Revisiting the identification of wintertime atmospheric circulation regimes in the Euro-Atlantic sector

Article

Published Version

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1002/qj.3818

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

RESEARCH ARTICLE

Revisiting the identification of wintertime atmospheric circulation regimes in the Euro-Atlantic sector

Swinda K.J. Falkena[1] [ID] | Jana de Wiljes[1,2] | Antje Weisheimer[3,4] | Theodore G. Shepherd[5]

[1]Department of Mathematics and Statistics, University of Reading, Reading, UK

[2]Institute for Mathematics, University of Potsdam, Potsdam, Germany

[3]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK

[4]National Centre for Atmospheric Science (NCAS), Department of Atmospheric, Oceanic and Planetary Physics (AOPP), University of Oxford, Oxford, UK

[5]Department of Meteorology, University of Reading, Reading, UK

Correspondence
S.K.J. Falkena, Department of Mathematics and Statistics, University of Reading, Whiteknights, PO Box 220, Reading RG6 6AX, UK
Email: s.k.j.falkena@pgr.reading.ac.uk

Abstract

Atmospheric circulation is often clustered in so-called circulation regimes, which are persistent and recurrent patterns. For the Euro-Atlantic sector in winter, most studies identify four regimes: the Atlantic Ridge, Scandinavian Blocking and the two phases of the North Atlantic Oscillation. These results are obtained by applying k-means clustering to the first several empirical orthogonal functions (EOFs) of geopotential height data. Studying the observed circulation in reanalysis data, it is found that when the full field data are used for the k-means cluster analysis instead of the EOFs, the optimal number of clusters is no longer four but six. The two extra regimes that are found are the opposites of the Atlantic Ridge and Scandinavian Blocking, meaning they have a low-pressure area roughly where the original regimes have a high-pressure area. This introduces an appealing symmetry in the clustering result. Incorporating a weak persistence constraint in the clustering procedure is found to lead to a longer duration of regimes, extending beyond the synoptic time-scale, without changing their occurrence rates. This is in contrast to the commonly used application of a time-filter to the data before the clustering is executed, which, while increasing the persistence, changes the occurrence rates of the regimes. We conclude that applying a persistence constraint within the clustering procedure is a better way of stabilizing the clustering results than low-pass filtering the data.

KEYWORDS

atmospheric circulation regimes, information criteria, k-means clustering, persistence

1 | INTRODUCTION

The study of atmospheric circulation, or weather, regimes has a long history. Starting from the 1940s, when the German weather service developed a set of weather types classifying the daily synoptic circulation (Deutscher Wetterdienst, 2019), the concept of weather regimes as an expression of the low-frequency variability of the atmospheric circulation has been a topic of research. The underlying concept is that the weather itself is a stochastic

process, whose statistics are strongly conditioned on the weather regime. From around 1990 onwards, different clustering methods have been used to identify these persistent and recurrent circulation patterns (e.g. Mo and Ghil, 1987; Molteni *et al.*, 1990; Vautard, 1990), primarily focussing on the wintertime Northern Hemisphere. Later specific sectors of the Northern Hemisphere, primarily the Euro-Atlantic sector (e.g. Michelangeli *et al.*, 1995; Kageyama *et al.*, 1999) and the Pacific-North American sector (e.g. Straus *et al.*, 2007; Riddle *et al.*, 2013; Amini and Straus, 2018), as well as the Southern Hemisphere (e.g. Mo, 2000) have been studied, along with the relation of circulation regimes with, for example, climate change (Corti *et al.*, 1999) and regional weather (Cassou *et al.*, 2005). Also more limited areas have been considered (e.g., Robertson and Ghil, 1999).

Initially different clustering methods, such as hierarchical clustering (e.g. Cheng and Wallace, 1993), using the analysis of the probability density function (e.g. Kimoto and Ghil, 1993), or k-means clustering (e.g. Michelangeli *et al.*, 1995), were used to identify the atmospheric circulation regimes. The last method, k-means clustering, has subsequently become the most-used approach for identifying the regimes in atmospheric data (Hannachi *et al.*, 2017). The number of clusters k has to be set *apriori*, making finding the optimal number of regimes part of the problem. Commonly used methods to do this are the verification of significance by using synthetic datasets (e.g. Dawson *et al.*, 2012) or looking at the reproducibility and consistency when the algorithm is run multiple times (e.g. Michelangeli *et al.*, 1995). Nearly always the data are first projected onto the first several Empirical Orthogonal Functions (EOFs), after which the clustering algorithm is applied to the time series of these EOFs (e.g. Vautard, 1990; Ferranti *et al.*, 2015). In addition, some studies apply a low-pass time filter to the data to remove the high-frequency, noisy oscillations and focus on the low-frequency behaviour (e.g. Straus *et al.*, 2007; Grams *et al.*, 2017). This method enforces a higher persistence of the regimes compared to standard k-means clustering which is independent of the time-ordering of the data.

Since clustering methods represent a projection of the data to a lower-dimensional state space, applying clustering to the already filtered data of EOFs means a projection of the data is done twice. Thirty years ago, this approach was necessary because computational limitations did not allow use of the full field dataset. However this is no longer a constraint. Nevertheless most studies continue to follow the original approach and use EOFs. As EOFs give the modes associated with the most variability, while clusters give the recurrent patterns, the means of dimension reduction is quite different.

The question thus arises of what the effect of this double filtering is on the resulting atmospheric circulation regimes. Similarly, applying a low-pass filter to remove the high-frequency behaviour before the cluster analysis also means the data are filtered twice. This is likely not only to affect the persistence, but also the occurrence of the found regimes and possibly the clusters themselves, thus also raising the question of how strong this effect is.

In this paper we compare the results of k-means clustering using EOF data with the results found for the full field data, for the case of the wintertime Euro-Atlantic sector. We pay special attention to the optimal number of clusters k, as it has a large influence on the regimes that are found. Furthermore, we study the effect of time-filtering on the found regimes by comparing it with an adapted k-means clustering algorithm that incorporates a constraint on the regime duration to enforce persistence of the regimes. This novel algorithm, which is discussed in Section 2.1, does not change the data themselves, but only the method to identify the regimes. Both comparisons give insight into the effect of filtering the data, either using EOFs or applying a low-pass filter, before applying the k-means clustering algorithm. We start by explaining the methods used, followed by a discussion of the differences and similarities of the results. In the end the findings are summarized and discussed.

2 | METHODS

We use 500 hPa geopotential height data from ERA-Interim (Dee *et al.*, 2011) on a 2.5° by 2.5° longitude–latitude grid for a domain covering the Euro-Atlantic sector, 20° to 80°N and 90°W to 30°E. Daily data (00:00 UTC) are considered for the months December to March using 39 years of data (1979–2018). Deviations from a fixed background state are used throughout this period. The main argument for considering a fixed background state instead of a seasonally varying one is that, when applying cluster analysis, the data used are preferably as complete as possible to avoid any type of bias. This means that few to no assumptions, such as a seasonal cycle, are made in preparing the data to retain the information present in the data. Or, phrased differently, how can one compare two days if they are deviations with respect to a different background state? The risk of this approach is that seasonality affects the regimes that are found and introduces a bias in the occurrence and persistence. Thus there is a trade-off to be made between obtaining as large a sample size as possible whilst minimizing such effects. The rationale for the choice of the period December–March is based on the difference in background state between the

different months and is explained in the Supplementary Information. Based on this analysis, we do not expect the found regimes to be sensitive to the removal of the seasonal cycle. Differences in the occurrence and persistence of the regimes cannot be ruled out, and their behaviour throughout winter is discussed in the Supplementary Information.

2.1 | Clustering methods

The method used for the identification of circulation regimes in this study is k-means clustering using the standard Euclidian distance (L_2-norm; Jain, 2010). This method is applied to both the full field dataset as well as the time series of the first 5, 10, 15 and 20 EOFs. Furthermore, the method is also applied to the full field data after applying a 5- and a 10-day low-pass filter to remove high-frequency oscillations. The results for this time-filtered dataset are compared with those obtained by applying an adapted k-means algorithm to the unfiltered data that incorporates a persistence constraint in the clustering procedure itself. This persistent clustering method is described in what follows.

Given a dataset $\{x_t\}_{t \leq T} \in \mathbb{R}^m$, with t time, m the dimension of the data and T the span of the data given in time, the aim of any clustering method is to find a set of k cluster centres that accurately describe the dataset based on some measure. Let $\Theta = (\theta_1, \ldots, \theta_k)$ be the set of parameters describing the k cluster centres in either EOF or grid-point space. Here Θ represents the different circulation regimes for 500 hPa geopotential height anomaly data $\{x_t\}_{t \leq T}$. To assess how well the cluster centres represent the data, a model distance functional $g(x_t, \theta_i)$, giving the distance between a cluster centre and a data point, is required. We use the standard L_2-distance weighted by the cosine of latitude (Chung and Nigam, 1999). In addition we consider the affiliation vector $\Gamma = (\gamma_1(t), \ldots, \gamma_k(t))$, which indicates the weight of a certain cluster at some point in time. In practice $\gamma_i(t)$ is nearly always either zero or one, indicating to which cluster that point belongs. This is because a linear optimization problem always has an optimal solution on the boundary of the admissable set (Cottle and Thapa, 2017). For this reason, the affiliation vector is in general not considered when k-means clustering is discussed. Here we do consider this vector because it allows for the incorporation of persistence in the clustering procedure.

The task of identifying the atmospheric circulation regimes best representing the data means one has to find the optimal parameters for the cluster centres Θ and the affiliations of the data Γ. This is done by minimizing the averaged clustering functional (Franzke *et al.*, 2009)

$$\mathbf{L}(\Theta, \Gamma) = \sum_{t=0}^{T} \sum_{i=1}^{k} \gamma_i(t) g(x_t, \theta_i), \tag{1}$$

subject to

$$\sum_{i=1}^{k} \gamma_i(t) = 1, \quad \forall t \in [0, T], \quad \gamma_i(t) \geq 0,$$

$$\forall t \in [0, T], \quad i = 1, \ldots, k. \tag{2}$$

This is what the k-means procedure is doing implicitly, where $\gamma_i(t)$ is assumed to be zero or one. Finding the minimum of this functional minimizes the within-cluster variance as $\mathbf{L}$ is a measure of the distance between the cluster centres and the data points assigned to it. Because the within-cluster variance is minimized simultaneously for all clusters, the distance between data points assigned to different clusters becomes large. In other words: the between-cluster variance is maximized.

This clustering functional does not yet incorporate any persistence; an arbitrary reshuffling of the data leads to exactly the same result. To include persistence in the clustering method, we add a constraint on Γ that limits the number of transitions between regimes that is allowed (de Wiljes *et al.*, 2014). This constraint on the number of transitions between regimes, or switches, that are permitted throughout the whole time-series is:

$$\sum_{i=1}^{k} \sum_{t=0}^{T-1} |\gamma_i(t+1) - \gamma_i(t)| \leq C, \quad i = 1, \ldots, k, \tag{3}$$

for some constant C. The value of C gives twice the number of switches allowed (for a transition $\theta_1 \rightarrow \theta_2$ both the switch out of θ_1 and the one into θ_2 are counted), so e.g. an average cluster length of 5 days corresponds to $C = 2 \times \#\text{days}/5 \approx 1900$. In Table 1 the average regime duration corresponding to several values of C are given. Note that for a first-order Markov process, which is a relatively accurate assumption for the unconstrained regime dynamics, the average regime duration T_{av} can be related to the e-folding time scale T_e according to $T_{av} = 1/(1-p)$ and $T_e = -1/\log(p)$, where p is the self-transition probability (see Supplementary Information for details). Hence T_e is slightly shorter than T_{av} (e.g. 4.8 days for an average duration of 5.3 days), and they approach each other as $p \rightarrow 1$. The rationale behind the constraint is that in a chaotic atmospheric circulation not every data point can be straightforwardly assigned to a cluster. Some data points can be between clusters or outliers, for example, transitioning between clusters or extreme events. K-means clustering assigns these points to the nearest cluster (by

2804

Quarterly Journal of the
Royal Meteorological Society   RMetS   FALKENA ET AL.

TABLE 1 The value of C with corresponding average regime duration.

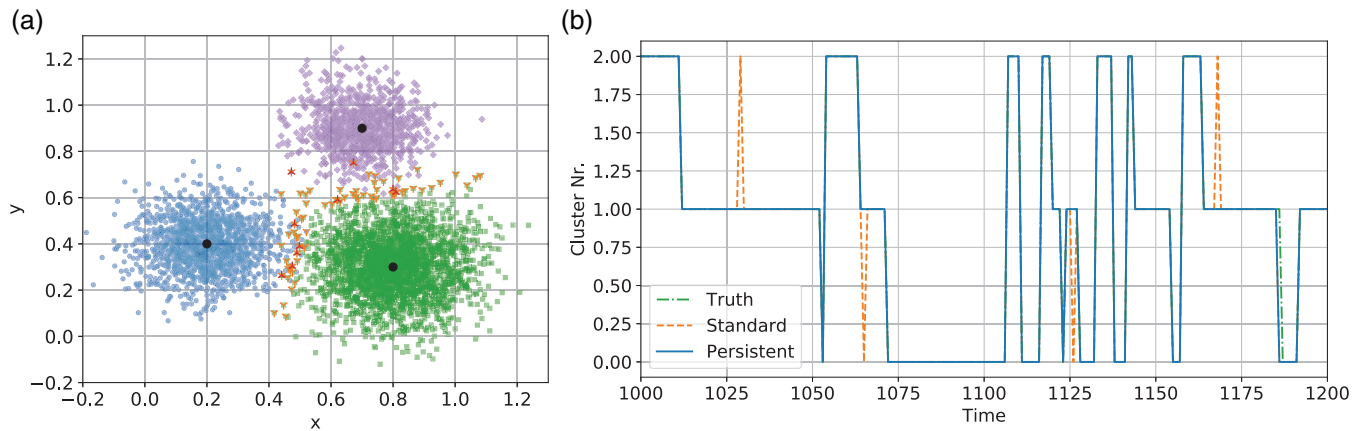| C	600	800	1000	1200	1400	1600	1800	2000	2200
Average duration (days)	15.8	11.8	9.5	7.9	6.8	5.9	5.3	4.7	4.3

FIGURE 1 The toy model showing (a) the three clusters (blue, green and purple) and (b) the assignment of the data to the different clusters in time. In (a) the data points wrongly assigned by standard k-means clustering are indicated by a Y-shape (orange) and the points that are still wrongly assigned when the persistence constraint is incorporated are indicated by an inverted Y-shape (red). (b) shows a part of the transition sequence which indicates in which cluster the data belong. The truth is shown dash-dotted (green, visible at the far right), the result for standard k-means dashed (orange), and the result when the persistence constraint is incorporated solid (blue). Note that the lines overlap most of the time.

distance), while it can be more sensible to assign them to the same cluster as their neighbours if the distance to that cluster is also quite small. This is exactly what the constraint in Equation 3 is doing for reasonable values of C.

The minimization of the clustering functional **L** taking into account the persistence constraint is done in two steps which are iterated until convergence:

1. For fixed Θ, minimize **L** over all possible values of Γ.
2. For fixed Γ, minimize **L** over all possible values of Θ.

The first part is done by linear programming using the Gurobi package for Python (Gurobi Optimization, 2019). The second part is done by k-means clustering. The computation is terminated when the difference between consecutive **L** becomes smaller than a set tolerance.

To make the effect of the incorporation of this constraint on the final clustering result more insightful, a very simple toy model is presented (details given in the Supplementary Information). Consider a system of three 2D clusters which are normally distributed around their respective centres, each with a different variance. They transition into each other according to a persistent transition matrix, meaning there is a high probability of a cluster transitioning to itself. The clusters are shown in Figure 1a. When applying the standard (unconstrained)

k-means algorithm, the data that are assigned to the wrong cluster are located between the different clusters. In the assignment of the data to the clusters, this leads to sudden jumps into one cluster and directly back to the original cluster, as can be seen in Figure 1b, which leads to the identification of a too short persistence. When the persistence constraint is incorporated, many of the wrongly assigned data points are now assigned to the correct cluster. Furthermore, we see that the short transitions into and directly out of a cluster are removed. Thus the persistence found using the constrained algorithm is closer to that of the real system.

Because k-means clustering can only identify local minima, we run the algorithm 500 times with different random initial conditions for the standard k-means algorithm. The initial condition at every location in space is drawn independently from a normal distribution around zero with the same standard deviation as the data. Note that this means there is no correlation in space, so as to not make any assumptions on the spatial patterns of the regimes. The tolerance used depends on k and is $0.0001/k^2$ for the full field data and $0.001/k^2$ for the EOF data. The algorithm including the persistence constraint is run 100 times with different initial conditions; the reduced number of runs is due to increased computation time by the incorporation of linear programming, which is similar for the EOF and full field data. The final result is chosen to

be the one with the smallest clustering functional **L**. This method in general works reasonably well, but even for the Lorenz 63 system (Lorenz, 1963) the "correct" clusters cannot be found for every realisation (simulation with different initial condition) when no persistence constraint is incorporated, as can be seen in the Supplementary Information. The data close to the cluster centres is always assigned correctly, but there is a substantial uncertainty in assigning the data further away from the cluster centres. Thus it is important to be careful when applying k-means clustering and not blindly trust the result, especially as the method assigns every data point to a cluster even if it actually is between different clusters. The incorporation of the persistence constraint improves this aspect of the k-means result.

2.2 | Information criteria

For k-means clustering, the number of clusters k has to be set *apriori* and the question is how to determine the best value for k. The main methods used to this end in the identification of atmospheric circulation regimes are verification by synthetic datasets (e.g. Straus *et al.*, 2007; Straus *et al.*, 2017), using a classifiability index (e.g. Michelangeli *et al.*, 1995; Plaut and Simonnet, 2001), and looking at the similarity of runs with different initial conditions (e.g. Jung *et al.*, 2005). An alternative method is to use an information criterion (e.g. O'Kane *et al.*, 2013), which is widely used in for example biological sciences (e.g. Volinsky and Raftery, 2000; Posada and Buckley, 2004; Arnold, 2010). An information criterion is a tool from model selection which is used to identify the optimal model (Burnham and Anderson, 2004); it strikes a balance between how well the model fits the data and the number of parameters needed, to prevent over-fitting. The optimal balance is where the information criterion is minimal. As the clusters are effectively a model representing the data, the concept can be applied here as well. This tool for identifying the optimal number of clusters has already been used in many applications and more theoretical studies (e.g. Chen and Gopalakrishnan, 1998; Fraley and Raftery, 1998; Cobos *et al.*, 2014). In addition to allowing to identify the optimal number of clusters k, an information criterion also allows for finding the best constraint value C when persistence is incorporated in the clustering procedure.

The two information criteria that are used most widely are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Burnham and Anderson, 2004). The AIC is based in information theory and is an approximation of how different two probability distributions (one for the data, one for the model) are

(Akaike, 1973). It is given by

$$\text{AIC} = -2\log(\mathcal{L}(\hat{\theta}|\text{data})) + 2K, \quad (4)$$

where $\mathcal{L}(\hat{\theta}|\text{data})$ is the likelihood of the optimal model $\hat{\theta}$ given the data, which measures how well the model fits the data, and K the number of parameters in the model. For cluster analysis, the number of parameters is determined by the number of clusters, their dimension and the length of the data time series (expressions are given in the Supplementary Information). In contrast, the BIC is based on the limiting behaviour of Bayes estimators, which minimizes the expectation value of the loss (e.g., error), and reads (Schwarz, 1978):

$$\text{BIC} = -2\log(\mathcal{L}(\hat{\theta}|\text{data})) + K\log(n), \quad (5)$$

where n is the sample size, here being the dimension of the data (number of grid points or number of EOFs) times the number of days. Just as with the AIC, the BIC strikes a balance between the (log-)likelihood, i.e. how well the clusters fit the data, and the number of parameters in the model. For both the AIC and BIC, we refer to the second term as the penalty term since it penalizes the use of many parameters in finding the optimal model to prevent over-fitting.

To compute the values of both information criteria, the log-likelihood is needed. Assuming the errors of the model are independent and normally distributed, the log-likelihood term can be written as (Burnham and Anderson, 2004)

$$-2\log(\mathcal{L}(\hat{\theta}|\text{data})) = n\log(\hat{\sigma}^2), \quad (6)$$

where $\hat{\sigma}^2 = \sum \hat{\epsilon}_t^2/n$ is the error variance for residuals $\hat{\epsilon}_t$, the latter being the difference between the cluster centres and the data for every grid point or EOF. This allows for a straightforward computation of both information criteria using the clustering functional.

The only difference between the AIC and the BIC is how they penalize the number of parameters in the model. The penalty term in the BIC takes into account the sample size, while the term in the AIC does not. This means the penalty term in the BIC is stronger with respect to the number of parameters, accounting for an assumed higher variability in a high-dimensional dataset, which increases the chances of over-fitting. Note that the number of parameters is different between the EOF and full field data since the number of parameters needed to identify a cluster centre is set by the dimension of the data. The dimension of EOF data is significantly smaller than that of full field data. As a consequence, the penalty term of the BIC is stronger relative to the log-likelihood term for EOF data compared

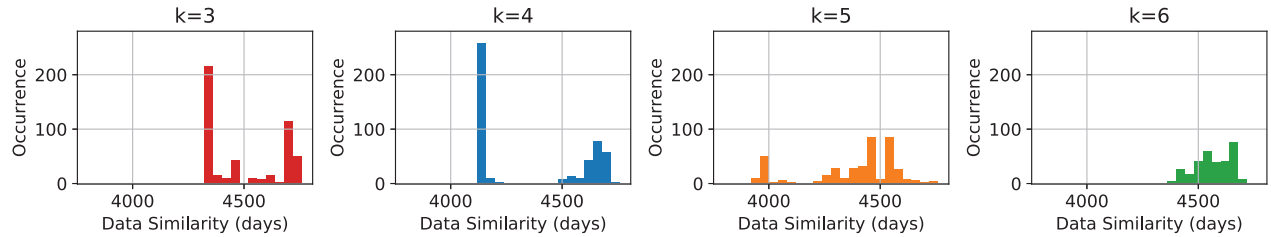(a) The clustering functional L.
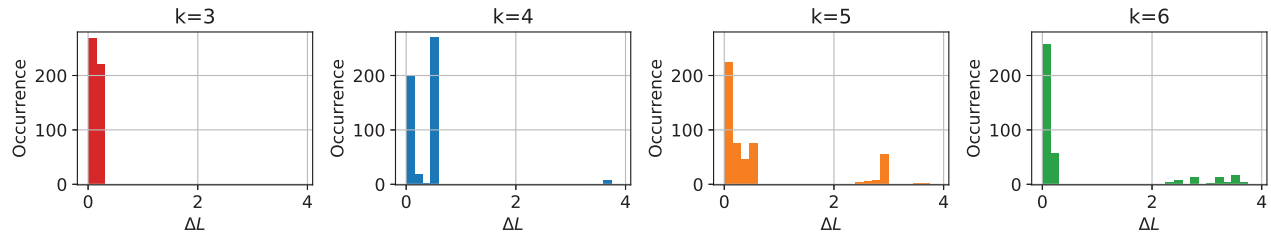
(b) The data similarity.

FIGURE 2 Histograms for (a) the difference of the clustering functional with its minimum value $\Delta \mathbf{L} = \mathbf{L}_{run} - \mathbf{L}_{min}$ and (b) the data similarity with respect to the optimal (minimal $\mathbf{L}$) result using the first 20 EOFs for $k = 3, \dots, 6$.

to full field data. This is illustrated in Table S2 using 20 EOFs. For this reason the BIC is expected not to perform well for EOF data, by which we mean that it will identify a very low number of clusters k to be optimal. On the other hand, the penalty term in the AIC likely is too weak to yield a realistic optimal k for the full field data. A high number of clusters is expected to be found optimal, well beyond what is physically reasonable and suitable. When using either of these criteria, one always has to judge whether the result is sensible for the purpose of the study.

3 | RESULTS

First we compare the results of the regimes found using the EOF data and the full field data. In this we focus on the optimal number of regimes using the information criteria discussed. Second we discuss the difference between enforcing persistence in the clustering method and the use of time-filtered data on the occurrence and persistence of the regimes.

3.1 | Number of regimes

The standard number of wintertime regimes identified over the Euro-Atlantic sector in literature is four (e.g. Vautard, 1990; Cassou, 2008; Dawson and Palmer, 2015). Few studies question this number (e.g. Fereday *et al.*, 2008). This optimal number of four clusters has always been found in the context of EOF data. The most-used argument for four regimes being optimal is based on how consistent,

or similar, the results of the k-means algorithm are for different (random) initial conditions. A way in which this is commonly assessed is the classifiability index introduced by Michelangeli *et al.* (1995), which uses pattern correlation to determine how similar two sets of regimes are. The value of the index is computed for both the data and a synthetic dataset with the same statistics, to see whether there exists a k for which the result is significantly different from a noise model. Another method to assess the similarity is to look at the assignment of the data to the different regimes (Fereday *et al.*, 2008). This data similarity measure considers the regimes for all data (instead of only their aggregate) and thus possibly provides more information than the pattern correlation.

Here we briefly examine the data similarity, as well as the spread in the clustering functional $\mathbf{L}$, for the EOF regimes with the main aim being to verify the reliability of the found regimes. Histograms for $\Delta \mathbf{L} = \mathbf{L}_{run} - \mathbf{L}_{min}$, being the difference of $\mathbf{L}$ for each run with the minimal value over all runs of the clustering algorithm with different initial conditions, and the number of days assigned to the same regime (as the best result) are shown in Figure 2. The first aspect to note is that for some k not only a global minimum is found, but also a local one (e.g. for $k = 4$ in Figure 2a). In addition we see that the found regimes can be quite different in their assignment of data to certain regimes, indicating the found regimes are significantly different. Instead of only combining all accurate results (near the global minimum of $\mathbf{L}$) by computing the average data similarity (similar to the classifiability index; Michelangeli *et al.*, 1995) we also look at its variance. The values for $k = 3, \dots, 6$, given on the left side of Table 2, show that

TABLE 2 The mean (μ) and variance per cluster (σ^2/k) of the data similarity for data with $\mathbf{DL} = \mathbf{L}_i - \mathbf{L}_{i+1}$, where the $\mathbf{L}_i$ are sorted from small to large, below a set threshold for both the EOF and full field results. For the full field results, the values for odd and even years are also given. The number of runs that are below the threshold is shown as well.

| | 20 EOFs: DL < 0.01 | | | Full field: DL < 0.0005 | | | | | | | | |
| | All years | | | All years | | | Odd years | | | Even years | | |
k	μ	σ^2/k	#data	μ	σ^2/k	#data	μ	σ^2/k	#data	μ	σ^2/k	#data
3	4,643	3,649	254	4,552	2,109	485	2,342	875	409	2,246	140	156
4	4,658	330	197	4,607	1,440	201	2,359	452	248	2,255	37	423
5	4,509	978	265	4,660	149	204	2,187	3,564	274	2,243	132	137
6	4,571	1,103	315	4,581	790	316	2,296	322	60	1,911	10,478	210
7	—	—	—	3,686	37,440	417	1,998	12,145	165	2,113	1,470	170

the variance is lowest for $k = 4$, which goes together with a high average. This is consistent with the results found in literature. Interestingly the mean for $k = 6$ is slightly higher than found for $k = 5$, while the opposite would be expected as more clusters allow for more variability.

Next we turn to the distribution of $\Delta\mathbf{L}$ and the data similarity for the full field data, as shown in Figure 3. The first thing to note is that the distributions of $\Delta\mathbf{L}$ look similar to those for the EOF data, indicating that using the high-dimensional full field data does not reduce the chance of finding the optimal regimes. Some differences with the EOF result do occur for the data similarity, most notably the increased similarity for $k = 5$. Looking at the results for the global minimum we find that both $k = 5$ and $k = 6$ show a smaller variance than $k = 4$, in contrast to the EOF results. Especially for $k = 5$ the difference is substantial. These differences indicate that by performing an EOF analysis some information is lost, resulting in a stronger consistency of k-means for $k = 5$ and $k = 6$ when using the full field data.

In addition we look at the results of the clustering algorithm for a subset of the data, which is a standard approach to test the robustness of clustering methods to, for example, identify cats on photos (Jain, 2010). Here the results for the datasets of odd and even years are studied, as stationarity of the dataset cannot be assumed. The differences between the results for odd and even years are found to be large, indicating either $k = 4$ or $k = 6$ having the smallest variance. Also for $k = 5$ differences between the two sets of years are large. These ambiguous results raise the question whether half the dataset is of sufficient length to draw reliable conclusions about the clustering results. This also means that non-stationarity of the regimes due to e.g. climate change is difficult to study accurately using clustering methods applied to reanalysis datasets.

We refrain from drawing definite conclusions about the optimal number of clusters k from the above discussion on consistency, as there is some debate about its suitability

for this purpose (Philipp *et al.*, 2007). Instead, we use the AIC and BIC to identify the optimal number of regimes. The AIC is used when considering the EOF results, as it is expected to give better results in that case as discussed in Section 2.2. In Figure 4a the AIC is shown for using 5, 10, 15 and 20 EOFs to identify the circulation regimes. A minimum at $k = 4$ is found when 20 EOFs are used, although the AIC is also small for $k = 3$ and $k = 5$. For lower numbers of EOFs, the optimal number is found to be lower, while a higher number of EOFs leads to a higher optimum for k. This is to be expected because the use of a limited number of EOFs means that some variability of the original data is neglected. This loss of variability is larger when fewer EOFs are used, and as a consequence fewer clusters are needed to account for the variability of the EOF data. The BIC has its minimum at $k = 2$ for every number of EOFs considered, indicating the penalty term for the number of parameters is indeed too strong for the EOF data (Section 2.2). Based on the AIC we conclude that $k = 4$ is indeed the optimal number when using 20 EOFs, which corresponds with results from literature. However, for other numbers of EOFs the optimal k according to the AIC is different, meaning this conclusion is not unambiguous.

The optimal number of regimes identified by an information criterion when using the full field dataset is not the same. The BIC is more suitable for the full field data than the AIC due to the dependence of the penalty term on the sample size, making it stronger than the penalty term of the AIC (Section 2.2). In Figure 4b both the AIC and BIC are shown for the full field data. The BIC points towards an optimum of $k = 6$. The AIC does not show a minimum in the range considered as the penalty term is not strong enough for the high-dimensional full field data. Therefore, we base our decision on the optimal number of regimes on the BIC and find $k = 6$ to be optimal.

The regimes that are obtained by using ten or more EOFs and the full field data do not differ substantially

(a) The clustering functional L.
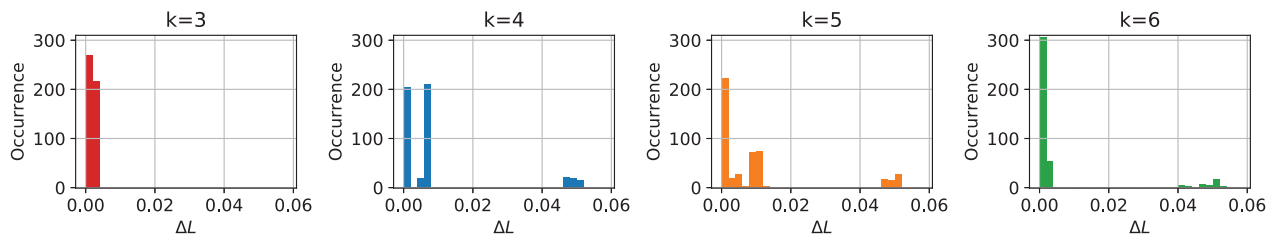
(b) The data similarity.
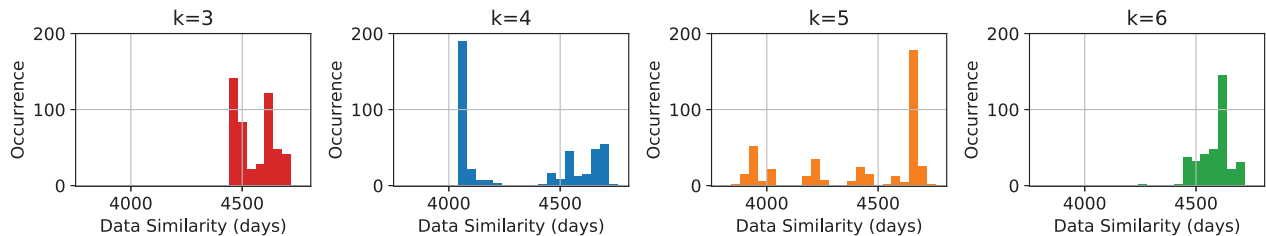
FIGURE 3   As Figure 2, but using the full field data.

FIGURE 4   Information criteria for both the full field and EOF datasets for the range $k = 2, \ldots, 10$. (a) shows the AIC for different numbers of EOFs, and (b) the AIC and BIC for the full field data.

(a) The AIC for different numbers of EOFs.

(b) The AIC and BIC for the full field data.
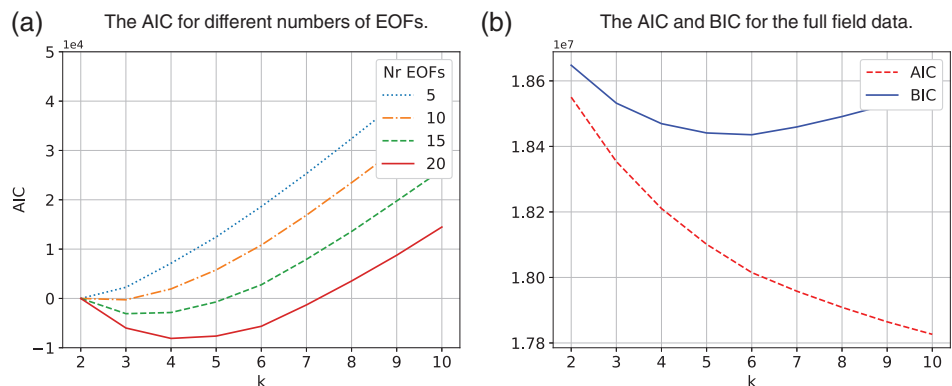
for the same k. Similarly the occurrence rate and transition probabilities of the regimes do not differ substantially. In Figure 5 the four regimes known from literature (e.g. Straus *et al.*, 2007; Hannachi *et al.*, 2012) are shown as obtained by applying k-means clustering on the full field data for $k = 4$. They are the Atlantic Ridge (AR), Scandinavian Blocking (SB) and the two phases of the North Atlantic Oscillation (NAO). The transition probability of a regime to itself (or daily re-occurrence rate) and overall occurrence rate of these regimes can be found in Table 3. The positive phase of the NAO is the most frequently occurring regime, followed by SB. The high occurrence of the NAO+ regime may reflect the fact that it is the only regime associated with a northern low pressure area. Both phases of the NAO are found to be most persistent, i.e. transition to themselves with the highest probability, while the AR exhibits the least persistence. We note that the occurrence rates obtained are similar to those found in literature despite not using a seasonally varying background state.

In Figure 6 the regimes found using k-means clustering on the full field data for $k = 6$ are shown. The first four regimes are in essence the same as those found for $k = 4$ in Figure 5. Small differences occur in the location of the maximum high or low pressure area for the AR and NAO+. The two additional regimes found have a low pressure area either in the central Atlantic or over Scandinavia. The first thus can be identified as the opposite phase of the AR with a pattern correlation of −0.57 and we refer to it as AR−, while the original regime is denoted by AR+. Similarly we refer to the second additional regime as SB−, as it represents the opposite phase of the SB regime (from now on denoted by SB+), where the pattern correlation −0.49 is slightly lower. Note that the pattern correlation of the two phases of the NAO is higher for $k = 6$, with the value being −0.93, versus −0.59 when using $k = 4$. The use of six clusters thus introduces a pleasing symmetry in the found regimes, with an equal number of regimes having a high and low pressure area in the north of the domain.

TABLE 3 The values of the occurrence rate, self-transition probability, e-folding time and average regime duration for the unconstrained result for both $k = 4$ and $k = 6$.

	$k = 4$				$k = 6$					
	AR	**SB**	**NAO+**	**NAO−**	**AR+**	**SB+**	**NAO+**	**NAO−**	**AR−**	**SB−**
Occurrence (%)	21.3	26.8	31.5	20.4	15.6	19.6	16.9	15.5	16.3	16.1
Self-transition probability	0.756	0.792	0.850	0.849	0.712	0.748	0.751	0.847	0.787	0.730
e-folding time (days)	3.6	4.3	6.2	6.1	2.9	3.4	3.5	6.0	4.2	3.2
Average duration (days)	4.1	4.8	6.7	6.6	3.5	4.0	4.0	6.5	4.7	3.7
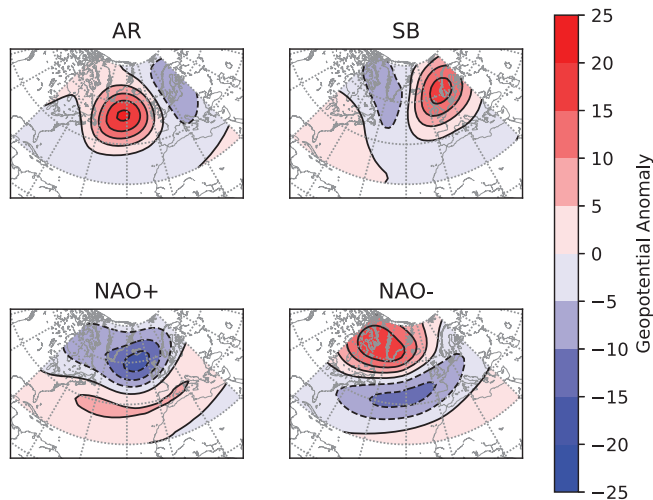
FIGURE 5 The clustering result of the standard k-means algorithm applied to the full field data for $k = 4$.

FIGURE 6 The clustering result of the standard k-means algorithm applied to the full field data for $k = 6$.

The occurrence rate and self-transition probability of the six regimes show different behaviour than found for $k = 4$, as can be seen in Table 3. Instead of the NAO+, the SB+ is found to be the most frequently occurring regime. The NAO+ is the second ranked regime in occurrence, albeit with a small but significant difference relative to SB+. The other four regimes show similar occurrence rates. When looking at the self-transition probabilities, the NAO− remains the most persistent, with exactly the same probability. However the NAO+ does lose some of its persistence, reducing its self-transition probability to a rate similar to that of the SB+. The AR− is found to be the second most persistent regime and the AR+ remains the least persistent. In Table 3 the e-folding time and average regime duration computed using the self-transition probability are also given for reference. Furthermore, we note that the occurrence rate and self-transition probabilities change throughout the winter months. Figures showing this seasonal behaviour are given in the Supplementary Information. This variation could reflect the effect of the seasonal cycle in the data, as well as intrinsic variability in the regime behaviour.
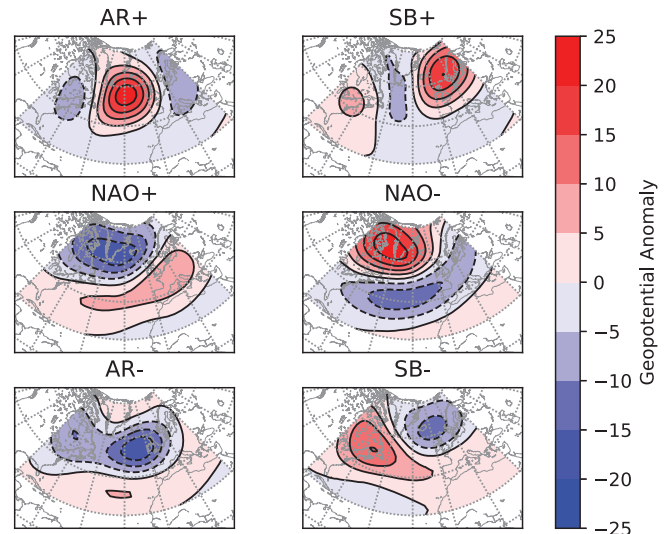
3.2 | Persistence

In Section 2.1 two methods to enforce persistence of the atmospheric circulation regimes have been discussed. The first method is the standard approach of applying a time-filter, and the second method is to include a persistence constraint in the clustering algorithm itself. Note that the average regime duration over the unconstrained regimes following Table 3 is 5.5 days for $k = 4$ and 4.4 days for $k = 6$, which means the persistence constraint will only affect the result when it is below either $C \approx 1800$ or $C \approx 2200$ for $k = 4$ or $k = 6$ respectively (Table 1). The regimes found using these two methods do not differ substantially from those found and discussed in the previous section. When a time-filter is applied to the data, the regimes are found to be slightly weaker, meaning the maximum and minimum geopotential height anomaly are smaller, but they do not show a visible difference in the configuration of high and low pressure areas. For the results using a persistence constraint, differences in the regimes only emerge for very strong (unrealistic) constraints in
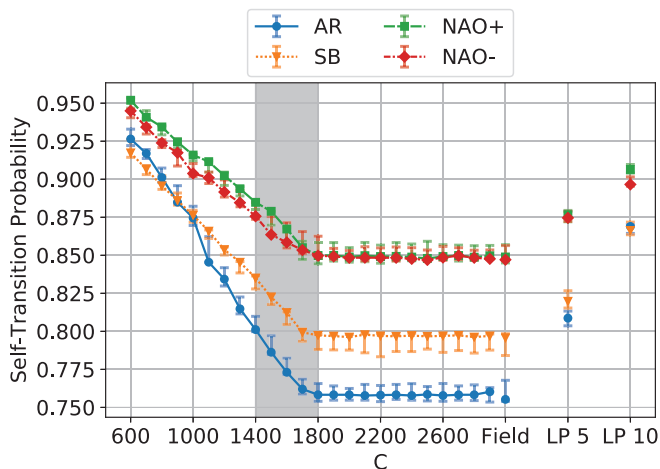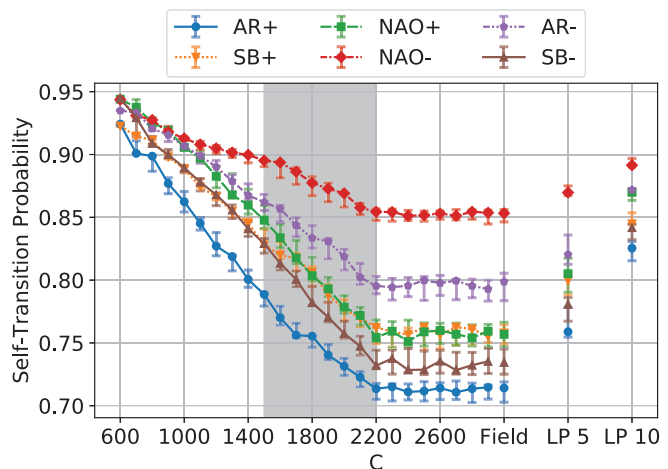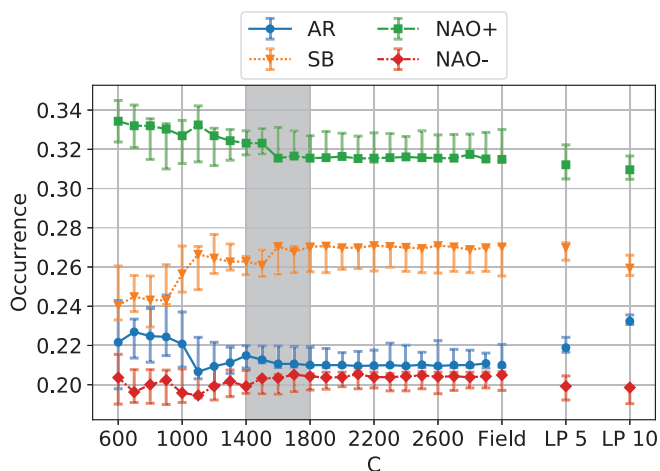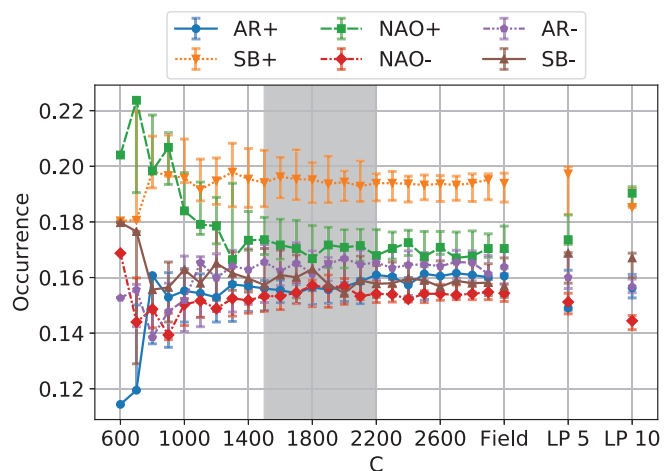
(a) The self-transition probabilities for $k = 4$.

(b) The self-transition probabilities for $k = 6$.

(c) The occurrence rates for $k = 4$.

(d) The occurrence rates for $k = 6$.

FIGURE 7 The (a, b) self-transition and (c, d) occurrence probabilities of the different regimes for (a, c) $k = 4$ and (b, d) $k = 6$ for the clustering results including the persistence constraint depending on the value of C (corresponding average regime durations can be found in Table 1). On the right side of each panel, the values for the unconstrained algorithm (Field) and the 5- and 10-day low-pass filters are shown. The error bars indicate the maximum and minimum value of occurrence/self-transition probabilities for clustering results with a slightly larger **L** (bounds for the difference are {0.00968, 0.00936, ..., 0.00232, 0.002, 0.002, 0.002} decreasing with increasing C, which are chosen sufficiently small to give similar regimes according to the data correspondence). The grey bands indicate the region in which the persistence constraint is considered to act as a good filter.

the form of slight shifts in the location of the centres of high and low pressure. For weak (realistic) constraints, the regimes found are the same as for the unconstrained method and no weakening is found. By a "realistic" constraint we mean one that does not force data points into regimes which are a large distance away, but only switches those data points that are between different regimes, as can be seen in the toy example in Section 2.1. In practice these are constraints corresponding to an average regime duration below ~ 9 days (the corresponding C can be found in Table 1).

The effects of the time-filtering and persistence constraint method on the self-transition probability are shown in Figures 7a ($k = 4$) and 7b ($k = 6$). On the left in each

panel, the results for the constrained algorithm are shown for various C and as expected the self-transition probability increases with decreasing C. The smaller the value of C, the fewer switches between regimes are allowed. The increase of the self-transition probability with decreasing C is approximately linear for all regimes, and starts at the "raw" self-transition probability of the regimes. Consistently with the values computed using the average regime duration in Table 3, we find that the constraint starts to affect the self-transition probabilities around either $C = 2200$ for $k = 6$ (4.3 days) or $C = 1800$ for $k = 4$ (5.3 days). Note that, when the constraint starts to affect the results, the regime dynamics can no longer be described as a first-order Markov process and care must be taken in

interpreting the average regime duration and corresponding self-transition probabilities.

Comparing the results for time-filtered data with those of the constrained method in Figure 7, we see that using a 5-day low-pass filter corresponds to a constraint of roughly 2000 for $k = 6$ and 1400 for $k = 4$. This difference is mainly due to the stronger effect of the constraint for a larger number of clusters. For the 10-day filter, the corresponding values of C are approximately 1400 and 1100 for $k = 6$ and $k = 4$ respectively. Note that the self-transition probability of certain regimes differs slightly between the two methods. For example, the AR+ regime is found to increase its self-transition probability more strongly for the time-filtered data.

The occurrence rates of the different regimes are shown in Figures 7c and 7d for $k = 4$ and $k = 6$ respectively. We start by looking at the results of the constrained algorithm. The occurrence rate remains the same as for the unconstrained data, even for constraint values significantly stronger than the "raw" persistence of the data. Only for very low C (strong constraints) do the occurrence rates start to differ. This indicates that the method indeed causes a switching of the "between-cluster" points to the cluster of their neighbours instead of the cluster they are slightly closer to, as expected from the results of the toy model (Section 2.2). We regard the constraint as being "weak" so long as the occurrence rates are not affected, and help to identify true physical persistence. For these values of C we consider that the persistence constraint acts as a good filter, indicated by the grey bands in Figure 7. In contrast, the results for the time-filtered data show significant differences in the occurrence rates of the regimes. Especially for the 10-day filter, the differences, for example the AR+ regime ($k = 4$) or the NAO+ regime ($k = 6$), are substantial. As this is the standard filter used in literature (e.g. Straus *et al.*, 2017), it raises the question of how reliable the occurrence rates are and whether they are not solely a feature of the method used. In contrast, the inclusion of the constraint within the clustering procedure itself does not lead to such a bias and therefore provides a more robust way of finding persistent regimes, that is, of isolating the signal from the noise.

When using the constrained algorithm, one of the choices that needs to be made is which constraint value C is best to use. Here we base this choice on the BIC, as we did for finding the optimal number of clusters k. In Figure 8 the BIC is shown for both $k = 4$ and $k = 6$. For $k = 4$ the minimum is found for $C = 1400$ and for $k = 6$ it is found for $C = 1500$. These constraint values correspond to an average regime duration of 6.8 and 6.3 days respectively. The lower end of the region in which the BIC is close to its minimum coincides with the point beyond which smaller values of C start to affect the occurrence rates, giving a
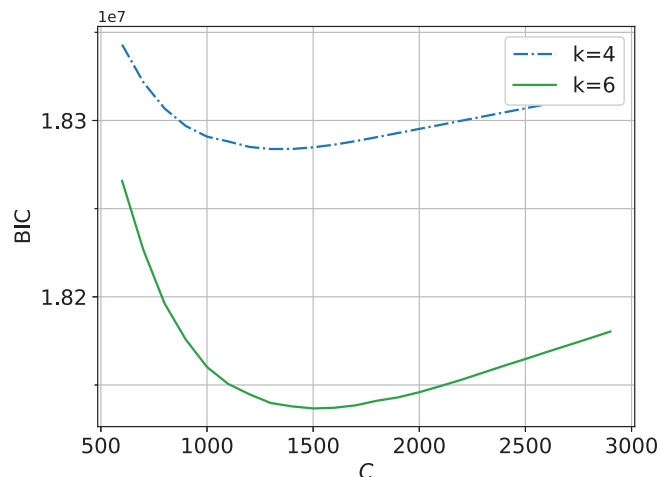

FIGURE 8 The BIC for the clustering with persistence constraint for $k = 4$ and 6 and C in the range $\{600, \ldots, 2900\}$.

lower bound for the region where the persistence constraint is considered to act as a good filter as indicated by the grey bands in Figure 7. This increases the confidence of the optimal value of C being around these values. Interestingly, the optimal average regime duration for $k = 4$ and $k = 6$ differs by less than 10% ($\Delta C \approx 1500 - 1400 = 100$), whereas without the persistence constraint the average duration differs by 20% ($\Delta C \approx 2200 - 1800 = 400$). This confirms that the persistence constraint is helping identify a physical signal that is less dependent on the number of clusters chosen. The range where the BIC is very close to its minimum is between 6.3 ($C = 1500$) and 7.9 ($C = 1200$) days for $k = 4$. For $k = 6$ this range is from 5.9 ($C = 1600$) to 6.8 ($C = 1400$) days. Twice this time-scale, which is the minimum for recurrence of a regime (i.e., for regimes A and B we have $A - B - A$), thus corresponds roughly to a period of 12 to 14 days. This is somewhat longer than the time-scale of synoptic weather systems (Blackmon *et al.*, 1977; Boljka *et al.*, 2018). As this is an average, there are a substantial number of longer-lasting regimes showing persistence well beyond the synoptic time-scale.

4 | CONCLUSION AND DISCUSSION

In this study we have shown, using an information criterion and further arguments based on the consistency of the clustering result, that the traditional number of four clusters is not optimal for representing wintertime Euro-Atlantic weather regimes when full field data are used. The traditional approach of applying clustering to the first few EOFs involves a loss of information, which affects the number of regimes that best represent the data. The optimal number of regimes for the full field data was

identified using the Bayesian Information Criterion (BIC), which strikes a balance between how well the regimes fit the data and the number of parameters needed to describe them.

We find that for the full field data, six regimes is the optimal choice. The two additional regimes are the opposite phases of the Atlantic Ridge and Scandinavian Blocking, introducing a pleasing symmetry in the found clusters. Furthermore, the dominant occurrence of the NAO+ when there are only four clusters, which likely is due to it being the only regime with a low pressure area in the north, is reduced by the addition of two regimes that also have this feature. Therefore, six regimes allow for more variability in their representation of the circulation and prevent all data with a more zonal flow from projecting onto the NAO+.

Furthermore, we looked into ways to enforce persistence of the regimes. A common approach in literature is to apply a low-pass filter to remove high-frequency oscillations and focus on the persistent behaviour (e.g. Bao and Wallace, 2015; Straus *et al.*, 2017). This alters the data to which the clustering algorithm is applied, just as the use of EOFs does. We have shown that this leads to a significant change in the occurrence rate of the circulation regimes. A new method, which incorporates a persistence constraint in the algorithm itself, does not change the data while still enforcing persistent regimes. The results for this approach do not exhibit the change in occurrence rate found for the time-filtered data, as long as the constraint is not too strong, while still having an increased self-transition probability. Therefore this method leads to a more robust and unbiased result than for the time-filtering approach.

A choice that needs to be made in this adapted clustering method is the value of the constraint C. Using the BIC the optimal value of C is found to lie around an average regime duration of six to seven days. Interestingly, this matches the point beyond which smaller values of C start to affect the occurrence rates. Thus it can be viewed as a more accurate estimate of the physical persistence of the regimes than that provided by the raw data without the persistence constraint. Double this value, which is the minimum for recurrence of a regime, is thus slightly longer than that of synoptic weather systems (Blackmon *et al.*, 1977; Boljka *et al.*, 2018). This shows that the atmospheric circulation indeed exhibits persistence beyond the synoptic time-scale, suggesting the presence of predictable low-frequency modes.

Both results indicate that care must be taken when applying filtering methods (EOFs, low-pass filters) to the data before a clustering algorithm is applied. Clustering itself provides a means of dimension reduction, by projecting onto components representing recurrent patterns in the data. Since this is a method of filtering the data, it seems ill-advised to apply this to already filtered EOF data, as it is not clear what the effect of this double filtering has on the result. A similar argument holds for applying a time-filter to the data before clustering. Information is lost in this procedure, introducing a bias in the resulting circulation regimes and their occurrence rates.

SUPPORTING INFORMATION

Supporting Information is available online. It contains a discussion of the effects of seasonality on the regimes, technical details of the time-filter, the toy model, the information criteria and relation between the e-folding time and the average regime duration, as well as additional figures of distributions of ΔL and the data similarity, and clustering results for the Lorenz 63 system.

ORCID

Swinda K.J. Falkena https://orcid.org/0000-0002-2317-9213

REFERENCES

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle, pp. 267–281 in Second International Symposium on Information Theory, Tsahkadsor, Armenia, 2–8 September 1971. Petrov, B.N., Csáki, F. (eds), Akadémiai Kiadó, Budapest.

Amini, S. and Straus, D.M. (2018) Control of storminess over the Pacific and North America by circulation regimes. *Climate Dynamics*, 52(7), 4749–4770. https://doi.org/10.1007/s00382-018-4409-7

Arnold, T.W. (2010) Uninformative parameters and model selection using Akaike's information criterion. *Journal of Wildlife Management*, 74(6), 1175–1178. https://doi.org/10.2193/2009-367

Bao, M. and Wallace, J.M. (2015) Cluster analysis of Northern Hemisphere wintertime 500 hPa flow regimes during 1920–2014. *Journal of the Atmospheric Sciences*, 72(9), 3597–3608. https://doi.org/10.1175/jas-d-15-0001.1

Blackmon, M.L., Wallace, J.M., Lau, N.-C. and Mullen, S.L. (1977) An observational study of the Northern Hemisphere wintertime circulation. *Journal of the Atmospheric Sciences*, 34, 1040–1053.

Boljka, L., Shepherd, T.G. and Blackburn, M. (2018) On the coupling between barotropic and baroclinic modes of extratropical atmospheric variability. *Journal of the Atmospheric Sciences*, 75, 1853–1871. https://doi.org/10.1175/JAS-D-17-0370.1

Burnham, K.P. and Anderson, D.R. (2004) Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33(2), 261–304. https://doi.org/10.1177/0049124104268644

Cassou, C. (2008) Intraseasonal interaction between the Madden–Julian Oscillation and the North Atlantic Oscillation. *Nature*, 455(7212), 523–527. https://doi.org/10.1038/nature07286

Cassou, C., Laurent, T. and Phillips, A.S. (2005) Tropical Atlantic influence on European heat waves. *Journal of Climate*, 2805–2811. https://doi.org/10.1175/JCLI3506.1

Chen, S.S. and Gopalakrishnan, P.S. (1998). Clustering via the Bayesian information criterion with applications in speech recognition, pp. 645–648 in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP'98, https://doi.org/10.1109/ICASSP.1998.675347.

Cheng, X. and Wallace, J.M. (1993) Cluster analysis of the Northern Hemisphere wintertime 500 hPa height field: spatial patterns. *Journal of the Atmospheric Sciences*, 50(16), 2674–2696.

Chung, C. and Nigam, S. (1999) Weighting of geophysical data in principal component analysis. *Journal of Geophysical Research*, 104(D14), 16925–16928.

Cobos, C., Muñoz-Collazos, H., Urbano-Muñoz, R., Mendoza, M., León, E. and Herrera-Viedma, E. (2014) Clustering of web search results based on the cuckoo search algorithm and Balanced Bayesian Information Criterion. *Information Sciences*, 281, 248–264. https://doi.org/10.1016/j.ins.2014.05.047

Corti, S., Molteni, F. and Palmer, T.N. (1999) Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature*, 398(6730), 799–802. https://doi.org/10.1038/19745

Cottle, R.W. and Thapa, M.N. (2017) *Linear and Nonlinear Optimization*. Springer, Berlin.

Dawson, A. and Palmer, T.N. (2015) Simulating weather regimes: impact of model resolution and stochastic parameterization. *Climate Dynamics*, 44(7-8), 2177–2193. https://doi.org/10.1007/s00382-014-2238-x

Dawson, A., Palmer, T.N. and Corti, S. (2012) Simulating regime structures in weather and climate prediction models. *Geophysical Research Letters*, 39(21), 1–6. https://doi.org/10.1029/2012GL053284

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F. (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137, 553–597. https://doi.org/10.1002/qj.828

Deutscher Wetterdienst (2019). Grosswetterlagen Forecast (GWL). Available at: https://www.dwd.de/EN/research/weatherforecasting/met_applications/nwp_applications/grosswetterlagen_forecast.html; accessed 15 May 2020.

de Wiljes, J., Putzig, L. and Horenko, I. (2014) Discrete nonhomogeneous and nonstationary logistic and Markov regression models for spatiotemporal data with unresolved external influences. *Communications in Applied Mathematics and Computational Science*, 9(1), 1–46. https://doi.org/10.2140/camcos.2014.9.1

Fereday, D.R., Knight, J.R., Scaife, A.A., Folland, C.K. and Philipp, A. (2008) Cluster analysis of North Atlantic–European circulation types and links with tropical Pacific sea surface temperatures. *Journal of Climate*, 21(15), 3687–3703. https://doi.org/10.1175/2007jcli1875.1

Ferranti, L., Corti, S. and Janousek, M. (2015) Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, 141, 916–924. https://doi.org/10.1002/qj.2411

Fraley, C. and Raftery, A.E. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8), 578–588. https://doi.org/10.1093/comjnl/41.8.578

Franzke, C., Horenko, I., Majda, A.J. and Klein, R. (2009) Systematic metastable atmospheric regime identification in an AGCM. *Journal of the Atmospheric Sciences*, 66(7), 1997–2012. https://doi.org/10.1175/2009JAS2939.1

Grams, C.M., Beerli, R., Pfenninger, S., Staffell, I. and Wernli, H. (2017) Balancing Europe's wind-power output through spatial deployment informed by weather regimes. *Nature Climate Change*, 7(8), 557–562. https://doi.org/10.1038/NCLIMATE3338

Gurobi Optimization (2019) *Gurobi Optimizer Reference Manual*. Available at: http://www.gurobi.com/documentation; accessed 15 May 2020.

Hannachi, A., Woollings, T. and Fraedrich, K. (2012) The North Atlantic jet stream: a look at preferred positions, paths and transitions. *Quarterly Journal of the Royal Meteorological Society*, 138, 862–877. https://doi.org/10.1002/qj.959

Hannachi, A., Straus, D.M., Franzke, C.L.E., Corti, S. and Woollings, T. (2017) Low-frequency nonlinearity and regime behavior in the Northern Hemisphere extratropical atmosphere. *Reviews of Geophysics*, 55(1), 199–234. https://doi.org/10.1002/2015RG000509

Jain, A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jung, T., Palmer, T.N. and Shutts, G.J. (2005) Influence of a stochastic parameterization on the frequency of occurrence of North Pacific weather regimes in the ECMWF model. *Geophysical Research Letters*, 32(23), 1–4. https://doi.org/10.1029/2005GL024248

Kageyama, M., D'Andrea, F., Ramstein, G., Valdes, P.J. and Vautard, R. (1999) Weather regimes in past climate atmospheric general circulation model simulations. *Climate Dynamics*, 15(10), 773–793. https://doi.org/10.1007/s003820050315

Kimoto, M. and Ghil, M. (1993) Multiple flow regimes in the Northern Hemisphere winter. Part II: sectorial regimes and preferred transitions. *Journal of the Atmospheric Sciences*, 50(16), 2645–2673.

Lorenz, E.N. (1963) Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20, 130–141.

Michelangeli, P.-A., Vautard, R. and Legras, B. (1995) Weather regimes: recurrence and quasi-stationarity. *Journal of the Atmospheric Sciences*, 52(8), 1237–1256.

Mo, K.C. (2000) Relationships between low-frequency variability in the Southern Hemisphere and sea surface temperature anomalies. *Journal of the Climate*, 13, 3599–3610. https://doi.org/10.1175/1520-0442(2000)013<3599:RBLFVI>2.0.CO;2

Mo, K.C. and Ghil, M. (1987) Statistics and dynamics of persistent anomalies. *Journal of Atmospheric Sciences*, 44(5), 877–901.

Molteni, F., Tibaldi, S. and Palmer, T.N. (1990) Regimes in the wintertime circulation over Northern Extratropics. I: Observational evidence. *Quarterly Journal of the Royal Meteorological Society*, 116, 31–67. https://doi.org/10.1002/qj.49711649103

O'Kane, T.J., Risbey, J.S., Franzke, C., Horenko, I. and Monselesan, D.P. (2013) Changes in the metastability of the midlatitude Southern Hemisphere circulation and the utility of nonstationary cluster analysis and split-flow blocking indices as diagnostic tools. *Journal of the Atmospheric Sciences*, 70(3), 824–842. https://doi.org/10.1175/JAS-D-12-028.1

Philipp, A., Della-Marta, P.M., Jacobeit, J., Fereday, D.R., Jones, P.D., Moberg, A. and Wanner, H. (2007) Long-term variability of daily North Atlantic–European pressure patterns since 1850 classified by simulated annealing clustering. *Journal of Climate*, 20(16), 4065–4095. https://doi.org/10.1175/JCLI4175.1

Plaut, G. and Simonnet, E. (2001) Large-scale circulation classification, weather regimes, and local climate over France, the Alps and western Europe. *Climate Research*, 17, 303–324. https://doi.org/10.3354/cr017303

Posada, D. and Buckley, T.R. (2004) Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5), 793–808. https://doi.org/10.1080/10635150490522304

Riddle, E.E., Stoner, M.B., Johnson, N.C., L'Heureux, M.L., Collins, D.C. and Feldstein, S.B. (2013) The impact of the MJO on clusters of wintertime circulation anomalies over the North American region. *Climate Dynamics*, 40(7-8), 1749–1766. https://doi.org/10.1007/s00382-012-1493-y

Robertson, A.W. and Ghil, M. (1999) Large-scale weather regimes and local climate over the western United States. *Journal of Climate*, 12(6), 1796–1813. https://doi.org/10.1175/1520-0442(1999)012<1796:LSWRAL>2.0.CO;2

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.

Straus, D.M., Corti, S. and Molteni, F. (2007) Circulation regimes: chaotic variability versus SST-forced predictability. *Journal of Climate*, 20(10), 2251–2272. https://doi.org/10.1175/JCLI4070.1

Straus, D.M., Molteni, F. and Corti, S. (2017). Atmospheric regimes: the link between weather and the large-scale circulation, Chapter 4 in Nonlinear and Stochastic Climate Dynamics, DOI 10.1017/9781316339251.005. Franzke, C.L.E., O'Kane, T.J. (eds), Cambridge University Press, Cambridge, UK.

Vautard, R. (1990) Multiple weather regimes over the North Atlantic: analysis of precursors and successors. *Monthly Weather Review*, 118, 2056–2081.

Volinsky, C.T. and Raftery, A.E. (2000) Bayesian information criterion for censored survival models. *Biometrics*, 56(1), 256–262. https://doi.org/10.1111/j.0006-341X.2000.00256.x