January are displayed (Figure 11). The main difference between the observed and simulated spatial correlations for January is related to the AR regime during lead time 0. In fact, the area affected by a positive impact of AR on wind speed in January doesn't correspond to the observed area (Figure 11, bottom right). Such a difference is mostly due to the low pattern correlation between S4 simulated and observed regime anomalies for AR in January (see Figure 5), anomalies which are also employed to measure the impact (along with the wind speed anomalies). The other three regimes show little or no differences, as the patterns of the simulated impact closely resemble the observed patterns (right column in Figure 11), even for high lead months (columns to the left), without showing any decrease of the intensity of the impact with the lead time. Such a

high degree of similarity translates into a high spatial correlation coefficient, which in January is always over 0.7 for these three regimes (Figure 12). Monthly impact maps for the rest of the months across the year, together with the impact maps on 2-m temperature can be found at the ESS-BSS web catalogue[3].

The spatial correlation coefficients between simulated and observed impact maps on 10-m wind speeds for all possible combinations of lead time and start data are shown in Figure 12. Most of the impact correlations are above 0.6 (the threshold between the red and the blue colours), even for lead times of 6 months. Values below this threshold are usually associated to the same target months: all WRs have a quite low correlation value for September, for October in the case of NAO- and BL, and for May and June for BL (diagonal lines with blue triangles in Figure 12). Regarding the AR, target months of May and December also shown a very low value, but this can be due to in these months the observed WR doesn't resemble to the spatial structure of a typical DJF AR regime. Highest correlations are measured for the target months of January, February and March, most specially for both NAO+ and NAO-.

Globally, correlation values are lower than those shown in Figure 5 of regime anomalies, because S4 accuracy in reproducing the observed impact depends on both its accuracy in simulating the observed SLP fields and in simulating the observed wind speed fields; thus, impact correlations take into account the forecast errors of two variables instead of one, increasing the total error. Similar findings were measured for the WRs impact on 2-m temperature, whose impact maps are also available in the ESS-BSC web catalogue.

[3] http://www.bsc.es/ESS/catalogue

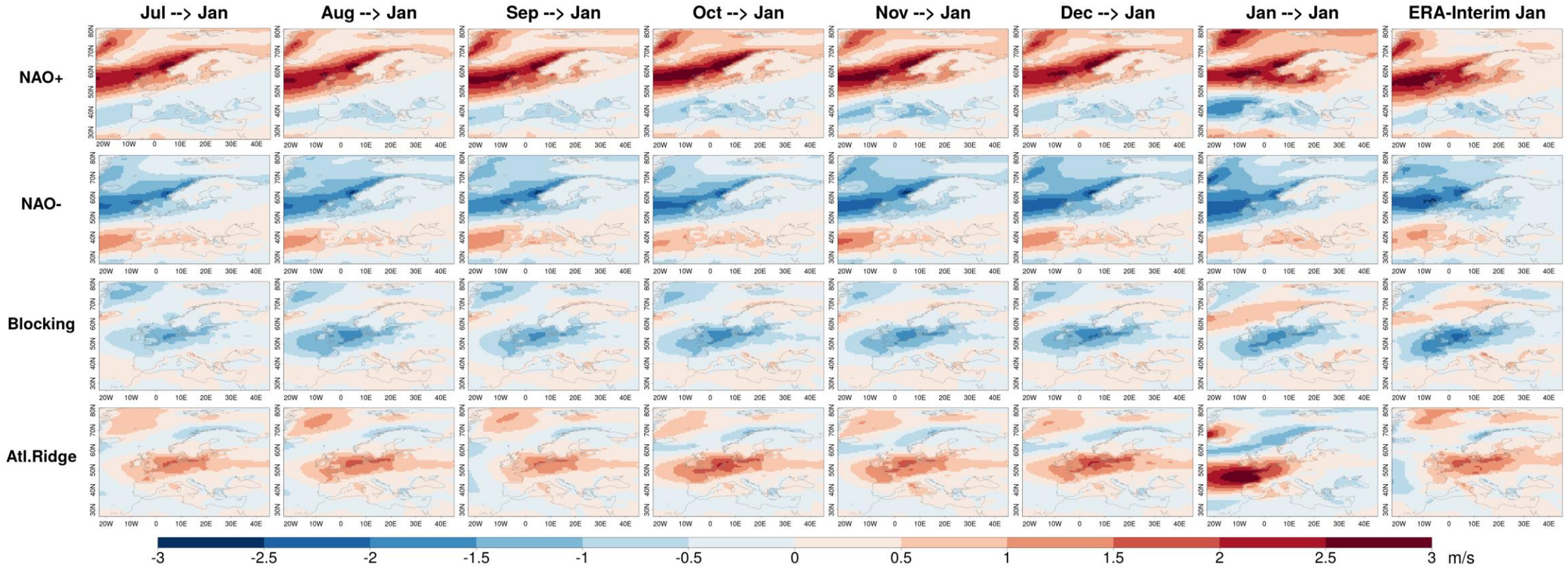

***Figure 11**. WRs impact on 10-m wind speed for target month of January.*

Simulated versus observed impact of the WRs on 10-m wind speed (in m/s) in Europe over the period 1981-2015, for target month of January. From left to right, the start dates vary from July to January (lead times decreases from 6 to 0); being the last column on the right the ERA-Interim impact. Positive values show areas with positive impact (higher wind speed than the average), while negative values show areas with negative impact.

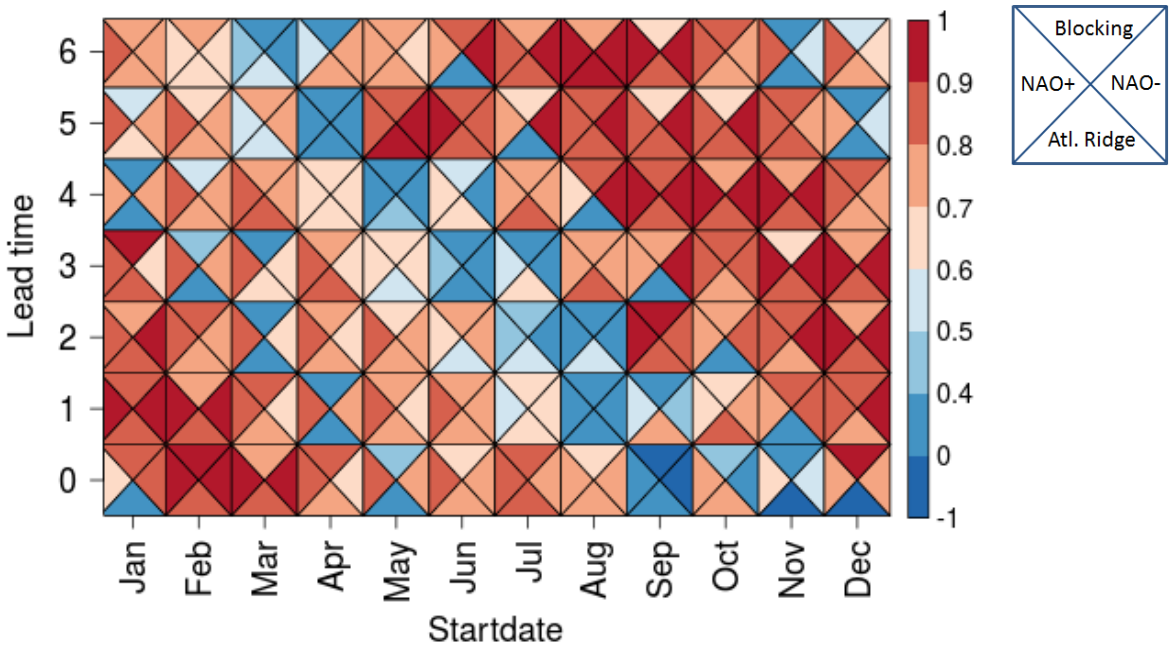

***Figure 12**. Wind Speed's impact correlation.*

As in Figure 5, but for the correlation values between simulated and observed impact of the WRs on 10-m wind speed.

4. Conclusions

The main aim of this work was to compare two monthly WR classifications, one derived from simulated SLP data of the seasonal forecast system ECMWF S4, and another proceeding from ERA-Interim reference reanalysis. In this way, the skill of S4 in simulating the observed main characteristics of the atmospheric circulation (summarized by the WRs) was validated. To this end, the four Euro-Atlantic WRs were classified by means of a 3-months running k-means cluster analysis applied to daily SLP anomalies of ERA-Interim reanalysis and S4 separately. The assessment between simulated and observed WRs was performed with several verification metrics: pattern correlation, frequencies correlation, average frequency bias, persistence bias, transition probabilities bias and 10-metre wind speed and 2–metre temperature impact pattern correlation, for each start date and lead time individually.

One of the main challenges this work faced was to develop a methodology to obtain the four monthly Euro-Atlantic WRs throughout the year. In fact, for some target months and WRs, (notably AR in December and May), the traditional k-means clustering technique cannot adequately represent the observed winter (DJF) regime structures, as also evidenced by different authors (Fereday et al. 2008, Neal et al. 2016). This is the reason why in the bibliography most of the studies are focused on winter or summer seasons separately (Vrac et al. 2014). We demonstrated that following a 3-month running clustering approach, centred in the middle month, the resulting WRs have a structure more similar to the winter (DJF) structure than the ones classified with a traditional single-month clustering. However, even with this method, there are still two ambiguous observed regimes, the AR in December and in May. Therefore, in these two months of AR, it is not possible to assess the performance of the WRs with S4 correctly. For the rest of the regimes and months, some general conclusions can be drawn about how the S4 reproduces the observed WRs.

It is worth mentioning that all verification metrics exhibit similar errors and bias for all lead times referred to the same target month, instead of increasing the errors with the lead time. Overall, the spatial patterns of the simulated regimes are remarkably well simulated by S4 ($r>0.9$) for the majority of start dates, lead times and regimes, except during September, October and November, probably due to the high SLP variability during autumn. Notably, spring and summer season perform almost as well as the winter season. Moreover, S4 skilfully reproduces the average monthly frequencies of occurrence of each WR over period 1981-2015 for the ensemble mean, even for high lead times (up to six months in advance). However, S4 doesn't adequately reproduce the monthly frequency time series at lead times greater than zero. Such low skill might be attributed to the intrinsic unpredictability of the regimes, and not to a model fault (Ferranti et al. 2015). This lack of skill prohibits using S4 to provide reliable monthly forecasts of the frequency of occurrence of the WRs beyond the first month. Furthermore, S4 forecasts often tend to underestimate the monthly frequency of occurrence and persistence of the NAO+ and NAO- regimes, and to overestimate the monthly frequency of BL and AR regimes. Finally, S4 greatly underestimates the transition probabilities of NAO+ regime to NAO- regime and vice versa, and overestimates the transition probability from NAO+/NAO- regimes to BL or AR.

5. Appendix

The validation process assumes that the observed monthly weather regimes are always representative of the cluster space; however, such a condition is not always satisfied, particularly during spring and autumn months (being SLP data highly variable during these transition months), and/or when are few years of daily data available (Fereday et al. 2008). In fact, European WRs are usually classified during winter season (DJF), during the extended winter season (NDJFM), or during summer season (JJA). Outside these seasons, the cluster analysis is less performant. Thus, to provide robustness to the cluster analysis, several decades of reanalysis data should be introduced; however, in this work, only 35 years could be employed (1981-2015), the common period between ERA-Interim and S4 data. For this reason, some observed monthly WRs are not always representative of the cluster space, as can be seen in the red points of the Taylor diagrams in Figure 13 with low correlations.

Red points in Figure 13, in fact, represents the similarity between monthly regime anomalies patterns (obtained with k-mean clustering) and those defined for winter (DJF) season. The 3-months running cluster was introduced to increase the regime's representativeness, by adding more SLP data to the k-means clustering. In this way, the monthly WRs pattern anomalies more closely resemble the winter spatial structure of the four WRs, as shown by the blue points in Figure 13, which are more closely concentrated near the black point (corresponding to DJF patterns) than the red points.

To measure quantitatively the performance of each type of approach (clustering or running clustering), the average of the correlation coefficients and of the root mean square error (RMSE) for the 12 months was measured independently for each of the two methodologies and shown in Table 1. It is evident that the 3-months running clustering generates regimes with spatial patterns more similar to the winter DJF patterns, both in term of correlations and of RMSE. For example, the negative correlation measured for BL in November becomes a highly positive correlation (>0.95) when classifying regimes with the running clustering.

AR is the regime whose monthly patterns are most different from the winter (DJF) ones, as can be seen by the large spread of the points in Figure 13. Without the running clustering, two months present negative correlations with the winter regimes: January and October. With the running cluster, only one month present negative correlations (December), while another one has a slightly positive one (May). Overall, AR correlations increase on average of +0.16 when adopting the running clustering approach, and the RMSE increases (improves) on average of -0.68. Consequently, the S4 assessment of the simulated WRs has to be done with great caution for the two December and May AR ambiguous regimes, since it is not possible to compare them with the corresponding two AR regimes simulated by S4. A possible solution is to classify the observed regimes employing a probabilistic reanalysis with many ensemble members (for example, with the upcoming ERA5 reanalysis), in a way similar to S4 WR classification, which demonstrated to have enough SLP data to define each regime unambiguously.

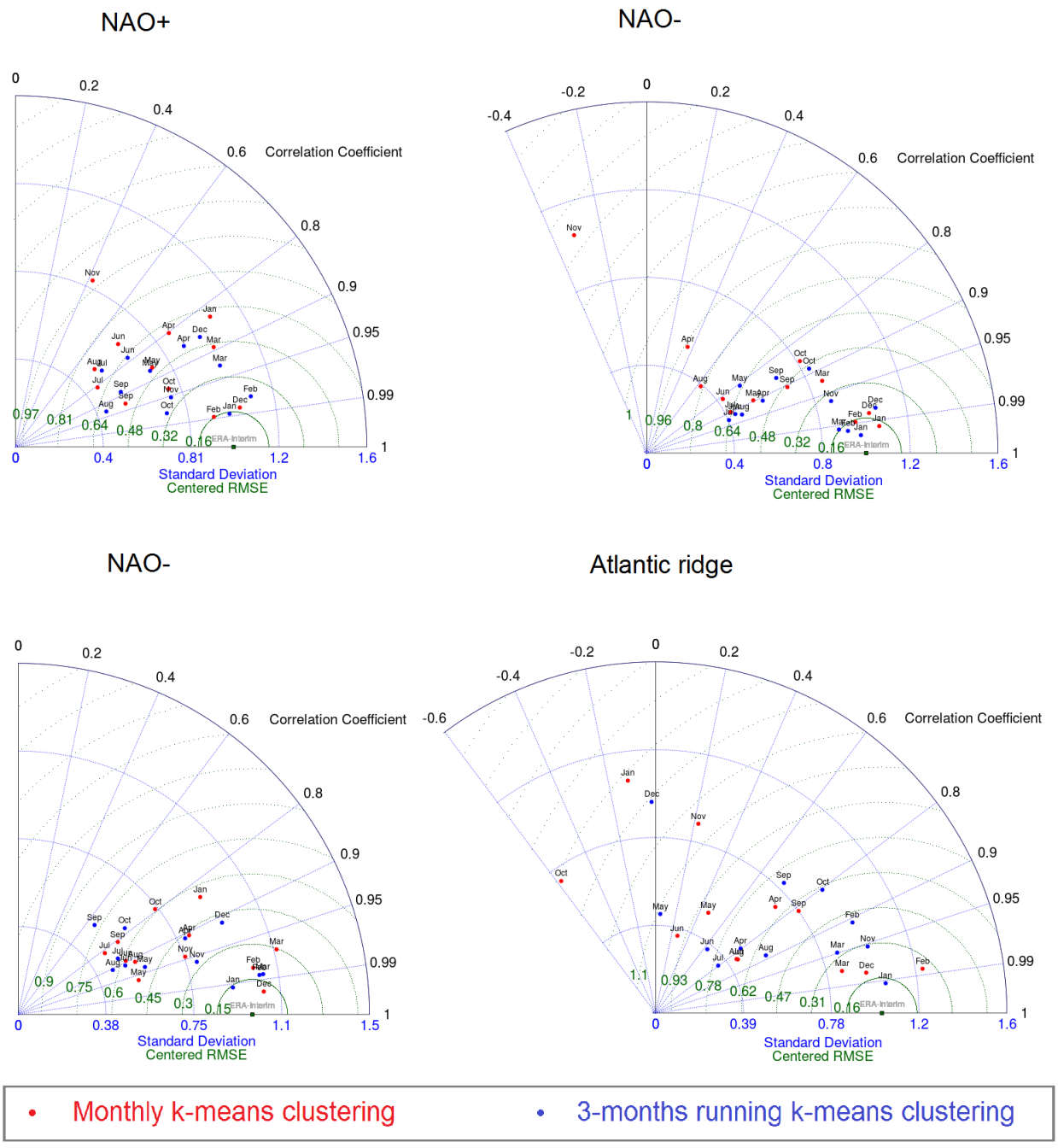

***Figure 13**. Taylor diagrams of the monthly regime patterns for the two clustering methodologies.*

Taylor diagrams of the comparison between ERA-Interim monthly regime anomalies computed with or without the 3-months running cluster. Each point represents a different month, and each diagram a different regime. Blue (red) points represent the WRs computed with (without) the running cluster (blue points correspond to black points in Figure 3). Point closer to the black point on the bottom axis labelled "ERA-Interim" represents regime anomalies that more closely resemble the winter (DJF) ones, in term of spatial correlation, standard deviation ratio and RMSE.

Note that even if the two problematic clusters were included in the validation of S4, the terminology adopted in the text for the winter regimes (NAO+, NAO-, BL and AR) is employed even during these two months when the AR's spatial structure doesn't correspond to that of the AR winter (DJF) structure.

	NAO+	NAO-	BL	AR
r *(running clustering)*	0.90	0.90	0.93	0.70
r *(clustering)*	0.82	0.90	0.74	0.54
Δr	***+0.08***	***0.00***	***+0.19***	***+0.16***
RMSE *(running clustering)*	3.15	3.42	2.33	3.41
RMSE *(clustering)*	3.73	3.46	3.26	4.10
ΔRMSE	***-0.58***	***-0.04***	***-0.93***	***-0.68***

Table 1*. Comparison between the two clustering methodologies.*

Correlation (r) or RMSE values shown in the Taylor diagram in Figure 13 averaged over all months, for each regime and methodology separately (clustering or running clustering). Both Δr and ΔRMSE show a clear improvement of the running clustering approach compared to the simple clustering, except for the NAO- regime, which only improves slightly.

6. Acknowledgements

The authors acknowledge funding support from the RESILIENCE (CGL2013-41055-R) project, funded by the Spanish Ministerio de Economía y Competitividad (MINECO). We are also grateful to Dr. Christian Viel of Météo France for the helpful discussions regarding this work.

7. References

Cassou, C., Terray, L. & Phillips, A., 2006. Weather Regimes and European heat waves 2003 a case study. *JPL OSE Meeting*, pp.35–42.

Corti, S., Molteni, F. & Palmer, T.N., 1999. Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature*, 398 (April), pp.799–802.

Dawson, A., Palmer, T.N. & Corti, S., 2012. Simulating regime structures in weather and climate prediction models. *Int.J.Clim.*, 39 (November), pp.1–6.

Dee, D.P. et al., 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656), pp.553–597.

Doblas-Reyes, F. et al., 2009. Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts. *Int.J.Clim*, 1559(July), pp.1538–1559.

Fereday, D.. et al., 2008. Cluster Analysis of North Atlantic – European Circulation Types and Links with Tropical Pacific Sea Surface Temperatures. *American Meteorological Society*, 21(1), pp.3687–3703.

Ferranti, L. & Corti, S., 2011. New Clustering Products. *ECMWF Newsletter*, 127(September), pp.6–11.

Ferranti, L., Corti, S. & Janousek, M., 2015. Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, 141(688), pp.916–924.

Fil, C. & Dubus, L., 2005. Winter climate regimes over the North Atlantic and European region in ERA40 reanalysis and DEMETER seasonal hindcasts. *Tellus, Series A: Dynamic Meteorology and Oceanography*, 57(3), pp.290–307.

Hafez, Y.Y. & Almazroui, M., 2014. Recent Study of Anomaly of Global Annual Geopotential Height and Global Warming. *Atmospheric and Climate Sciences*, 4, 347–357(July), pp.347–357.

Hartingan, J.. & Wong, M.., 1979. Algorithm AS 136 A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, 28(1), pp.100–108.

Hurrell, J.W. & Deser, C., 2009. North Atlantic climate variability: The role of the North Atlantic Oscillation. *Journal of Marine Systems*, 78(1), pp.28–41.

Mahlstein, I. et al., 2015. Estimating daily climatologies for climate indices derived from climate model data and observations. *Journal of Geophysical Research Atmospheres*, pp.2808–2818.

Michelangeli, P.-A., Vautard, R. & Legras, B., 1995. Weather Regimes: Recurrence and Quasi Stationarity. *Journal of the Atmospheric Sciences*, 52(8), pp.1237–1256.

Molteni, F. et al., 2011. The new ECMWF seasonal forecast system (System 4). *European Centre for Medium-Range Weather Forecasts, internal report*.

Neal, R. et al., 2016. A flexible approach to defining weather regimes and their application in weather forecasting over Europe. *Meteorological Applications*, 1, pp.1–12.

Pelly, J. & Hoskins, B., 2003. A New Perspective on Blocking. *American Meteorological Society*, (1983), pp.743–755.

Tibaldi, S. & Molteni, F., 1989. On the operational predictability of blocking.pdf. *Tellus*, 42a, pp.343–365.

Vrac, M., Vaittinada Ayar, P. & Yiou, P., 2014. Trends and variability of seasonal weather regimes. *International Journal of Climatology*, 34(2), pp.472–480.

Yiou, P. & Nogaj, M., 2004. Extreme climatic events and weather regimes over the North Atlantic: When and where? *Geophysical Research Letters*, 31(7), pp.1–4.