

# Spatio-temporal climate regionalization using a self-organized clustering approach

Mihaela I. Chidean<sup>1</sup> · Antonio J. Caamaño<sup>1</sup>  · Carlos Casanova-Mateo<sup>2</sup> · Julio Ramiro-Bargueño<sup>1</sup> · Sancho Salcedo-Sanz<sup>3</sup>

Received: 2 March 2018 / Accepted: 23 December 2019 / Published online: 13 February 2020  
© Springer-Verlag GmbH Austria, part of Springer Nature 2020

## Abstract

The authors present a novel self-organized climate regionalization (CR) method that obtains a spatial clustering of regions, based on the explained variance of physical measurements in their coverage. This method enables a microscopic characterization of the probabilistic spatial extent of climate regions, using the statistics of the obtained clusters. It also allows for the study of the macroscopic behaviour of climate regions through time by using the dissimilarity among different cluster size probability histograms. The main advantages of the presented method, based on the Second-Order Data-Coupled Clustering (SODCC) algorithm, are that SODCC is robust to the selection of tunable parameters and that it does not require a regular or homogeneous grid to be applied. Moreover, the SODCC method has higher spatial resolution, lower computational complexity, and allows for a more direct physical interpretation of the outputs than other existing CR methods, such as Empirical Orthogonal Function (EOF) or Rotated Empirical Orthogonal Function (REOF). These facts are illustrated with an example of winter wind speed regionalization in the Iberian Peninsula through the period (1979 – 2014). This study also reveals that the North Atlantic Oscillation (NAO) has a high influence over the wind distribution in the Iberian Peninsula in a subset of years in the considered period.

## 1 Introduction

Climate regionalization (CR) is defined as the process of dividing a given area into smaller regions, in such a way that they are somehow homogeneous with respect to a specified climatic variable (Badr et al. 2015). CR is a key point in climate studies, since it allows explaining small-scale climate events in terms of the spatio-temporal mechanisms which produce them. CR has been specifically applied to

palaeo-climatic problems (Knapp et al. 2002), precipitation trends, floods and drought events (Comrie and Glenn 1998; Baeriswyl and Rebetez 1997; Burn 1989), numerical models improvement for climate studies (Argüeso et al. 2011; Regonda et al. 2016), or climate change studies (Önol and Semazzi 2009), among others.

There are a number of well-known linear analysis techniques for obtaining high-quality CR. Empirical Orthogonal Function (EOF) analysis, also known as Principal Component Analysis (PCA), is one of the most standard techniques in climatology with direct application in CR. EOF analysis tries to identify natural spatio-temporal variability of observations (Jolliffe 2002). The idea behind EOF analysis is to identify a set of orthogonal eigenfunctions which accounts for most of the system's total variance (von Storch and Zwiers 1999). Thus, EOF analysis tries to obtain the dominant modes of variability, in turn reducing the data space by only considering those EOFs which cover a large percentage of the total variance. EOF analysis has been intensely used in CR (White et al. 1991; Comrie and Glenn 1998; Baeriswyl and Rebetez 1997). The basic idea is to use EOF or Rotated Empirical Orthogonal Function (REOF) to define and interpret clusters of different climatic variables,

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00704-019-03082-6>) contains supplementary material, which is available to authorized users.

---

✉ Antonio J. Caamaño  
antonio.caamano@urjc.es

- <sup>1</sup> Department of Signal Theory and Communications, Universidad Rey Juan Carlos, Madrid, Spain
- <sup>2</sup> Department of Civil Engineering: Construction, Infrastructures and Transports, Universidad Politécnica de Madrid, Madrid, Spain
- <sup>3</sup> Department of Signal Processing and Communications, Universidad de Alcalá, Madrid, Spain

at different spatial resolutions of the grid. However, different studies have pointed out some issues in the use of EOF analysis (Dommenges and Latif 2002). For example, in Kim et al. (2015), it is shown that EOF analysis usually considers stationarity in the covariance of the data, when some of the most important climatic variables show cyclic components. This could lead to misleading results, and the authors propose the use of cyclostationary EOF analysis in these cases. Another important problem when applying EOF analysis is due to the orthogonality constraint of the procedure. It is shown that unphysical modes may appear due to this constraint (Lian and Chen 2012), producing results difficult to explain. Different works have shown that this issue may be alleviated by using REOF analysis, a method which has been shown to perform much better than the classical EOF approach in many particular cases. In fact, previous studies have shown how REOF analysis is able to avoid unphysical modes of EOFs while keeping the important and robust spatio-temporal patterns of the data (Richman 1986). On the other hand, REOF analysis has its own set of problems: (1) the database size necessary to obtain a good calculation of the REOFs, (2) the selection of the number of modes involved in their calculation and last, but not least, (3) the selection of the rotation criteria (Jolliffe 2002, p. 271) all of which make the results highly dependent on subjective decisions.

Non-linear clustering techniques have also been applied to CR problems. These include methods such as hierarchical clustering (Ward 1963; Unal et al. 2003; Badr et al. 2014, 2015; Nojarov 2017),  $k$ -means (Cassou et al. 2004; Carvalho et al. 2016; Zhang et al. 2016), multivariate statistical techniques (Shahriar et al. 2015), or fuzzy clustering (Sarma and Hazarika 2014; Irwin 2015). The main advantages of these methods are that they allow a very high spatial resolution of the grid in climate problems, with a tunable similarity criterion (different from the variance as in EOFs). However, there are some issues related to clustering techniques in CR, such as a more difficult interpretation that can lead to different conclusions due to the use of different similarity criteria or schemes (Cassou et al. 2004). Also, the need for tuning some specific parameters of the methods (for example the number  $k$  of clusters in the  $k$ -means approach) might be an issue.

In this paper, a novel self-organized heuristic clustering technique is proposed for CR problems. Specifically, the Second-Order Data-Coupled Clustering (SODCC) algorithm is used (Chidean et al. 2015a), a self-organized clustering approach that uses statistical characteristics of the measured data to geographically group similar nodes. To this end, the proposed algorithm uses signal subspace dimension to determine the minimum amount of linearly independent components in each cluster, i.e. the number of eigenvalues that explain most of the variance in the

data. It approaches the clustering problem from a bottom-up perspective, in which neighbouring nodes are initially grouped and, solely based on the signal subspace dimension of the covariance matrix, it decides whether to fuse to other cluster or not. This procedure is repeated until there is no free node left in the system. This procedure is robust to the selection of specific tuning parameters, apart from those that define the statistical distribution of eigenvalues in covariance matrices.

The proposed SODCC algorithm builds a bridge between linear (EOFs and REOFs) and non-linear analysis techniques (hierarchical,  $k$ -means... clustering) as it uses characteristics from the latter to build spatial structures comparable with the former. Moreover, by using SODCC, it is also possible to analyze the temporal and spatial structures of the dataset, with a higher spatial resolution of the grid. This analysis permits a direct comparison with climate patterns such as the North Atlantic Oscillation (NAO).

The rest of the paper is structured as follows: Section 2 includes a detailed description of the SODCC algorithm and its analysis from a theoretical point of view. Section 3 covers the case study considered in this work, where the regionalization and spatio-temporal analysis of wind speed data in the Iberian Peninsula is carried out using the SODCC algorithm and robustness to the selection of tunable parameters is demonstrated. Finally, Section 4 closes the article by giving some concluding remarks.

## 2 Methods

In this section, the clustering algorithm proposed in this work is detailed for CR. First, the system model is described, providing mathematical definitions for the network and dataset. Next, the SODCC algorithm is described and the procedure to analyze climatological data using the aforementioned algorithm. In the following, the operation of SODCC is analyzed from a theoretical point of view.

### 2.1 System model

Consider a set of  $N$  geographical locations determined by their corresponding latitude and longitude coordinates to be the measuring stations (or grid nodes) for a given climate variable, i.e. wind speed. Let  $T_s$  be the uniform sampling interval, namely the time interval between two consecutive data measurements. Each measuring station can be modeled as an element of a network with  $N$  nodes and  $C$  connections among them.

Let  $M_T$  be the total number of data measurements per node and  $\mathbf{x}_m \in \mathbb{R}^N$ , with  $\mathbf{x}_m = [x_m(1), x_m(2), \dots, x_m(N)]^T$ ,

be the vector<sup>1</sup> of data measurements for all nodes in a specific time instant  $m = 1, 2, \dots, M_T$ . By assembling the  $M_T$  data vectors in a data matrix, we obtain  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_{M_T}]$ , which is the dataset measured by the entire network. The covariance matrix of a data matrix  $\mathbf{X}$  is defined as  $\Sigma = \mathbf{M}_X \mathbf{M}_X^T / M_T$ , being  $\mathbf{M}_X$  the centered data matrix.

With this system model, the spatio-temporal correlations of the climate variable can be analyzed considering the complete set of  $N$  nodes and using traditional climate analysis such as EOFs or REOFs (Jolliffe 2002). However, different approaches can be taken into account, such as the recently proposed clustering algorithm SODCC that we consider in this work. SODCC organizes the nodes in groups in terms of the statistics of the measured data (Chidean et al. 2015a). In the following, the details of the SODCC operation and the procedure to apply it to climate variables are further explained.

## 2.2 Clustering algorithm: SODCC

The SODCC algorithm has been previously proposed and described in detail for wireless sensor network applications (Chidean et al. 2015a).<sup>2</sup> However, previous works have shown that it is also suitable for climate data analysis, e.g. in Chidean et al. (2015b), where the structure of temperature field in Europe from more than 100 measuring stations was studied using this approach. In this section, we focus on the general description of the SODCC and its operation from the climate application point of view.

The SODCC algorithm organizes the nodes of a network in logical groups, also known as clusters, following a data-driven criterion, i.e. uses data statistics to decide. The outcome of SODCC is a set of clusters, each one containing  $N_i$  nodes. Note that all  $N$  nodes of the network must belong to one and only one cluster and that each cluster may have a different cluster size  $N_i$ . The final cluster size is determined by the minimum amount of nodes that explain a minimum of 90% of the data variance of that region, by means of the calculation of the corresponding eigenvalues. For example, small cluster sizes appear in regions with high data correlation and large cluster sizes appear in regions with low data correlation.

The fact that it is possible to explain a minimum amount of data variance for all the obtained clusters leads to a more relevant characteristic of SODCC: the phase transition of the data matrices of each cluster is achieved, i.e. it is possible to properly perform the eigendecomposition

within each cluster. For details about the phase transition of correlation matrices with a finite number of samples, refer to Appendix A.

The SODCC algorithm follows a bottom-up approximation, where the nodes start belonging to small clusters and, by means of cluster fusion processes, they end up belonging to larger clusters. The algorithm has two phases that are different in both objective and operation:

1. Cluster initialization—random initialization of small clusters that act as seeds for the final set of clusters.
2. Cluster growing—all non-final (or non-stable) clusters take part in cluster fusion processes until a stopping criterion is reached and the final set of clusters is determined.

### 2.2.1 Cluster initialization

The objective of this phase is the random initialization of the set of clusters, so that the second phase can use them, perform its operation, and output the final set of clusters.

Given the  $N$  nodes that form the network, there are three possible and exclusive status for any of them:

1. Cluster initialization (CH)—this is the first node of a given cluster.
2. Normal node—the node belongs to a cluster and it is not the CH of that cluster.
3. Role-free node—the node does not belong to any cluster in any of the previous ways.

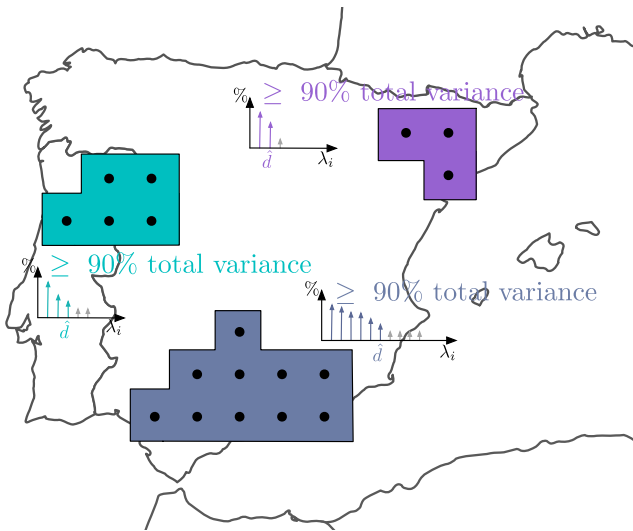
The first phase of the SODCC algorithm also requires two preset parameters in order to form the first set of clusters. Parameter  $P$  indicates the probability that a given role-free node becomes CH and parameter  $N^{1st}$  indicates the maximum initial cluster size. These parameters are set in order to obtain clusters that are small enough to act as “seeds” of the final clusters and large enough to not interfere with the operation of the second phase.

The results presented in this work are obtained considering  $P = 0.35$  and  $N^{1st} = 3$ , that were also considered in previous works (Chidean et al. 2015a, b). However, Section 3 includes a brief discussion regarding method robustness with respect to these parameters.

The operation of this first phase is summarized in Algorithm 1. In short, role-free nodes independently decide to become CH according to  $P$ . Next, the new CHs search for role-free nodes in their neighbourhood and, if it is possible, include up to  $N^{1st} - 1$  role-free nodes into their own cluster. “Neighbourhood” allows multiple definitions in a network; however, in this work can be understood as nodes located within a short distance of a given node, e.g. the eight nearest nodes in a squared regular grid as the one sketched in Fig. 1.

<sup>1</sup>A vector is a matrix consisting of a single column of elements and the symbol  $(\cdot)^T$  indicates the transpose operation.

<sup>2</sup>Note that the algorithm source code can be found at <https://github.com/MihChi/SODCC>.



**Fig. 1** Example of clusters obtained by SODCC. The embedded graphs schematize the sorted eigenvalues  $\lambda_i$ , highlighting the  $\hat{d}$  eigenvalues that explain a minimum of 90% of the variance in each cluster

---

#### Algorithm 1 Cluster initialization.

---

- 1: Role-free nodes become CH according to probability  $P$
  - 2: **if** the node is a new CH **then**
  - 3:   Search role-free nodes in the neighborhood
  - 4:   Include up to  $N^{1st} - 1$  role-free nodes into own cluster
  - 5:   Role-free nodes that belong to the new cluster turn into “normal nodes”
  - 6: **end if**
- 

Due to chance, it is possible that there are still some role-free nodes after this algorithm loop. For this case, the SODCC algorithm allows the repetition of this cluster initialization phase multiple times (up to  $N^{rep}$  times). Eventually, after a given number of repetitions, all remaining role-free nodes turn into CH and, unable to find available role-free nodes, their cluster is formed by a single node.

This random cluster initialization may affect the interpretation of results. Therefore, in order to cancel its effect, the SODCC algorithm has to be applied multiple times and the results have to be analyzed in a probabilistic manner.

### 2.2.2 Cluster growing

The objective of this phase is to determine the final set of clusters, by means of cluster fusions processes until a stopping criterion is achieved. The SODCC stopping criterion is independently evaluated for each non-final cluster after each cluster fusion process and uses the covariance matrix  $\Sigma_i \in \mathbb{R}^{N_i \times N_i}$  of the data measured by the  $N_i$  nodes that form the  $i$ th cluster. To ensure the achievement of the stopping criterion, SODCC does not

consider cluster division processes. The detailed operation of this second phase is summarized in Algorithm 2. At the beginning, for the  $i$ th cluster, the minimum amount of data per node  $M_i$  to achieve the phase transition is calculated, that is, according to matrix perturbation theory (Nadler 2008), the minimum number of entries per node needed for the successful construction of a covariance matrix of a given size. Next, the covariance matrix  $\Sigma_i$  is calculated, and its eigenvalues are estimated. Following, the number of eigenvalues  $\hat{d}$  (sorted in decreasing order) that account for a minimum of 90% of explained variance of the dataset measured by all the nodes forming  $i$ th cluster is determined. In this work, the value of the  $\hat{d}$  is calculated by means of the Fast Subspace Decomposition (FSD) algorithm (Xu and Kailath 1994). For details about the FSD algorithm, please refer to Appendix B.

---

#### Algorithm 2 Cluster growing.

---

##### for each cluster:

- 1: Calculate  $M_i$ , which is the amount of data values required to achieve phase transition for this cluster and to be able to properly calculate the eigenvalues
  - 2: Calculate  $\Sigma_i$  using  $M_i$  data values per node and estimate the eigenvalues
  - 3: Determine  $\hat{d}$  using the FSD algorithm
  - 4: **if**  $\hat{d} \geq N_i$  **then**
  - 5:   Fuse to a neighbor cluster
  - 6:   Change state of the node from CH to normal node that belongs to the new cluster
  - 7: **end if**
- 

Finally, based on the values of  $N_i$  and  $\hat{d}$ , the stopping criterion is evaluated. More specifically, the  $i$ th cluster fulfills the stopping criterion if  $\hat{d} < N_i$  and the SODCC algorithm ends its operation for this cluster. Otherwise, this cluster must fuse to a neighbouring cluster (minimum euclidean distance) in order to grow in size and to span a larger spatial region and fulfill the stopping criterion in future algorithm iterations. Again, in the present work, cluster fusion is based on distance between nodes and a cluster can only fuse to those clusters located in the neighbourhood of its nodes.

The relation between the stopping criterion of the SODCC and the final cluster sizes has been shown in Chidean et al. (2015a). In the following, we summarize the main idea, also represented in Fig. 1. Given the final set of clusters, the small clusters are expected to appear in areas where there exists a high data correlation, as only data from few nodes are needed to determine the  $\hat{d}$  eigenvalues with 90% of the explained variance. For example, in Fig. 1, the cluster shaded in purple is final as the sum of the explained variance of the first  $\hat{d} = 2$  eigenvalues is greater than 90%, indicating that the  $N_i = 3$  nodes of the cluster have

high data correlation. On the other hand, large clusters are expected to appear in areas where the cross-correlation of the data among nodes is low. Therefore, the SODCC algorithm ensures that closely correlated data series are clustered together (more than 90% of the variance can be explained with  $\hat{d} < N_i$  eigenvalues) and that the cluster configuration reflects the underlying data spatio-temporal correlations.

### 2.3 Theoretical analysis of SODCC

In the previous subsection, we have described an algorithm that drives the clustering in the different stages. The cluster fusion processes can be directly assimilated with diffusion-limited aggregation processes (in terms of the cluster border interaction) for irreversible clusters (i.e. once they form, they do not diminish in size). This assimilation provides us with the means to propose a theoretical analysis of the cluster size probability (CSP) theoretical distribution obtained for the final set of clusters.

An initial formulation for this distribution can be proposed following a dynamic scaling function (Vicsek and Family 1984):

$$n(t, N_i) \sim t^{-w} N_i^{-\tau} f\left(\frac{N_i}{t^z}\right) \tag{1}$$

where  $t$  is the time throughout the clustering process until the final distribution is achieved and  $N_i$  is the number of nodes that form the  $i$ th cluster. The importance of time in the present theoretical formulation lies on its scaling dependence, independent from that of the cluster size. Equation 1 is formed by three terms that describe the cluster fusion processes through a power-law (with  $w > 0$ ), the final CSP theoretical distribution through a different power-law (with  $0 < \tau < 2$ ), and the characteristic cluster size (with  $z > 1$ ), respectively. Function  $f(x)$  has power-law behaviour for small  $x$ , i.e.  $x^{2-\tau}$  for  $x \ll 1$  and  $f(x) \ll 1$  for  $x \gg 1$ . In the context of climatology, the scaling functions for cluster size distribution are related to stable spatio-temporal characteristics that exponentially relate the spatial extent of physical phenomena to their temporal stability. In the present case, the cluster second-order statistics features define both relations. In climatology, the extensive study of the EOF and their is the closest relative to the present approach. Generally, EOF is used to characterize global sets of data. In the present study, a possible interpretation would be the local determination of EOF and their temporal stability.

The CSP theoretical distribution for the final set of clusters (the equilibrium condition of the fusion processes) can be approximated by

$$n(t, N_i) \sim t^{\tau-2} N_i^{-\tau} \tag{2}$$

Equation 2 provides us with a power-law dependence on the cluster size from the smallest cluster size possible

( $N_i = 3$  in our case) to infinity. A normalization condition is necessary in order to define a proper probability distribution, i.e.

$$\sum_{N_i=3}^N n(t, N_i) = 1 \tag{3}$$

The time dynamics of the formation of the final set of clusters (governed by the second phase of the SODCC) are quantized at the characteristic times, when the phase transition condition for the covariance matrix is fulfilled (see Appendix A) and the FSD statistic is computed. Then, a fraction of the clusters with size  $N_i$  will grow in size, as they are fused to other clusters.

In a CR problem, where spatial correlations are studied, the minimum meaningful cluster size of the final set of clusters is  $N_i^{\min} = 3$ , where the SODCC determines two signal eigenvalues ( $\hat{d}_{\min} = 2$ ) that account for a minimum of 90% of explained variance and one noise eigenvalue. Therefore, the CSP theoretical distribution includes an exponential decay of the cluster size from  $N_i = 3$  onwards.

On the other hand, the theoretical maximum resolvable signal subspace dimension  $\hat{d}_{\max}$  is limited by the temporal size of the data  $M_T$ . For the final set of clusters, the condition for phase transition holds for all clusters. As the minimum amount of data required to properly calculate the data covariance matrix,  $M_i$  can be set to  $M_i = 4 \times N_i^{\max}$  and as at least one noise eigenvalue is determined, the maximum value for  $\hat{d}_{\max}$  is  $N_i^{\max} - 1$  (Chidean et al. 2015a). Clusters where the signal subspace dimension is  $\hat{d}_{\max}$  (the maximum obtained value considering the complete set of grid points and all time instants) will be the largest clusters in the grid with FSD convergence. In statistical terms, these clusters will be the largest connected structures in the grid that are bound with the 90% explained variance. The larger clusters (where  $N_i > \hat{d}_{\max} + 2$ ) are not bestowed with such tight relation and, therefore, are bound to be ephemeral through the different realizations of the SODCC algorithm, thus changing their shape, size, and even their location. That is the main reason for their exponentially decaying sizes in the CSP distribution.

Thus, the CSP theoretical distribution includes an exponential decay between cluster sizes  $\hat{d}_{\min} + 1$  and  $\hat{d}_{\max}$ , a sharp probability rise in  $N_i = \hat{d}_{\max} + 1$ , followed by an exponential decay with the same exponent. These behaviours are universal in clustering processes with fusion (Vicsek and Family 1984).

Considering the above arguments, we formulate the CSP theoretical distribution obtained for the final set of clusters as:

$$n(t, N_i) = \kappa \times t^{-w} \left( g(N_i, \hat{d}_{\min}) + g(N_i, \hat{d}_{\max}) \right) \tag{4}$$

where the modulation function

$$g(x, x_0) = \theta(x - x_0)(x - x_0)^{-\tau} \quad (5)$$

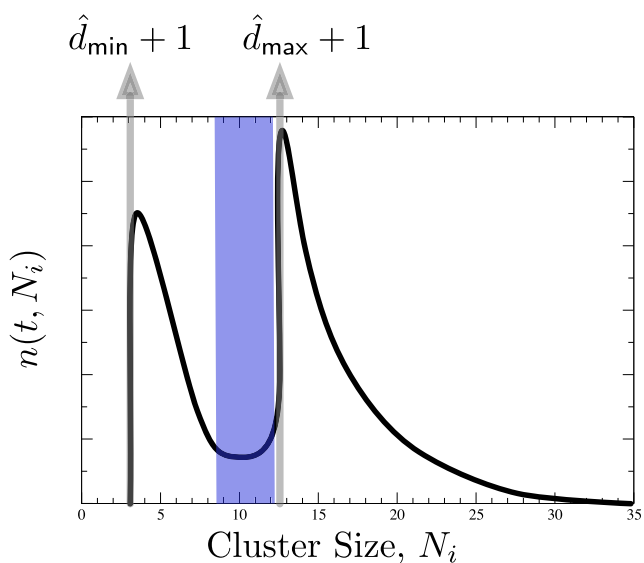
$\theta(x)$  is the Heaviside step function ( $\theta(x - x_0)$  denotes the  $\theta(x)$  function shifted  $x_0$  units in the abscissa axis) and

$$\kappa = \frac{\tau - 1}{(N - \hat{d}_{\min})^{1-\tau} + (N - \hat{d}_{\max})^{1-\tau} - 2} \quad (6)$$

is the normalization constant *s.t.* Equation 3 is satisfied.

In Fig. 2, we can see a generic CSP in which  $\hat{d}_{\min} + 1$  and  $\hat{d}_{\max} + 1$  are the most relevant features of the distribution. Due to the dynamics of cluster aggregation, the clusters with sizes in the trough of the  $N_i \in [\hat{d}_{\min} + 1, \hat{d}_{\max} + 1]$  interval are those in which their signal eigenvalues are relevant to the statistics of the data under study. Clusters with sizes beyond  $N_i = \hat{d}_{\max} + 1$  are the result of cluster aggregation of large, unstructured clusters with unresolved signal eigenvalues according to the FSD statistic. Thus, let the “focus zone” be the cluster size range where the CSP is able to resolve the climate regions (e.g. the shaded area in Fig. 2).

In this section, we have shown that the resulting CSPs do not have an arbitrary shape. On the contrary, they result from the cluster fusion process. The fitting of the experimental CSPs to their theoretical functions by means of the scaling parameters  $\tau$ ,  $\kappa$ , and  $w$  may reveal further details of the physical processes in a single month/year. These scaling parameters determine the statistics of the cluster sizes. Therefore, examining their variations through the years may lead to revealing the physical nature of the interactions among clusters. This task is out of the scope of the present work but it will be undertaken in the future.



**Fig. 2** Plot of a generic CSP  $n(t, N_i)$ , where  $\hat{d}_{\min} + 1$  and  $\hat{d}_{\max} + 1$  are indicated with grey arrows and the interval of interest of cluster size probability (focus zone) is shown in blue

### 3 Case study: wind resource in the Iberian Peninsula

In this section, we analyze the wind resource in the Iberian Peninsula in the framework of CR, in order to show the performance of the proposed SODCC. Before this, we discuss some previous works dealing with spatio-temporal analysis of wind speed and different approaches for its regionalization.

Spatio-temporal analysis of wind speed is significant for a number of problems related to climate, as indicator of circulation changes, and also for renewable energy resource analysis, among others. Very recent works deal with wind speed regionalization or spatio-temporal analysis of wind speed using reanalysis data. In Yu et al. (2015), the temporal variability of wind speed is studied in the USA, based on the climate forecast system reanalysis data. EOF analysis is used to find connections to NAO and ENSO patterns, which seem to control part of the wind speed variability across the USA. Another related work is Troccoli et al. (2012), which studies long-term variability of wind speed trends over Australia. In this case, wind speed observations and reanalysis data are used to obtain the wind speed trends.

Also, recent works on wind speed regionalization in the Iberian Peninsula have been published. In Azorin-Molina et al. (2014), the analysis of homogenization of wind speed in the Iberian Peninsula is considered. The study analyzes wind speed trends recorded at 67 land-based stations across Spain and Portugal for the period 1961–2011, finding a slight downward trend in wind speed. The paper also analyzes the possible impact of three atmospheric indices (NAO, Mediterranean oscillation, and western Mediterranean oscillation) and the role played by the urbanization growth in the observed decline of wind speed. Finally, in Santos-Alamillos et al. (2016), the effect of spatio-temporal balancing of wind resource in wind farms in Spain is studied. A regionalization approach based on EOF analysis is carried out in this case.

With this discussion on previous related works in mind, we start our analysis of wind speed in the Iberian Peninsula, by considering wind speed at 10-m data obtained from the ERA-Interim reanalysis dataset. We apply the SODCC clustering algorithm, from both spatial and temporal points of view, including an analysis of the results in the framework of climate indices such as the NAO. Next, we compare our findings with the results obtained by analyzing the dataset by means of classical data analysis in climatology, i.e. EOFs and REOFs. With this dual analysis, we are able to show the main benefits of the CR method proposed in this work. Finally, we demonstrate the robustness of the SODCC algorithm to initialization parameters.

### 3.1 Dataset

In this case study, we use wind speed at 10-m data obtained from the ERA–Interim reanalysis dataset (Dee et al. 2011). The geographical area is bounded by parallels 35°N and 44°N and by meridians 10°W and 6°E, demarcating the Iberian Peninsula and the Balearic Islands. The grid is set to 0.5° × 0.5°, making a total of  $N = 627$  nodes in the network, placed in a regular grid as shown in Fig. 3. Given this node location, the distance resolution is approximately 56 km. For this case study, we consider the wind speed data for January, in representation for the most windy season in the area of interest (Zishka and Smith 1980; Branick 1997). We consider the 1979–2014 time interval with a temporal resolution of  $T_s = 6$  h, making a total of  $M_{\text{month}} = 31 \times 4 = 124$  data values per month. As the analysis that we consider in this work is performed independently per year, we define the following set of data ( $T_j$ ) to clearly specify the time period under analysis:

$$T_j = \{\text{first time instant of } j \mid j = 1979, \dots, 2014\} \quad (7)$$

Given the regular grid of the nodes, we should expect edge effects in the border of the studied area. However, as it will be commented on Section 3.b.2, these effects are not apparent in the results.

### 3.2 Data analysis with SODCC

The signal processing method for CR presented in this work is based on the SODCC algorithm. The idea behind this method is that, for each realization, the statistics calculated for the set of clusters obtained reveal both the spatial and temporal modes of the climate variable. All of the software used in the present work for the SODCC analysis is original and it has been developed by some of the authors in Matlab

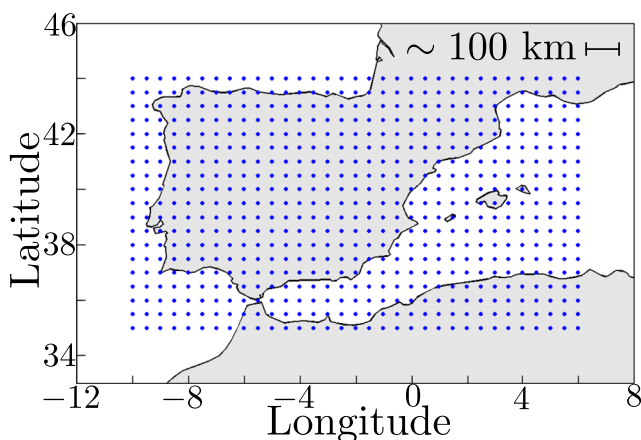


Fig. 3 Geographical location of the  $N = 627$  nodes (blue dots) of the network used in the present case study

and it is available on the following link (<https://github.com/MihChi/SODCC>).

The considered dataset includes January wind speed data between 1979 and 2014. Data of each particular year is independently analyzed, giving a total of 36 starting points for the SODCC algorithm. This analysis allows the assessment of the temporal evolution of the spatial structures of the data.

For each possible value of  $T_j \in T_J$ , in this case study, we perform 5000 independent simulations, applying the SODCC algorithm to the wind speed dataset of  $N = 627$  nodes. Such experiments allow the analysis of the results with sufficient statistical representativity that are not determined by specific realizations of the random initialization. As previously mentioned, the output of each independent SODCC realization is a set of clusters, formed by nearby nodes, where it is possible to estimate the eigenvalues that explain a minimum of 90% of the data variance.

Result analysis can be performed at node level and at set of clusters level for each  $T_j$ . In the first case, by analyzing how likely is each node of the grid to belong to a cluster of a given size, the spatial correlation of the data at multiple spatial scales is revealed. Moreover, changes over the different years will show both the temporal evolution of the data statistics and their spatial extent. The following section is devoted to the spatial analysis of the results by means of the node to cluster size probability (NCSP) histograms. In the second case, studying the complete set of clusters allows a more straightforward analysis of the temporal evolution of wind resources in the Iberian Peninsula. Section 3.2.2 includes the temporal analysis by means of the CSP histograms normalized to have unit area as approximations of the corresponding CSP theoretical distributions that were previously introduced in Section 2.2.3.

#### 3.2.1 Spatial analysis

The spatial analysis of the wind speed dataset can be performed by means of the node to cluster size probability (NCSP) histograms normalized to unit area, namely the probability of a given node to be associated with a cluster of a certain size  $N_i$  for the starting point of the simulation  $T_j \in T_J$ . These NCSP histograms can be displayed over a map, similar to the EOF and REOFs representations. We can represent one map for each combination of  $N_i$  and  $T_j$  enabling the analysis of the spatial extent of the data correlations with great detail. Moreover, this representation also allows the simultaneous analysis of various spatial extents by using multiple values of  $N_i$  for each value of  $T_j$  (e.g.  $N_i \in \{5, 6\}$  or  $N_i \in \text{focus zone}$ ).

The decision criterion of the SODCC organizes the nodes in terms of the explained variance, i.e. in a cluster of  $N_i$

a minimum of 90% of the data variance is explained by  $\hat{d}$  eigenvalues, being  $\hat{d} < N_i$ . Then, geographical areas where the measured data has higher spatial correlation will result in higher probability to form smaller clusters and lower probability to form larger clusters, while areas where the spatial correlation is lower the behaviour is opposite. In the present case study, from Figs. 4 and 5, we can identify the extent of the spatial correlation for each geographical areas. The complete set of figures showing the NCSP histograms for all possible cluster sizes ( $N_i \in [N_i^{\min}, N_i^{\max}]$ ) and for all starting points of the simulation ( $T_j \in T_J$ ) is provided as electronic supplemental material (file “2\_NCSP.pdf”). All these additional figures reaffirm our conclusions regarding the spatial correlation of the wind speed data.

Figures 4 and 5 show the NCSP for  $N_i = 7$  and  $N_i = 18$ , respectively, for a subset of  $T_j \in T_J$ , i.e. each figure indicates the probability of each node to belong to a cluster if size  $N_i$  in a given year. The representation of the NCSP over a map and showing the probability value using colors allows the spatial analysis of the data correlations: nodes located in areas with similar colors have similar data statistics. The selected cluster sizes are representative values from both inside and outside the CSP focus zone and will allow the analysis at different spatial scales, as the cluster size is directly related with the spatial extent. Both figures use identical color range in order to show the probability values and allow direct comparison. From these figures we can clearly identify two different behaviours regarding the spatial extent of the data correlations.

Figures 4 and 5 show the NCSP for  $N_i = 7$  and  $N_i = 18$ , respectively, for a subset of  $T_j \in T_J$ , i.e. each figure indicates the probability of each node to belong to a cluster if size  $N_i$  in a given year. The representation of the NCSP over a map and showing the probability value using colors allow the spatial analysis of the data correlations: nodes located in areas with similar colors have similar data statistics. The selected cluster sizes are representative values from both inside and outside the CSP focus zone and will allow the analysis at different spatial scales, as the cluster size is directly related with the spatial extent. Both figures use identical color range in order to show the probability values and allow direct comparison. From these figures, we can clearly identify two different behaviours regarding the spatial extent of the data correlations.

The fact that multiple nearby nodes have a higher probability to belong to a cluster of a given size (e.g. blue and green areas in Figs. 4 and 5) indicates that the statistics of the wind speed data of those locations are similar (from an explained variance point of view). However, while in Fig. 4, the examined clusters reveal the short range inter-node relations (clusters with size 7, with an isotropic distribution refer to an area of approximate radius of 87 km), Fig. 5 refers to longer range inter-node relations (areas with

approximate radius of 137 km). Even though the particular probabilities for both cluster sizes may be similar, the spatial spread of both is obviously different. This fact is also apparent in these figures due to contrast differences: lower for larger cluster sizes, higher for small cluster sizes.

A consistent feature can be observed in both figures. If a given node has a high probability of belonging to a small size, then the same node has lower probability of being assigned to a larger cluster, as the NCSP is calculated per node and its sum has to be 1. However, individual clusters are not necessarily isotropic, i.e. radially symmetric. As a matter of fact, individual clusters may adopt elongated or even non-convex shape. And self-averaging over multiple cluster sets makes indistinct the variations on the clusters frontiers.

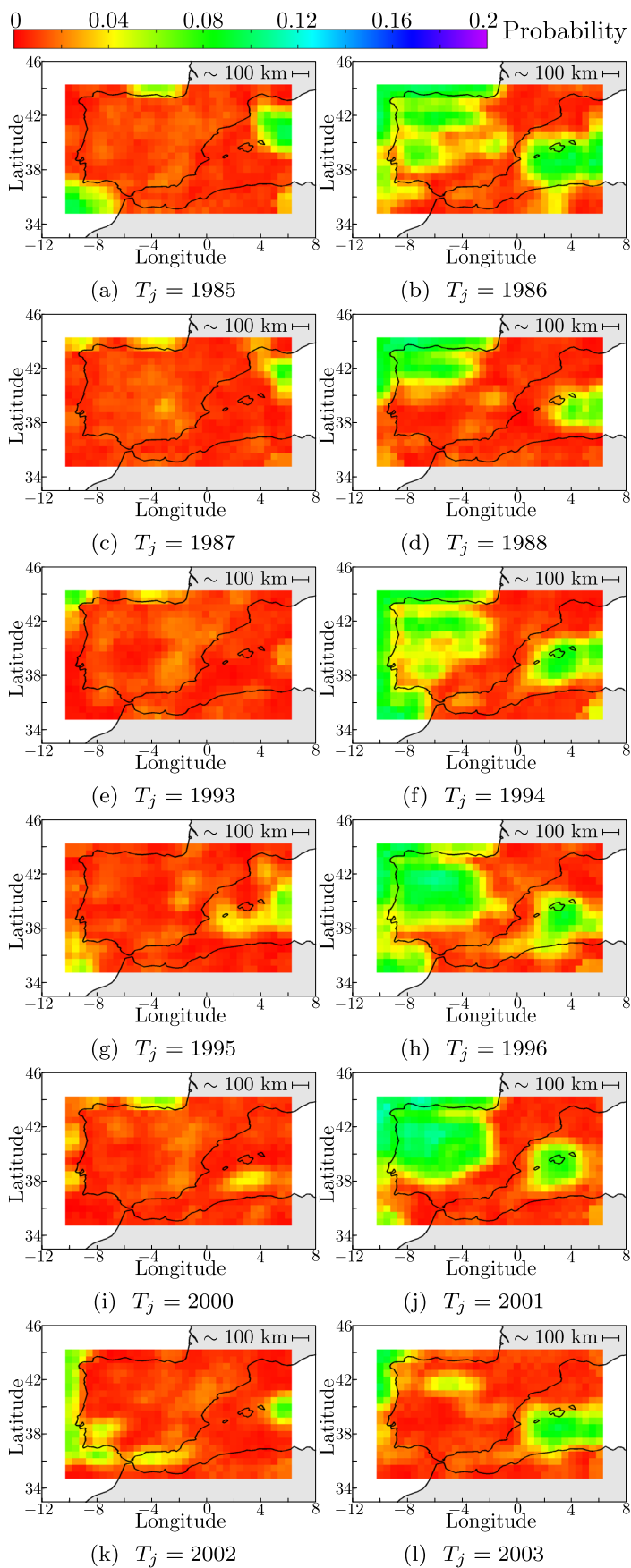
The particular spatial patterns that appear in Figs. 4 and 5 are revealing a shorter range explained variance in the area around the Balearic Islands for the subfigures plotted in the rightmost columns. As for the physical interpretation of these results, very similar variations in wind speed through time in neighbouring nodes are revealed as blue/green patches. Thus, from subfigures (a), (c), (e), (g), and (j) of both figures, we can clearly define short-range similarly varying areas. These homogeneously varying areas and their relation to macroscopic climate features will be addressed later in the temporal analysis of the SODCC.

One of the main benefits of the present methodology relies on the fact that cluster sizes are directly related to the explained variance of the data from the encompassed nodes. Smaller cluster sizes can be directly related to fewer eigenvalues being directly responsible for 90% of the variance of the data in the cluster (by construction). On the other hand, in larger clusters, a larger number of eigenvalues is needed to explain the same amount of variance of the data over a larger spatial extension. Thus, the fewer eigenvalues needed (small clusters), the higher the variance endowed to each of those eigenvalues than with respect to the set in larger clusters. In the present case, the larger variance in the smaller clusters (in wind speed) can be related directly to a higher amount of energy concentrated in smaller areas. This fact can be directly related to wind energy production in those areas (Chidean et al. 2018). Furthermore, a relative increase (decrease) of the population of smaller clusters in successive years would indicate the absence (presence) of dominant components at a larger scale. This facilitates not only the quantification of the correlation distance of turbulence in wind speed, but also their dominant locations at any given time.

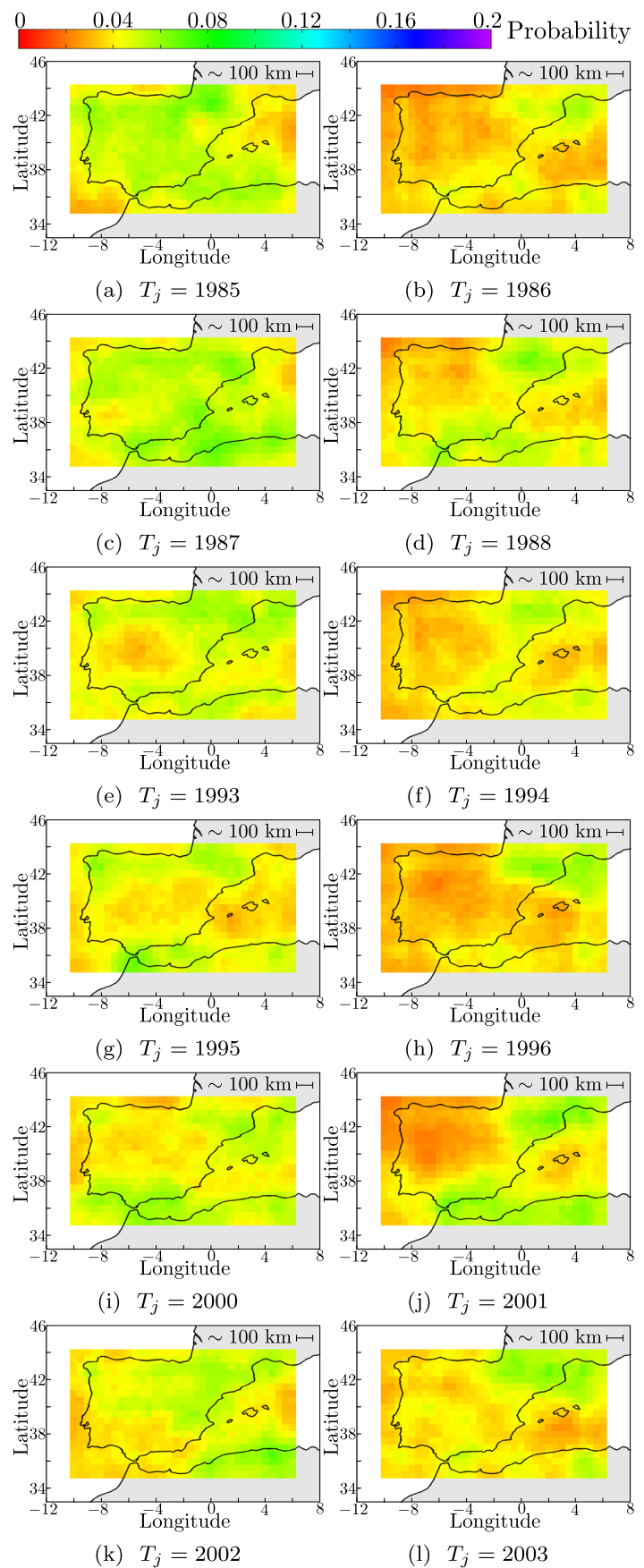
At this point, one further comment is required with respect to the edge effects. Although we expected edge effects to be apparent in the NCSP histograms, we did not find evidence of such. We performed spatial sensitivity analysis considering both the center and the borders of



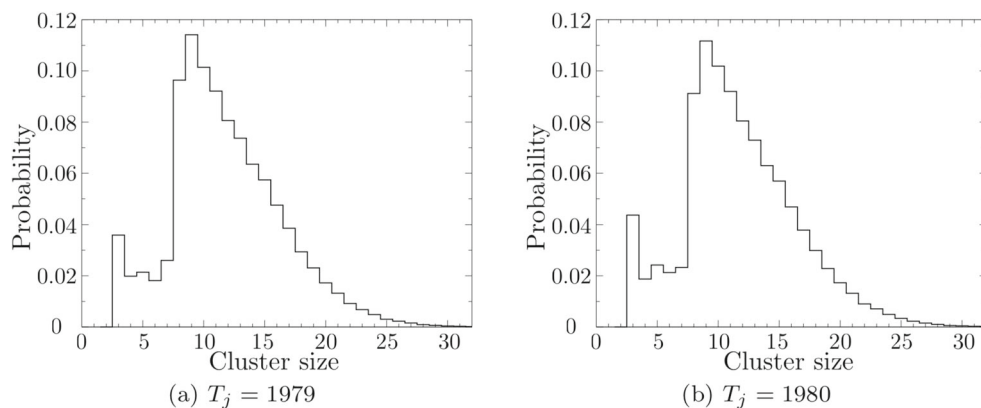
**Fig. 4** NCSP values for  $N_i = 7$  and multiple values of  $T_j$  represented over the map, showing the probability of a given node to belong to a cluster of size  $N_i = 7$ . This representation allows the spatial analysis of the data correlations



**Fig. 5** NCSP values for  $N_i = 18$  and multiple values of  $T_j$  represented over the map, showing the probability of a given node to belong to a cluster of size  $N_i = 18$ . This representation allows the spatial analysis of the data correlations



**Fig. 6** CSP histograms resulting from the SODCC clustering of the wind speed data from the  $N = 627$  nodes for  $T_j = 1979$  and  $T_j = 1980$



the grid, obtaining, up to experimental measurement error, virtually the same results. Thus, this type of analysis for large grids does not exhibit edge effects.

In conclusion, we can state that the CR based on the SODCC algorithm allows for the identification of regions of interest where the climatic data share similar statistics or belong to the same physical mode.

### 3.2.2 Temporal analysis

The temporal analysis of the results can be performed by means of the NCSP. However, due to the many possible cluster sizes (from cluster sizes 3 to 35 for each year under study), an alternative method based on the theoretical CSP distribution presented earlier is considered. The proposed method is based on the analysis of CSP histograms for each  $T_j$  and its inter-annual variations.

The CSP histogram reveals overall information about the behaviour of the results for each  $T_j$ . For example, Fig. 6 shows the CSP for  $T_j = 1979$  and  $T_j = 1980$ . As previously stated, these histograms reveal a bi-modal shape, with two exponential decays starting in  $\hat{d}_{\min} + 1$  and  $\hat{d}_{\max} + 1$ , respectively. The complete set of figures showing the CSP histograms is provided as electronic supplemental material (file “1\_CSP.pdf”).

Again, the temporal analysis of the results using the complete set of figures like the ones represented in Fig. 6 might be arduous and relevant details might be omitted. Thus, Fig. 7 combines all the obtained CSPs in a 3D representation, where the color gradient indicates the bar height (the probability of each cluster size) and the third axis is time. The orientation of the figure is chosen to better visualize the most representative differences over the time.

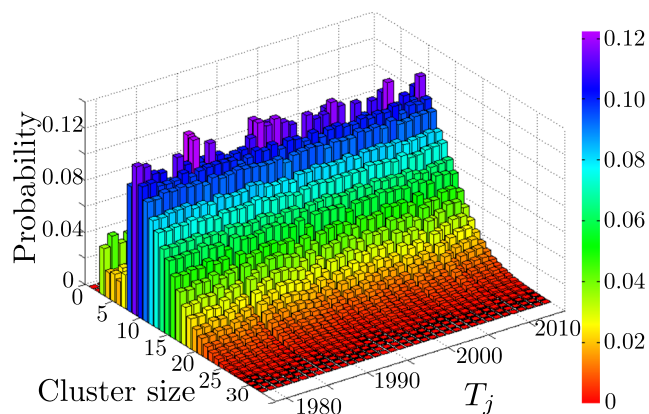
In this representation, the variations in time in the different CSPs are now apparent and are revealed as a series of hills and valleys (in time). For example, for a cluster size of 15 nodes, we can see a valley for  $T_j = 1981$  and a hill for  $T_j = 1991$ . However, some datasets might lead to less clear results and additional calculations have to be made

in order to obtain similar conclusions. Possibly one might criticize that visual analysis of the figures is not enough to offer convincing conclusions. Therefore, the next step in this methodology is to quantify the differences between the different CSPs. A canonical measure of the “distance” between any two probability mass functions is the Kullback-Leibler divergence (KLD) (Kullback and Leibler 1951). Thus, to quantify the variations in time between any two of the CSP histograms  $P_{\text{cluster}}(T_j)$  and  $P_{\text{cluster}}(T_k)$ , we examine their KLD computed as

$$D_{\text{KL}}(T_j, T_k) = \sum_{N_i=N_{\min}}^{N_{\max}} P_{\text{cluster}}(T_k|N_i) \log \frac{P_{\text{cluster}}(T_k|N_i)}{P_{\text{cluster}}(T_j|N_i)} \tag{8}$$

where  $P_{\text{cluster}}(a|b)$  states for the value of the CSP histogram for the year  $a$  and cluster size  $b$ . In order to avoid numerical issues and still calculate the KLD, if any  $P_{\text{cluster}}(a|b) = 0$  we replace this zero value by  $2^{-52}$ , the machine epsilon of the simulation and analysis equipment.

As stated above, the KLD is used here as a measure of dissimilarity among the different CSP histogram with



**Fig. 7** CSP histograms resulting from the SODCC clustering of the wind speed data for the  $N = 627$  nodes and for all values of  $T_j \in T_j$  (see Eq. 7), represented as a 3D-bar plot and color gradient to indicate the bar height

different years as reference, i.e. a low KLD value means that the CSP histograms from two different years are very similar. Thus, we will have as many KLD curves as reference years.

We calculate the KLD,  $D_{KL}(T_j, T_k)$ , for all possible combination of  $(T_j, T_k)$ , being  $T_j \in T_J$  and  $T_k \in T_J$ . Figure 8 a and b show all the calculated KLD curves, separated into two differentiated patterns depending on the reference year. This separation depends on both the decay exponent of the CSP histograms shown in Fig. 7 and the shape of each KLD curve. In Fig. 8a, the reference years  $T_j$  are

$$T_j^* \in \{1986, 1988, 1994, 1996, 2001, 2003, 2008, 2010, 2014\} \quad (9)$$

The representation of the KLD curves in separated plots is useful as it helps the analysis of the results by enhancing the differences between two distinct behaviours. For example, the  $T_j^*$  subset can be easily identified from Fig. 8b as the  $D_{KL}(T_j \notin T_j^*, T_k \in T_j^*)$  stand out as maxima. This subfigure also reveals a quasi-periodic pattern for the  $T_j^*$  subset, that appears in non-consecutive pairs every 4–6 years. On the other hand, in Fig. 8a,  $D_{KL}(T_j \in T_j^*, T_k)$  exhibit significantly higher values that confirm the lower degree of similarity to other years, as the years from  $T_j^*$  are the reference for the KLD.

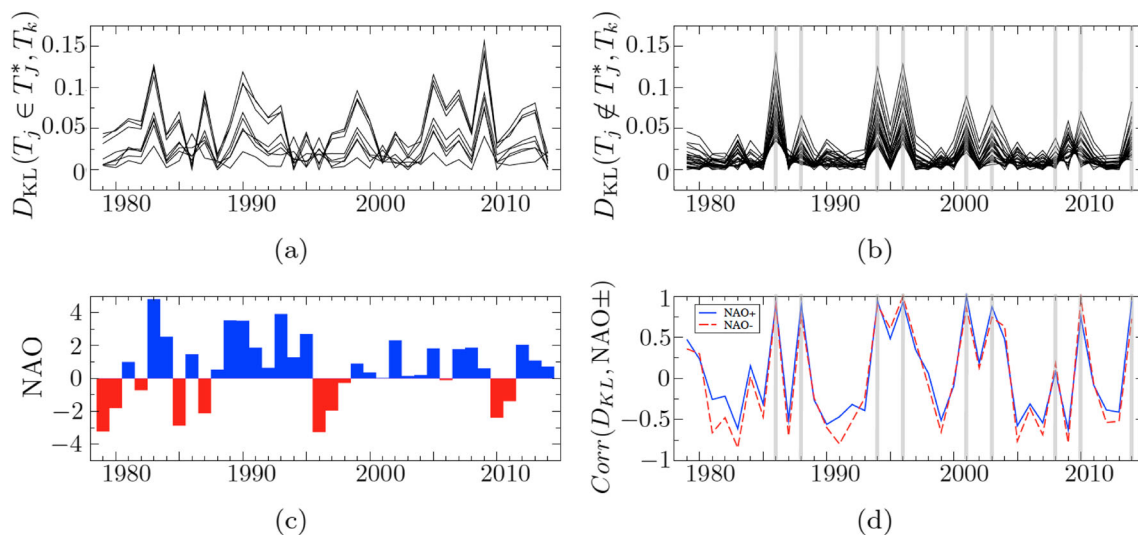
### 3.2.3 Correlation with climate indices

The KLD between the different CSP histograms offers the possibility of relating the obtained results with different

climate indices. Note that it is possible to calculate the correlation between the different KLD curves and any climate index related to the phenomenon under study. In this case, a good choice for a climate index to be correlated with wind speed in the Iberian Peninsula is the NAO (Working Group on Surface Pressure 2016). This climate index controls important processes related to Iberian Peninsula climate (Trigo et al. 2002), with deep effects not only on temperature (Prieto et al. 2002) or precipitation trends (Trigo et al. 2004), but also in renewable energy resources such as wind speed, as shown in Jerez et al. (2013). Of course, the NAO is not the only index with influence in the climate of the region, but different studies have pointed out a stronger relationship of the Iberian Peninsula climate with the NAO (Gimeno et al. 2002). On top of the already referred to works in the introduction of the present work, there are more recent studies that build on to that relation (Qu et al. 2012; Burningham and French 2013) and (Bierstedt et al. 2014).

To sum up the following analysis, we have encountered the difficulty of correlating a time series which can be positive or negative (the NAO) with a positive defined time series (the KLD). If we do that the usual way, both the correlation and the anticorrelation cancel each other, thus obtaining a numerical small correlation ( $< 0.5$ ) where a clear graphical correlation between signals can be observed. We have separated the positive and negative phases of the NAO as they have a clear climatological interpretation.

More specifically, we consider the January index of the NAO for the years under study, that is plotted in red (NAO–) and blue (NAO+) bars in Fig. 8c. Visual comparison between the KLD curves in Fig. 8a and the



**Fig. 8** Kullback-Leibler divergence  $D_{KL}(T_j, T_k)$  for  $T_k \in T_J$  and separating between (a)  $T_j \in T_j^*$  (see Eq. 9) and (b)  $T_j \notin T_j^*$ , plotted in black lines. c January index of the NAO between 1979 and 2014, plotted in red and blue bars for NAO– and NAO+, respectively.

**d** Correlation between all KLD curves and NAO index (normalized to their respective maximum values,  $\text{Corr}(D_{KL}, \text{NAO}+)_{\max} = 0.41$ ,  $\text{Corr}(D_{KL}, \text{NAO}-)_{\max} = 0.21$ ). In (b) and (d), years  $T_j \in T_j^*$  have been highlighted with grey lines

NAO index reveals a high degree of similarity. To attest this fact, we calculate the correlation between the KLD curves and the January index of the NAO, differentiating between NAO– and NAO+, and normalizing to the maximum obtained value.

This normalization is necessary as in this work we calculate the correlation between a real-defined time series (the NAO index) and a non-negative defined one (the KLD time series). Moreover, the separation between NAO+ and NAO– (e.g. by means of a half-wave rectification) is also useful because our signals of interest exhibit both correlation or anti-correlation, the cross correlation operator performs poorly (cancelling each other), and the actual similarity between the signals is not apparent. This non-linear operator (correlation a positive signal with a half-wave rectified real signal) has been described elsewhere (Eq. (3) in Doi and Fujita 2014) and it has been referred to as *cross-matching* operator.

By separating between NAO+ and NAO–, the total signal energy of the NAO series (i.e. the squared modulus of the signal integrated in time, such that the larger the variation of the NAO series between positive and negative values, the larger is its corresponding energy) is distributed between the two rectified signals (energy of the NAO signal =  $\sum_i |\text{NAO}_i|^2 = \sum_i |\text{NAO}_i^+|^2 + \sum_i |\text{NAO}_i^-|^2$ ), namely in the following percentage 68% and 32%, respectively. Therefore, the maximum value of the correlation between the KLD series and NAO+ signal would always result in value lower than unity, to be precise 0.68 for NAO+ (and 0.32 for NAO–) given the considered NAO series. In this work, the correlation values obtained between all the KLD series and NAO+ signal are  $\leq 0.41$  (and  $\leq 0.21$  for NAO–). Taking into account the half-wave rectification, these values correspond to actual correlation values of  $0.41/0.68 = 0.60$  between KLD and NAO+ and of  $0.21/0.32 = 0.66$  between KLD and NAO–.

As both actual correlation values are similar and in order to perform fair comparison avoiding different signal energy issues, in Fig. 8d, we represent the correlations series normalized to their maximum value. In order to clarify the relation to the KLD represented in Fig. 8b, we marked with vertical grey lines the reference years  $T_j \in T_j^*$  in both figures. From these results, we can explicitly state that the KLD with reference years  $T_j \in T_j^*$  is highly correlated with both phases of the NAO. Then, the spatial extent of the data correlations, observed in the NCSP histograms, is also very similar between years  $T_j \in T_j^*$  but very different to those of any other year  $T_j \notin T_j^*$ . As the KLD is used as a distance metric of “difference” between histograms, each KLD series quantifies differences between the CSPs of different years. Namely the higher the KLD, the more different the CSPs histograms and a given time series is dependent on  $T_j$ . For example, CSPs obtained for  $T_j \in T_j^*$

reveal a higher occurrence of small clusters, perceptible by the presence of multiple “valleys” in Fig. 7.

Therefore, KLD series calculated with  $T_j \in T_j^*$  as a reference, shown in Fig. 8a, have higher values on a sustained basis. These higher values are caused by the displacement of the CSPs histograms, that shift from high occurrence of small clusters to higher occurrence of large clusters. Moreover, it is noticeable that common patterns in Fig. 8a closely follow the shape of the January NAO index (depicted in Fig. 8c). Therefore, we can assert with confidence that almost all years with positive NAO index and most of the years with negative NAO index show a higher occurrence of larger clusters. We can also conclude that years  $T_j \in T_j^*$  are “tipping point” years, where no clear phase (either positive or negative) of the NAO index dominates in the Iberian Peninsula. This situation results in fewer dominant large-scale components and a larger short scale wind speed correlation, causing a more frequent and uneven distribution of turbulent cells. In the present work, these cells correspond to clusters sizes of 7 – 9 nodes covering convex clusters with a  $\approx 50$ -km radius.

In conclusion, a macroscopic time evolution of the climate regions with the SODCC analysis and its relation to NAO climate index has been established and the associated physical interpretation can be extracted from the results.

### 3.3 Comparison with EOFs and REOFs

In the following, we analyze the wind speed dataset using EOFs and REOFs, independently for each year considered in the dataset (from 1979 to 2014), using the methodology detailed in Björnsson and Venegas (1997).

The first step for this analysis is to check out whether the phase transition is achieved. In other words, we calculate the ratio between the amount of nodes and the amount of data available for each node and then we check if it is sufficient to properly calculate the EOFs. The considered dataset includes  $N = 627$  nodes in the network and  $M_{\text{month}} = 31 \times 4 = 124$  data values per node and month. The condition that establishes that the phase transition is achieved and therefore that the EOFs are properly calculated is the following:

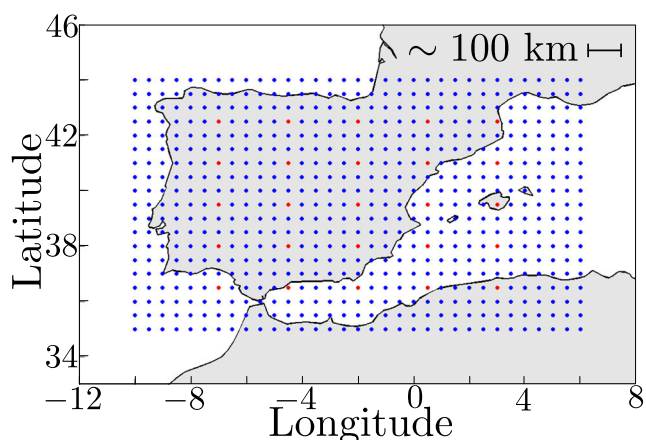
$$\frac{M_{\text{month}}}{N} \gtrsim 4 \tag{10}$$

Given the considered dataset, it is obvious that the phase transition is not achieved and then it is not possible to calculate the EOFs using all the geographical locations in the Iberian Peninsula. Therefore, it is necessary to use a subsample of the  $N$  possible nodes, in order to satisfy both the phase transition condition and the regular spatial grid requirement. Specifically, we consider the  $N_{\text{subsample}} = 25$  nodes highlighted in Fig. 9 to calculate the EOFs and the

REOFs. This spatial subsampling diminishes the spatial resolution of the grid from 56 to 168 km in the longitudinal axis and to 280 km in the latitudinal axis.

Regarding the REOF calculation, we consider the *varimax* criterion (Kaiser 1958) applied to the first three EOFs. Figure 10 shows the first EOF and the first REOF calculated over the considered dataset for several values of  $T_j \in T_J$ , revealing the different physical modes of the wind speed in the region of interest. The complete sets of figures showing the EOF and REOF are provided as electronic supplemental material (files “3\_EOF.pdf” and “4\_REOF.pdf”). As expected for this analysis, we can observe the differences between the EOFs and the REOFs, mainly caused by the inherent orthogonality obtained from the EOF calculation and relaxation of this constraint in the REOF case. It is commonly accepted that the physical modes are not in general orthogonal and, therefore, the EOF analysis is not the most suitable. On the other hand, the REOF analysis reveals the structure of the physical modes and the obtained patterns indicate underlying behaviour of the data, i.e. the wind speed in this work. Note, however, that the proposed SODCC algorithm also allows the study of the underlying data behaviour, and besides that, as we previously mentioned, it allows carrying out this study at multiple spatial resolutions of the grid.

In the following, we directly compare the REOF results with the results obtained from SODCC for the NCSP histograms, that were previously described. Figure 11 represents the results obtained for multiple values of  $T_j$  using the REOF analysis and the NCSP histograms considering cluster sizes from  $N_i = 5$  to  $N_i = 8$ . In this case, the NCSP histogram when multiple cluster sizes are considered represents the probability of a node to belong to clusters of multiple sizes, therefore is the sum of the histograms of individual cluster size. The complete set of



**Fig. 9** Geographical location and identification of the  $N_{\text{subsample}} = 25$  nodes (red dots) of the network used for the EOF and REOF analysis

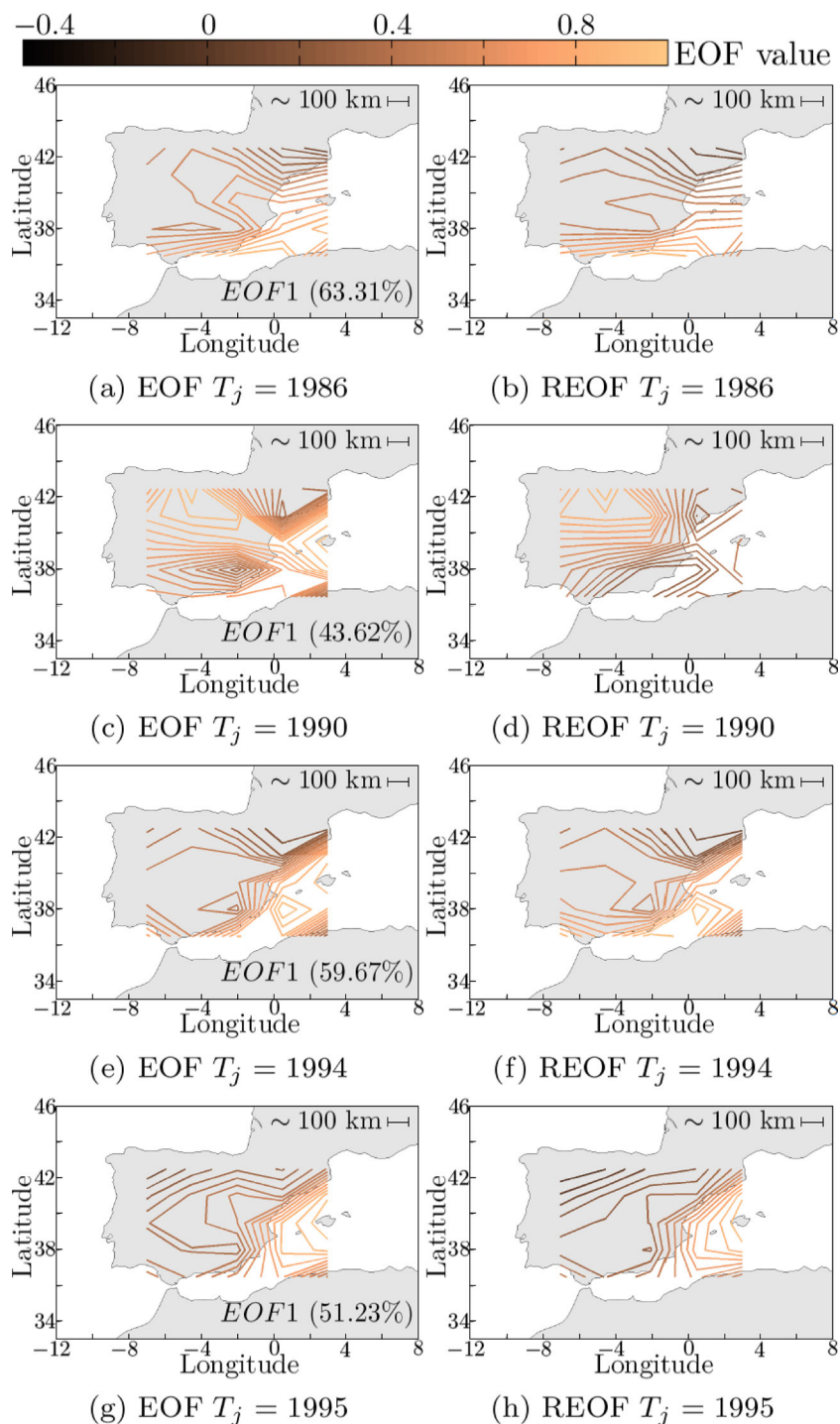
figures showing the REOF plots overlapped to the NCSP histograms is provided as electronic supplemental material (file “5\_NCSP\_REOF.pdf”). This range of cluster sizes represents the focus zone for the CSP histograms, it allows the study of short range spatial correlation, and limits the small variations between consecutive cluster sizes that can appear by chance. Moreover, the results that we may obtain with this analysis can be extrapolated to long range spatial correlations due to the dichotomy previously observed.

We can clearly see in Fig. 11 the marked change in the long to short range behaviour of clusters/cells (larger probability values) in years with either clear positive or negative NAO indices ( $T_j \notin T_J^*$ ) with respect to transition years ( $T_j \in T_J^*$ ) commented in the previous subsection. REOF values tend to show the sharpest transitions following the edges of probability change regions (red vs. green or blue) albeit with much less definition. Thus, where there is no dominant phase of the NAO, wind behaviour does not determine the sharp regional separation in climate zones. The Northeastern part of the Iberian Peninsula has marked long range behaviour (low probability for the NCSP distribution, i.e. orange/red regions) independent of the NAO phases (mainly due to the Pyrenees). On the other hand, the behaviour of the rest of the regions in the Iberian Peninsula evolves with time and shifts from long range corridors to short range turbulent cells (higher probabilities for the NCSP distribution, i.e. blue/green regions). The Northwestern region clearly shifts from long range correlations (with either negative or positive NAO) to short range (in years with transitional NAO).

A close analysis of the contour lines that represent the REOFs and the color patches that represent the NCSP histograms reveals that both results lead to comparable conclusions. Thus, it is possible to observe the same physical phenomenon using either REOF analysis or the SODCC analysis. However, the EOF and REOF calculation has several drawbacks that the SODCC algorithm is able to overcome. For example, there exists a high dependency between the REOF results and several criterion choices during their calculation, i.e. the number of EOFs used for their calculation or the rotation algorithm. Moreover, when calculating the EOFs and the REOFs, the phase transition has to be achieved for the autocorrelation matrix of the considered dataset. In other words, if a high spatial resolution is required for the grid then the dataset must include a very large amount of data. Moreover, if the dataset has not enough amount of data, the spatial resolution of the grid is compromised, as it has happened in this case study.

On the other hand, the SODCC algorithm does not depend on any parameter definition that could alter its final result. Moreover, a significantly higher spatial resolution for the grid can be obtained as the cluster organization follows a bottom-up approximation and the phase transition condition

**Fig. 10** Spatial patterns and percentage of the explained variance of the first EOF ((a), (c), (e), and (g)) and REOF ((b), (d), (f), and (h)) modes calculated for the  $N_{\text{subsample}} = 25$  spatial locations and multiple values of  $T_j$ , presented as dimensionless maps normalized to the  $[-1, 1]$  interval. Selection of years has been done in order to show no bias with respect to set selection, i.e.  $\{1987, 1994\} \in T_j^*$ ,  $\{1990, 1994\} \notin T_j^*$



has to be achieved at cluster level, requiring a significantly lower amount of data.

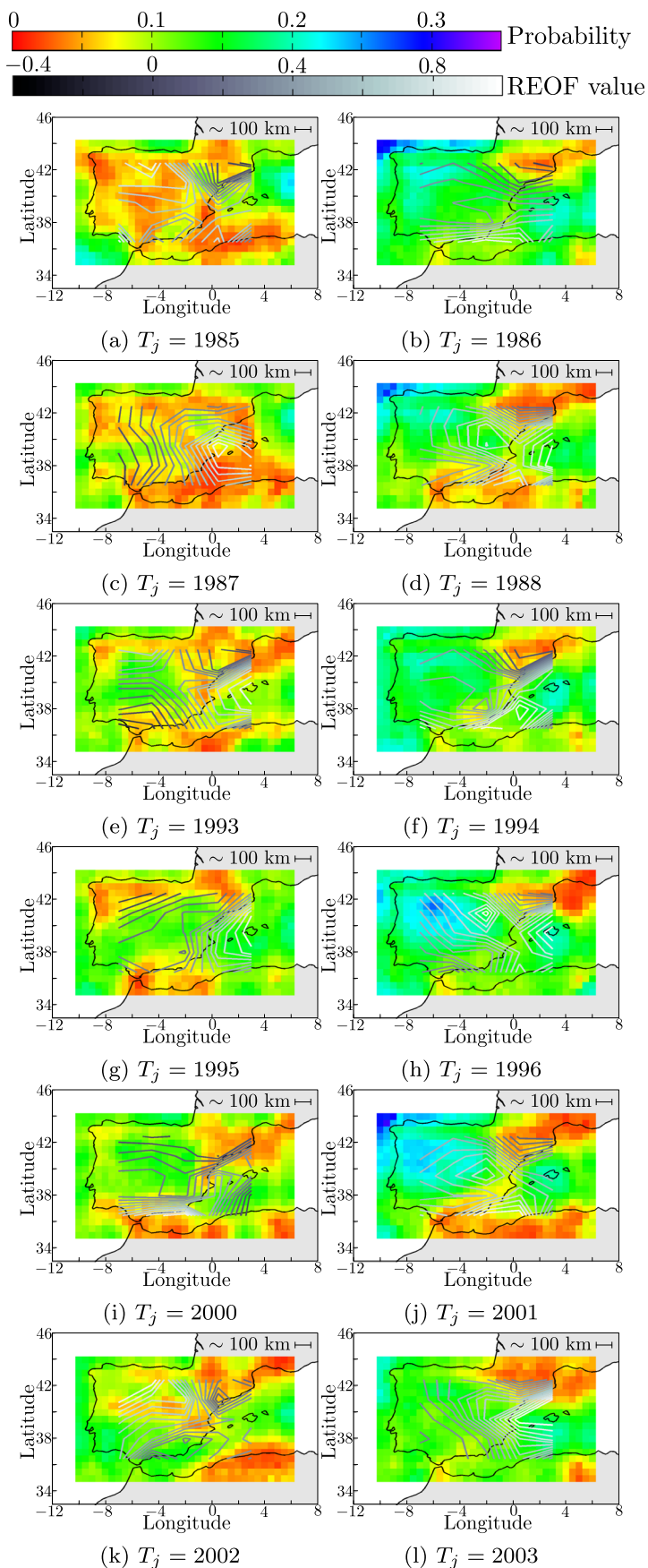
A further advantage of SODCC over REOF is the following. While the REOF analysis offers a panoramic view of the behaviour of the underlying data and its statistics, the SODCC algorithm allows the study of multiple spatial correlation extents. This benefit is achieved by analyzing different cluster sizes where the degree of correlation in

different geographical regions can be identified (recall that small clusters emerge in areas with high data correlation).

### 3.4 Robustness to initialization parameters

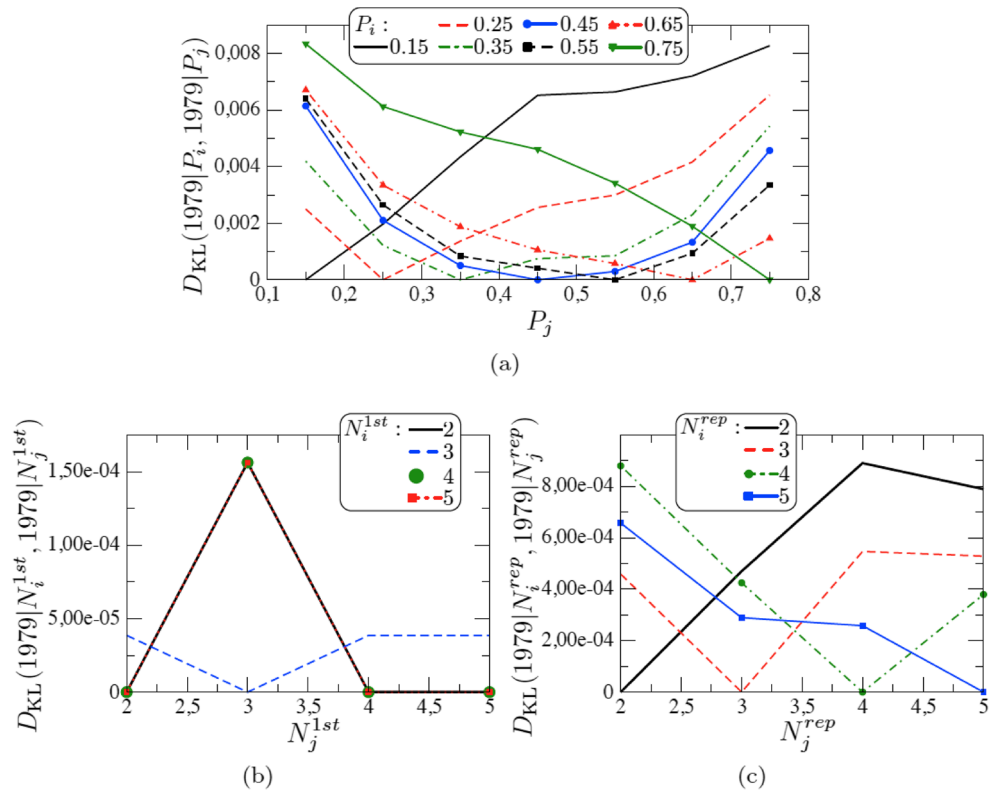
In the present section, we closely examine the robustness of the results with respect to the initialization parameters. We analyze the variation between CSP originated from

**Fig. 11** Comparison between results obtained with SODCC and REOF analysis for multiple values of  $T_j$ . The NCSP distributions are represented for  $N_i = 5$  to  $N_i = 8$  using the colorbar included at the top of this figure. The first REOF mode is represented as dimensionless map normalized to maximum 1. Colorbars are selected to facilitate the visualization. Left column corresponds to set of years  $T_j \notin T_j^*$ , right column corresponds to set of years  $T_j \in T_j^*$ . Climate regions can be clearly discerned in the NCSP distributions





**Fig. 12** KLD between CSPs obtained from simulations with different initialization parameters for year  $T_j = 1979$  with respect to the parameters used in the present work: **a** variation of the probability of a node being selected as CH,  $P_i$ , **b** variation of the maximum initial cluster size,  $N_i^{1st}$ , and **(c)** variation of the number of repetitions allowed for all the nodes to be included in any cluster,  $N_i^{rep}$



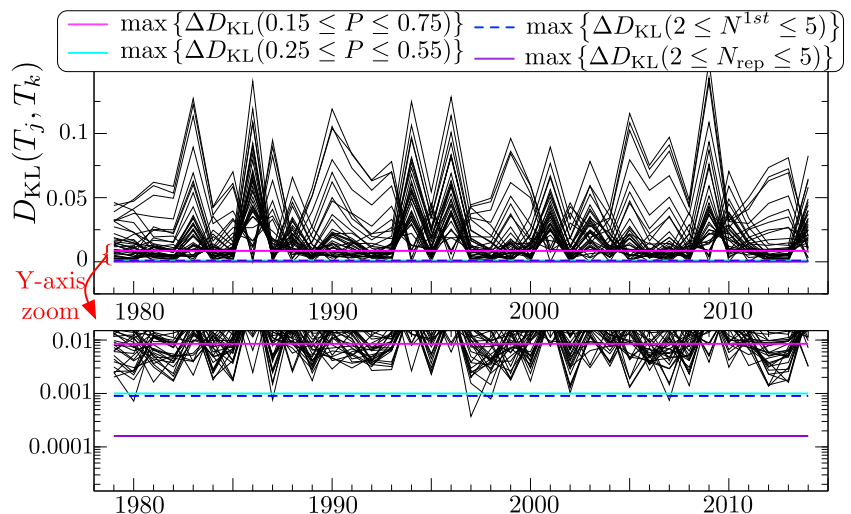
simulations with different initialization parameters, using the KLD metric. The initialization parameters under examination are as follows:

1.  $P$ —the probability of a node being selected as CH,  $P_i \in \{0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75\}$
2.  $N^{1st}$ —the maximum initial cluster size,  $N_i^{1st} \in \{2, 3, 4, 5\}$ , and
3.  $N^{rep}$ —the number of repetitions allowed for the first phase,  $N_i^{rep} \in \{2, 3, 4, 5\}$

Without any loss of generality, we have selected for the present test the first year in the series,  $T_j = 1979$ .

Figure 12 shows the resulting KLD calculated between the CSPs obtained varying the three tuning parameters. In Fig. 12a, we may observe that changes in the value of  $P_i$  lead to a convex evolution in KLD (with respect to  $P_i = 0.35$ , dash-dotted green line, the value that was used throughout this work) for all values in the  $0.35 \geq P_i \geq 0.65$  range. Furthermore, the maximum difference in KLD for that range is below  $10^{-3}$ . Thus, for reasonable values of  $P_i$  (neither too few nor too many initial clusters),

**Fig. 13** Upper panel: KLD for the different CSPs for all  $T_j \in T_j$  (joint representation of the curves in Fig. 8 (a) and (b)) along with the maximum obtained KLD values in Fig. 12 for  $0.15 \geq P_i \geq 0.75$  (magenta), reasonable values of  $0.25 \geq P_i \geq 0.55$  (cyan),  $2 \geq N_i^{1st} \geq 5$  (dashed blue), and  $2 \geq N_i^{rep} \geq 5$  (purple). Lower panel: same as the above but y-axis zoomed and in log-scale (Zero values of KLD are omitted for clarity)



selection of the specific value has no relevant impact on the obtained CSPs and, therefore, on the results of the present work. Nevertheless, extreme values of  $P_i$  are shown to have limited impact (KLD difference well below  $10^{-2}$ ).

In Fig. 12b, we may observe that the variation of the maximum initial cluster size,  $N_i^{1st}$  also leads to negligible differences between the CSPs, as the resulting KLD is well below  $10^{-3}$  for the considered range.

Finally, in Fig. 12c, we can see the resulting KLD calculated between CSPs obtained from simulations where different  $N_i^{rep}$  was used. Again, the resulting KLD is also below  $5 \cdot 10^{-4}$  indicating that the CSPs are very similar between them.

In order to clarify the overall impact of a purposefully bad selection of initial parameters (worst case scenario which determines the upper limit of variation of KLD), we have represented the maximum obtained KLD in each panel of Fig. 12 as horizontal lines in Fig. 13. Figure 13 also includes the KLD obtained for all  $T_j \in T_J$ , previously shown in Fig. 8 (a) and 8(b). It can be clearly seen that, even in the case of a bad selection of the tuning parameters, the impact on the resulting KLD patterns is well below the amplitude of the significant variations ( $\approx 10^{-4}$ ). Therefore, the correlation analysis presented in Section 3.2.3 would remain qualitatively invariant should a worst case scenario of initialization parameters arise.

Therefore, we have shown that the SODCC method is robust to the selection of the initialization parameters.

## 4 Conclusions

In this work, we have presented a novel methodology for self-organized CR based on the SODCC algorithm. This method organizes the regions by clustering the grid nodes in terms of the explained variance of physical measurements and their extent, i.e. in terms of the statistical characteristics of the time series. The main advantages of the SODCC algorithm, and hence of the CR method proposed, are that it is robust to the selection of tunable parameters and that it does not require any regular or homogeneous grid for the nodes. Moreover, the present method has higher spatial resolution of the grid, lower computational complexity, and a more direct physical interpretation of the outputs than other existing CR methods.

First, the operation of the SODCC algorithm only requires the selection of the margin of the FSD statistic to adjust to the corresponding chi-square distribution, which controls the amount of explained variance in a cluster. In this study, the margin was set such that, once the phase transition for the covariance matrix is achieved, the explained variance in each cluster is higher than 90%. This feature of SODCC is a benefit over other classical methods as the obtained result

is due to the self-organization nature of the algorithm and it is not necessary to determine *a priori* some key parameters, such as the number of EOFs used in the REOF calculation.

Second, the SODCC independence from the measurement nodes location is also a significant advantage over other classical CR methods. In the wind speed case study described in this paper, the SODCC algorithm has been applied to a regular grid. However, this method has been successfully applied to non-regular inhomogeneous grids (Chidean et al. 2015b).

The analysis of the obtained set of clusters enables a microscopic characterization of the climate regions and their spatial extent. The NCSP histograms show the probability of each node to belong to a given climate region (with joint explained variance of 90%) of a given size. These outcomes reveal the different physical modes of the data in the region of interest.

The SODCC outcomes also permit the study of the macroscopic behaviour of the different climate regions through time. The CSP histograms show the overall behaviour of the final set of clusters. The temporal analysis of these results can be performed by using a measure of similarity between the CSP histograms, such as the KLD.

This work also includes the analysis of the wind speed in the Iberian Peninsula in January for the period (1979 – 2014) as a specific case study. By applying the SODCC method to the present time series of data, it is possible to (1) characterize spatial probabilistic climate regions that have higher than 90% explained variance with high spatial resolution of the grid and (2) characterize the time evolution of the similarity of the studied area as a whole. Two distinct patterns emerged in the KLD curves initialized in the different years, pointing to a subset of years ( $T_j^*$ ) with distinct behaviour. Comparison between the KLD curves and both the positive and negative phases of the NAO showed higher correlation values for the  $T_j^*$  subset of reference years. The CR results corresponding to the  $T_j^*$  subset of years also showed distinctive features with large spatial correlations. These results are corroborated by means of the REOF analysis, with much lower spatial resolution of the grid however.

Some general results obtained in the present study are in the line pointed out by previous approaches. In Lorente-Plazas et al. (2015), a characterization of surface winds over the Iberian Peninsula is performed via a PCA and clustering approach in which 20 differentiated regions in the Iberian Peninsula. In the present work, we go beyond that characterization and perform a time analysis and a novel correlation with the NAO to show a high correlation of wind patterns in the Iberian Peninsula with such oscillation. This finding is supported, for example, in Jerez et al. (2013) where it indicates a close relationship of the wind speed resource with the NAO. Another work with comparable

results with this one is Azorin-Molina et al. (2014), where a study on surface wind speed trend in the Iberian Peninsula is carried out based on measuring stations data. A direct comparison of results is difficult to establish, due to the different methodologies applied, observation lengths, and data used. However, in Azorin-Molina et al. (2014), it is also recognized the relationship of the wind speed resource with the different phases of the NAO. But to add to these previous findings, we have shown that there is statistical evidence that the local wind speed clusters are large while the NAO phase is either positive or negative. However, in the transition years, where the NAO changes its sign, the local wind speed clusters diminish in size. Thus, the regionalization of wind speed zones in the Iberian Peninsula is shown to depend, not only in the geographical features, but also on the NAO phase.

Finally, as main conclusions of this work, we have found the following: (1) the climate regions obtained by means of the SODCC algorithm can be determined with higher spatial resolution of the grid than equivalent analysis methods, e.g. REOF analysis; (2) the SODCC outcomes facilitate the physical interpretation of the underlying phenomena; and (3) the NAO has a huge impact in the wind distribution in the Iberian Peninsula in only a subset of years in the studied period. This indicates that the SODCC is a valuable tool to study the microscopic spatial distribution of climate regions and also the macroscopic temporal evolution of those regions.

**Acknowledgements** ERA-Interim data provided courtesy of ECMWF. The authors would like to thank the editing and reviewing effort of the anonymous peers and the editor.

**Funding information** This work has been partially supported by the Autonomous Community of Madrid (Grant Ref. S2013/MAE-2835) and by the Ministerio de Economía y Competitividad of Spain (Grant Ref. TIN2017-85887-C2-2-P).

## Appendix A: Phase transition of correlation matrices with a finite number of samples

Using matrix perturbation analysis, it is possible to set a lower bound on the amount of samples to recover the largest eigenvalue of the corresponding correlation matrix (Nadler 2008).

Consider the data matrix  $\mathbf{X} \in \mathbb{R}^{N \times M}$ , being  $M$  the number of available data samples for each of the  $N$  measuring stations or grid sites that we aim to analyze. Its covariance matrix can be defined as  $\Sigma = \mathbf{M}_X \mathbf{M}_X^\top / M$ , being  $\mathbf{M}_X$  the mean-centered data matrix, i.e. transformation of matrix  $\mathbf{X}$  such that the mean of each row is zero.

Let the noise power of the measurement be defined as  $\sigma^2$  and the modulus of the first EOF of  $\Sigma$  as  $\|\mathbf{v}\|^2$ . The noise of the first EOF being simply a white noise stochastic physical

process that perturbs all EOFs equally in time and space. Thus, the Signal to Noise Ratio (SNR) needed to resolve the first EOF (the one with higher explained variance) can be defined as

$$\text{SNR}_v = \|\mathbf{v}\|^2 / \sigma^2 \tag{11}$$

It has been shown that the stochastic self-adjoint matrix  $\Sigma$  experiences a *phase transition* for  $M$  samples obtained from  $N$  measuring stations (for  $N/M = \text{constant}$  as  $N$  and  $M$  got to  $\infty$ ) s.t.  $M/N \geq \text{SNR}_v^{-2}$  (Nadler 2008, Eq. 2.19). The phase transition can be detected in the largest eigenvalue as it “arises” from the set of noise degenerate eigenvalues to a signal eigenvalue. This phase transition has been identified also in the fields of unsupervised learning (via the replica method) (Biehl and Mietzner 1994) and statistics (via the Stieltjes transform) (Baik and Silverstein 2006).

This, in fact, provides a threshold on the minimum amount of samples  $M$  needed to detect the eigenvalue with a minimum  $\text{SNR}_v$  as

$$M \geq \frac{N}{\text{SNR}_v^2} \tag{12}$$

Throughout the present work, we have used as working threshold to separate *signal eigenvalues* from *noise eigenvalues* that establishes that the minimum number of samples needed to obtain at least a signal eigenvalue is

$$M = 4 \times N \tag{13}$$

This assumption is valid as long as the only intervening noise in the mixture of the EOF is additive, white, and Gaussian.

Thus, we can see that the phase transition sets a limit on the minimum number of samples  $M$  for a  $\Sigma = \mathbf{M}_X \mathbf{M}_X^\top / M$  covariance matrix to be well conditioned, i.e. non-singular. By using a consistent threshold for the extraction of the EOFs, no matter the size of the originating covariance matrix (i.e. from small to large clusters or the complete network), we ensure the coherence and comparability of the results at any spatial scale.

## Appendix B: Fast subspace decomposition

In this work, we use the Fast Subspace Decomposition (FSD) algorithm to estimate the number of eigenvalues  $\hat{d}$  that account for a given amount of explained variance (Xu and Kailath 1994). The FSD algorithm estimates the first  $\hat{d}$  Rayleigh-Ritz (RR) eigenvalues and eigenvectors (spanning the signal subspace) up to the  $\hat{d} = d$  iteration, where  $d$  is the value to be estimated. FSD is based on the Lanczos method and has  $O(N^2d)$  computational complexity, much lower than that of the traditional eigendecomposition, that has order of  $O(N^3)$ .

For a data matrix  $\mathbf{X} \in \mathbb{R}^{N \times M}$  with  $\mathbf{M}_\mathbf{X}$  the centered data matrix and  $\Sigma = \mathbf{M}_\mathbf{X} \mathbf{M}_\mathbf{X}^\top / M$  the covariance matrix, the FSD statistic  $\varphi_{\hat{d}}$  is defined as (Xu and Kailath 1994):

$$\varphi_{\hat{d}} = M(N - \hat{d}) \log \left[ \frac{\sqrt{\frac{1}{N-\hat{d}} (\|\Sigma\|^2 - \sum_{n=1}^N \theta_n^2)}}{\frac{1}{N-\hat{d}} (\text{Tr}\Sigma - \sum_{n=1}^N \theta_n)} \right] \quad (14)$$

where  $\|\cdot\|$  is the Frobenius norm,  $\theta_n$  is the RR eigenvalues, and  $\text{Tr}\Sigma$  is the trace of matrix  $\Sigma$ . In each iteration, the  $\varphi_{\hat{d}}$  statistic is computed and for  $\hat{d} \geq d + 1$ ,  $\varphi_{\hat{d}}$  approaches a  $\chi^2$  distribution with  $(1/2)(N - \hat{d})(N - \hat{d} + 1) - 1$  degrees of freedom.

Finally, it has been proven that, for  $M$  samples, the following equation is valid (Xu and Kailath 1994)

$$\varphi_{\hat{d}} \leq \gamma_{\hat{d}} c(M) \quad (15)$$

where  $\gamma_{\hat{d}}$  is a threshold computed *a priori* as the end tail of the  $\chi^2$  distribution according to the required amount of explained variance. As in the present work we have considered that the minimum explained variance by the  $\hat{d}$  eigenvalues estimated by FSD is 90%, we calculated the

value  $\gamma_{\hat{d}}$  such that it accounts for 0.1 of the area of the corresponding  $\chi^2$  distribution. Also, function  $c(M)$  must comply with the following:

$$\lim_{M \rightarrow \infty} \frac{c(M)}{M} = 0 \quad \text{and} \quad \lim_{M \rightarrow \infty} \frac{c(M)}{\log \log M} = \infty \quad (16)$$

In practice, its asymptotic behaviour must be ‘‘slower’’ than linear but ‘‘faster’’ than  $\log \log$ ; functions such as  $c(M) = \log(M)$  or  $c(M) = \sqrt{\log(M)}$  can be used.

### Appendix C. Schematic of the method

In this work, we use the SODCC algorithm as the core of a climate data analysis procedure. In order to facilitate both the understanding of our results and conclusions and reproducibility of this analysis, in this appendix, we include a general scheme of the complete analysis performed in this work (Fig. 14). We also include detailed schematic representations of the performed simulations in Figs. 15 and 16 and of the analysis carried out in Fig. 17.

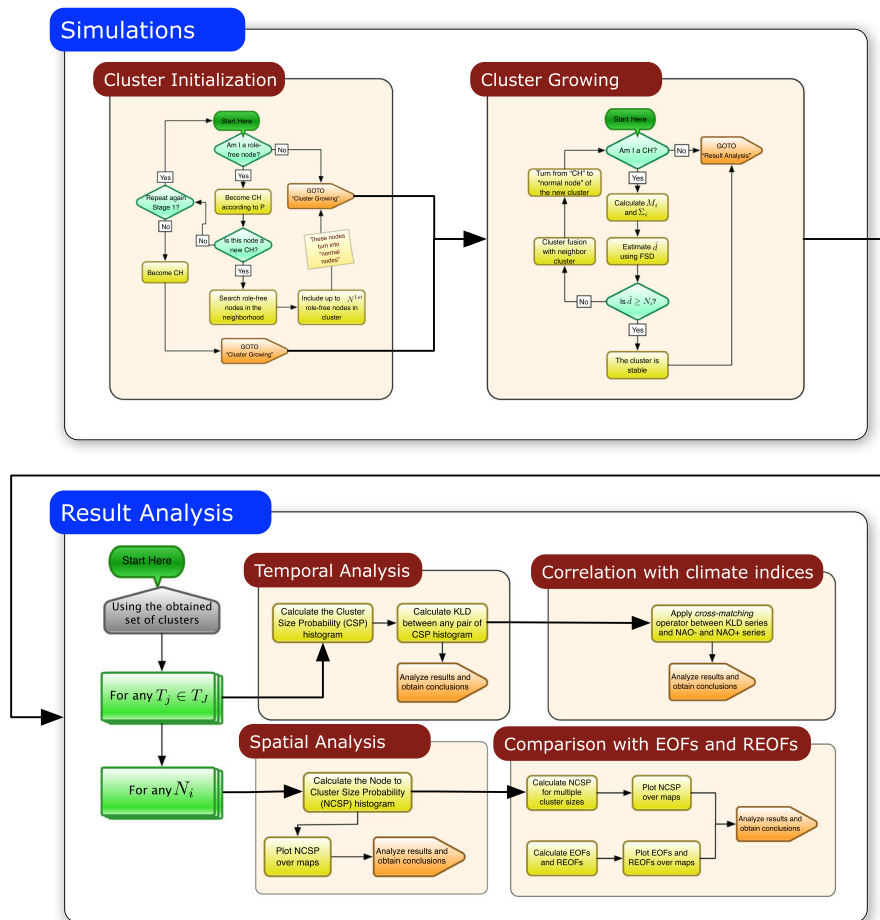
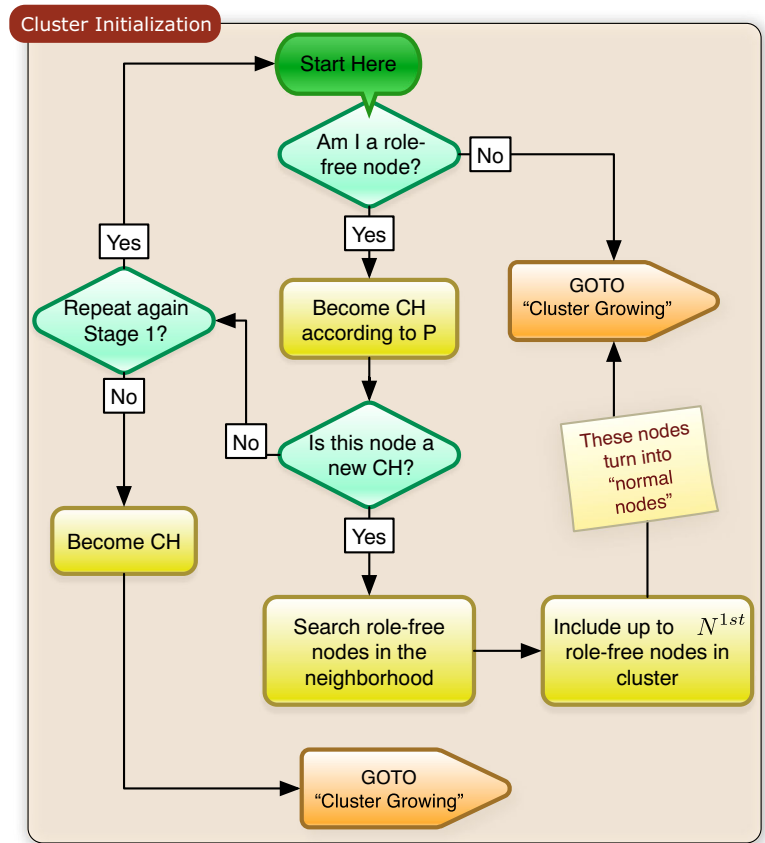
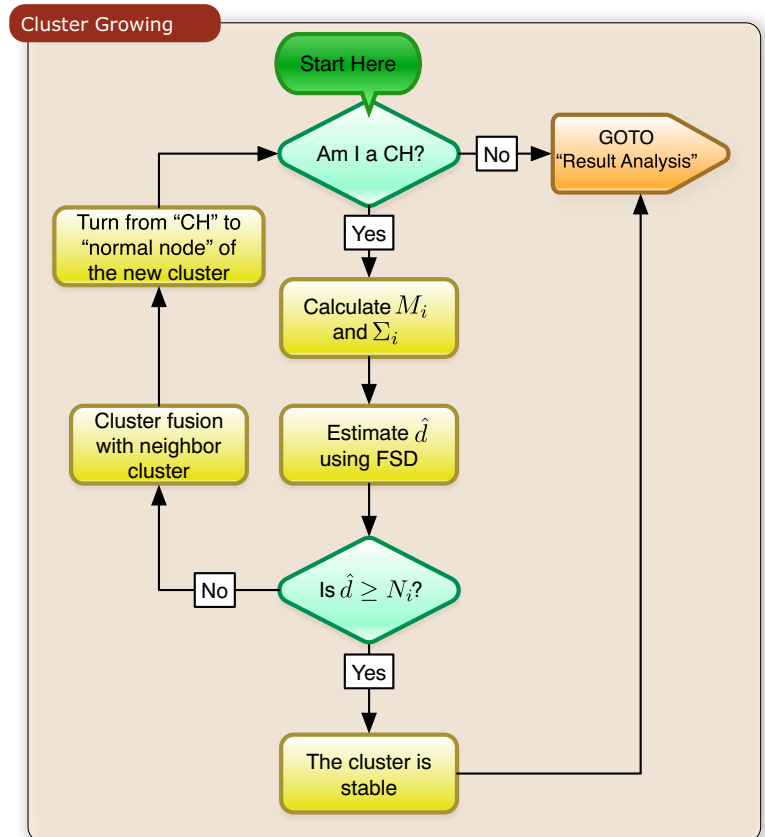


Fig. 14 Schematic representation of the simulations and result analysis performed in this work. Each inset is zoomed in Figs. 15, 16, and 17, respectively

**Fig. 15** Schematic representation of the cluster initialization stage of the SODCC algorithm



**Fig. 16** Schematic representation of the cluster growing stage of the SODCC algorithm



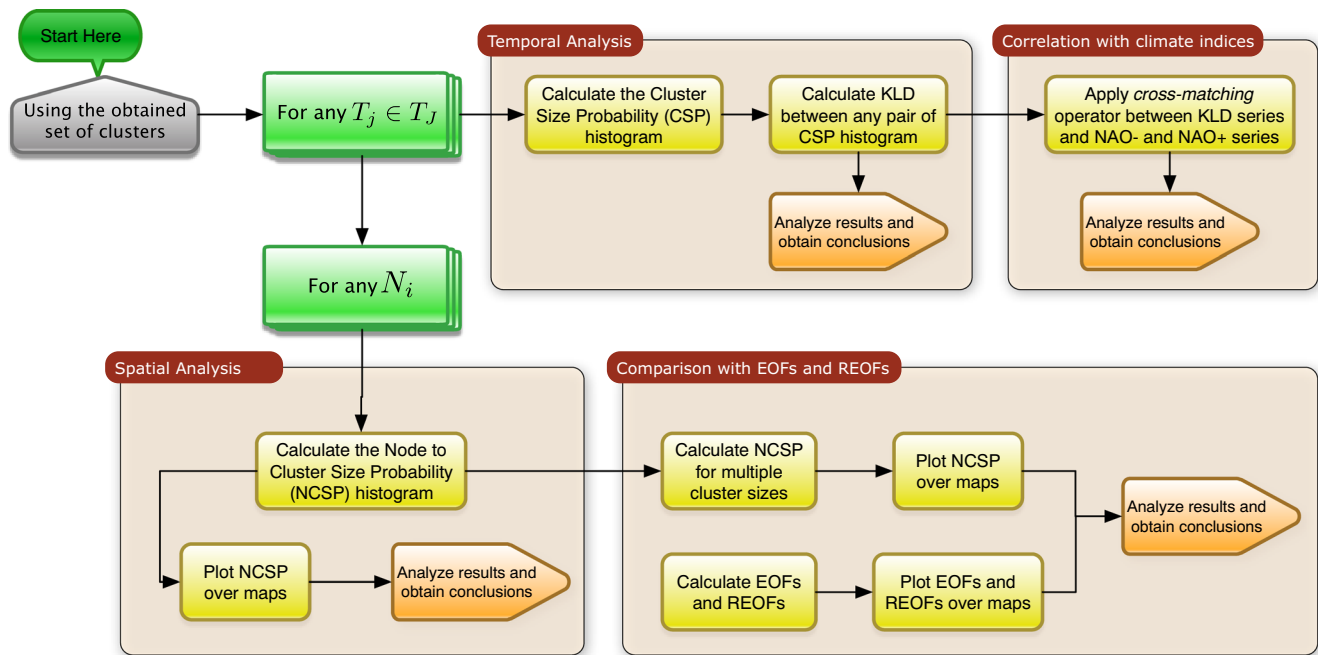


Fig. 17 Schematic representation of the result analysis performed in this work

## References

- Argüeso D, Hidalgo-Muñoz JM, Gámiz-Fortis SR, Esteban-Parra MJ, Dudhia J, Castro-Díez Y (2011) Evaluation of WRF parameterizations for climate studies over Southern Spain using a multistep regionalization. *J Climate* 24(21):5633–5651
- Azorin-Molina C, Vicente-Serrano SM, McVicar TR, Jerez S, Sanchez-Lorenzo A, López-Moreno JI, Revuelto J, Trigo R, Lopez-Bustins JA, Espírito-Santo F (2014) Homogenization and assessment of observed near-surface wind speed trends over Spain and Portugal, 1961–2011. *J Climate* 27(10):3692–3712
- Badr HS, Zaitchik BF, Dezfuli AK (2014) Climate regionalization through hierarchical clustering: options and recommendations for Africa. *AGU Fall Meeting Abstracts*
- Badr HS, Zaitchik BF, Dezfuli AK (2015) A tool for hierarchical climate regionalization. *Earth Sci Inform* 8(4):949–958
- Baeriswyl PA, Rebetez M (1997) Regionalization of precipitation in Switzerland by means of principal component analysis. *Theor Appl Climatol* 58(1–2):31–41
- Baik J, Silverstein JW (2006) Eigenvalues of large sample covariance matrices of spiked population models. *J Multivar Anal* 97(6):1382–1408
- Biehl M, Mietzner A (1994) Statistical mechanics of unsupervised structure recognition. *J Phys A Math Gen* 27(6):1885
- Bierstedt SE, Hünicke B, Zorita E, Wagner S, Gómez-Navarro JJ (2014) Variability of daily winter wind speed distribution over Northern Europe during the past millennium in regional and global climate simulations. *Clim Past* 2(12):317–338
- Björnsson H, Venegas S (1997) A manual for EOF and SVD analyses of climatic data. *CCGCR Report* 97(1):112–134
- Branick ML (1997) A climatology of significant winter-type weather events in the contiguous United States, 1982–94. *Weather Forecast* 12(2):193–207
- Burn DH (1989) Cluster analysis as applied to regional flood frequency. *J Water Resour Plan Manag* 115(5):567–582
- Burningham H, French J (2013) Is the NAO winter index a reliable proxy for wind climate and storminess in Northwest Europe? *Int J Climatol* 33(8):2036–2049
- Carvalho M, Melo-Gonçalves P, Teixeira J, Rocha A (2016) Regionalization of Europe based on a K-Means cluster analysis of the climate change of temperatures and precipitation. *Phys Chem Earth Parts A/B/C* 94:22–28
- Cassou C, Terray L, Hurrell JW, Deser C (2004) North Atlantic winter climate regimes: spatial asymmetry, stationarity with time, and oceanic forcing. *J Climate* 17(5):1055–1068
- Chidean MI, Morgado E, del Arco E, Ramiro-Bargueño J, Caamaño AJ (2015a) Scalable data-coupled clustering for large scale WSN. *IEEE Trans Wireless Commun* 15:4681–4694
- Chidean MI, Muñoz-Bulnes J, Ramiro-Bargueño J, Caamaño A. J, Salcedo-Sanz S (2015b) Spatio-temporal trend analysis of air temperature in Europe and Western Asia using data-coupled clustering. *Global Planet Change* 129:45–55
- Chidean MI, Caamaño AJ, Ramiro-Bargueño J, Casanova-Mateo C, Salcedo-Sanz S (2018) Spatio-temporal analysis of wind resource in the Iberian Peninsula with data-coupled clustering. *Renew Sustain Energy Rev* 81:2684–2694
- Comrie AC, Glenn EC (1998) Principal components-based regionalization of precipitation regimes across the southwest United States and northern Mexico, with an application to monsoon precipitation variability. *Climate Res* 10(3):201–215
- Dee DP, Uppala SM et al (2011) The ERA–interim reanalysis: configuration and performance of the data assimilation system. *Q J Roy Meteorol Soc* 137(656):553–597
- Doi T, Fujita I (2014) Cross-matching: a modified cross-correlation underlying threshold energy model and match-based depth perception. *Front Comput Neurosci* 8(127):1–15
- Dommenget D, Latif M (2002) A cautionary note on the interpretation of EOFs. *J Climate* 15(2):216–225
- Gimeno L, Ribera P, Iglesias R, de la Torre L, García R, Hernández E (2002) Identification of empirical relationships between indices

- of ENSO and NAO and agricultural yields in Spain. *Clim Res* 21(2):165–172
- Irwin SE (2015) Assessment of the regionalization of precipitation in two Canadian climate regions: a fuzzy clustering approach. Ph.D. thesis, The University of Western Ontario
- Jerez S, Trigo R, Vicente-Serrano SM, Pozo-Vázquez D, Lorente-Plazas R, Lorenzo-Lacruz J, Santos-Alamillos F, Montávez J (2013) The impact of the North Atlantic oscillation on renewable energy resources in southwestern Europe. *J Appl Meteorol Climatol* 52(10):2204–2225
- Jolliffe I (2002) *Principal component analysis*, 2nd edn. Springer
- Kaiser HF (1958) The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23(3):187–200
- Kim KY, Hamlington B, Na H (2015) Theoretical foundation of cyclostationary EOF analysis for geophysical and climatic variables: concepts and examples. *Earth Sci Rev* 150:201–218
- Knapp PA, Grissino-Mayer HD, Soulé PT (2002) Climatic regionalization and the spatio-temporal occurrence of extreme single-year drought events (1500–1998) in the interior Pacific Northwest, USA. *Quatern Res* 58(3):226–233
- Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* 22(1):79–86
- Lian T, Chen D (2012) An evaluation of rotated EOF analysis and its application to tropical Pacific SST variability. *J Climate* 25(15):5361–5373
- Lorente-Plazas R, Montávez JP, Jimenez PA, Jerez S, Gómez-Navarro JJ, García-Valero JA, Jimenez-Guerrero P (2015) Characterization of surface winds over the Iberian Peninsula. *Int J Climatol* 35(6):1007–1026
- Nadler B (2008) Finite sample approximation results for principal component analysis: a matrix perturbation approach. *Ann Stat*, 2791–2817
- Nojarov P (2017) Genetic climatic regionalization of the Balkan Peninsula using cluster analysis. *J Geogr Sci* 27(1):43–61
- Önol B, Semazzi FHM (2009) Regionalization of climate change simulations over the Eastern Mediterranean. *J Climate* 22(8):1944–1961
- Prieto L, García R, Díaz J, Hernández E, Del Teso T (2002) NAO influence on extreme winter temperatures in Madrid (Spain). *Annales Geophysicae* 20(12):2077–2085
- Qu B, Gabric AJ, Zhu J, Lin D, Qian F, Zhao M (2012) Correlation between sea surface temperature and wind speed in Greenland Sea and their relationships with NAO variability. *Water Sci Eng* 5(3):304–315
- Regonda SK, Zaitchik BF, Badr HS, Rodell M (2016) Using climate regionalization to understand climate forecast system version 2 (CFSv2) precipitation performance for the Conterminous United States (CONUS). *Geophys Res Lett* 43(12):6485–6492
- Richman M (1986) Rotation of principal components. *J Climatol* 6(3):293–335
- Santos-Alamillos F, Thomaidis N, Quesada-Ruiz S, Ruiz-Arias J, Pozo-Vázquez D (2016) Do current wind farms in Spain take maximum advantage of spatiotemporal balancing of the wind resource? *Renew Energy* 96:574–582
- Sarma AK, Hazarika J (2014) GCM based fuzzy clustering to identify homogeneous climatic regions of North-East India. *World Academy of Science, Engineering and Technology, International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering* 8(12):807–814
- Shahriar F, Montazeri M, Momeni M, Freidooni A (2015) Regionalization of the climatic areas of Qazvin province using multivariate statistical methods. *Mod Appl Sci* 9(2):123
- Trigo R, Osborn TJ, Corte-Real JM (2002) The North Atlantic oscillation influence on Europe: climate impacts and associated physical mechanisms. *Climate Res* 20(1):9–17
- Trigo R, Pozo-Vázquez D, Osborn TJ, Castro-Díez Y, Gámiz-Fortis S, Esteban-Parra MJ (2004) North Atlantic Oscillation influence on precipitation, river flow and water resources in the Iberian Peninsula. *Int J Climatol* 24(8):925–944
- Troccoli A, Muller K, Coppin P, Davy R, Russell C, Hirsch AL (2012) Long-term wind speed trends over Australia. *J Climate* 25(1):170–183
- Unal Y, Kindap T, Karaca M (2003) Redefining the climate zones of Turkey using cluster analysis. *Int J Climatol* 23(9):1045–1055
- Vicsek T, Family F (1984) Dynamic scaling for aggregation of clusters. *Phys Rev Lett* 52:1669–1672
- von Storch H, Zwiers FW (1999) *Statistical analysis in climate research*. Cambridge University Press
- Ward JH (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58(301):236–244
- White D, Richman M, Yarnal B (1991) Climate regionalization and rotation of principal components. *Int J Climatol* 11(1):1–25
- Working Group on Surface Pressure (2016) Download Climate Time-series. North Atlantic Oscillation (NAO). [https://www.esrl.noaa.gov/psd/gcos\\_wgsp/Timeseries/NAO/](https://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/NAO/). Accessed: 2016-01-22
- Xu G, Kailath T (1994) Fast subspace decomposition. *IEEE Trans Signal Process* 42(3):539–551
- Yu L, Zhong S, Bian X, Heilman WE (2015) Temporal and spatial variability of wind resources in the United States as derived from the climate forecast system reanalysis. *J Climate* 28(3):1166–1183
- Zhang Y, Moges S, Block P (2016) Optimal cluster analysis for objective regionalization of seasonal precipitation in regions of high spatial-temporal variability: application to Western Ethiopia. *J Climate* 29(10):3697–3717
- Zishka KM, Smith PJ (1980) The climatology of cyclones and anticyclones over North America and surrounding ocean environs for January and July, 1950–77. *Mon Weather Rev*, 387–401

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.