Global and Planetary Change 129 (2015) 45–55

Contents lists available at ScienceDirect

Global and Planetary Change

journal homepage: www.elsevier.com/locate/gloplacha

Spatio-temporal trend analysis of air temperature in Europe and Western Asia using data-coupled clustering

Mihaela I. Chidean [a], Jesús Muñoz-Bulnes [b], Julio Ramiro-Bargueño [a], Antonio J. Caamaño [a,*], Sancho Salcedo-Sanz [b]

[a] Dept. of Signal Theory and Communications, Universidad Rey Juan Carlos, Madrid, Spain
[b] Dept. of Signal Processing and Communications, Universidad de Alcalá, Madrid, Spain

ARTICLE INFO

Article history:
Received 6 December 2014
Received in revised form 27 February 2015
Accepted 18 March 2015
Available online 25 March 2015

Keywords:
Air temperature
Clustering techniques
Spatio-temporal trend
Climate change

ABSTRACT

Over the last decades, different machine learning techniques have been used to detect climate change patterns, mostly using data from measuring stations located in different parts of the world. Some previous studies focus on temperature as primary variable of study, though there have been other works focused on precipitation or even wind speed as objective variable. In this paper, we use the self-organized Second Order Data Coupled Clustering (SODCC) algorithm to carry out a spatio-temporal analysis of temperature patterns in Europe. By applying the SODCC we identify three different regimes of spatio-temporal correlations based on their geographical extent: small, medium, and large-scale regimes. Based on these regimes, it is possible to detect a change in the spatio-temporal trend of air temperature, reflecting a shift in the extent of the correlations in stations in the Iberian Peninsula and Southern France. We also identify an oscillating spatio-temporal trend in the Western Asia region and a stable medium-scale regime affecting the British Isles. These results are found to be consistent with previous studies in climate change. The patterns obtained with the SODCC algorithm may represent a signal of climate change to be taken into account, and so the SODCC could be used as detection method.

1. Introduction

The state-of-art for detection and attribution (D&A) of climate change problems usually considers two types of models: General Circulation Models (GCMs) and statistic models. The GCMs are mainly used to understand the dynamics of the physical components of the atmosphere, that are related to the climatic change phenomenon. The goal of GCMs is to obtain spatio-temporal climatic change patterns (usually global patterns), also known in the literature as fingerprints of the climatic change, and they can also be used to make predictions and long-term projections of climatic variables (Cramer et al., 2013). In D&A problems, the GCMs have been used with different temporal scales, from seasonal to decadal time-horizons. At present, some of the most sophisticated GCMs used to study climate change are developed by the WCRP's Working Group on Coupled Modeling (WGCM, 2014), which provides a multi-model context for carrying out coordinated climate model experiments, especially well-suited for D&A problems (Solomon et al., 2009).

On the other hand, the use of alternative statistic models has grown in the last years, due to they are able to obtain clear evidence of climate change in a fraction of the computational time needed by the GCMs. Several statistical models have been applied to different D&A problems, i.e. methods developed for econometric series (Kaufmann & Stern, 1997), also known as co-integration methods (Kaufmann & Stern, 2002; Kaufmann et al., 2011), or methods based on regression type approximations to evaluate climate change patterns using temperature data (Douglass et al., 2004; Stone & Allen, 2005).

Air temperature is a key parameter for the detection of climate change in certain areas, the assessment of its impact in different ecosystems (Alva-Basurto & Arias-González, 2014; Gomiero & Viarengo, 2014). Air temperature it is also related to the evaluation of the human activity, such as agriculture (Smith et al., 2009; Cobaner et al., 2014), health-care (Garske et al., 2013; Xu et al., 2014), and energy (Paniagua-Tineo et al., 2011; Jaglom et al., 2014). Temperature analysis in climate change studies may also involve problems of prediction, reconstruction or spatio-temporal analysis of the results; we are interested in the latter, where there are important previous works in the literature, published in the last few years. In Carrera-Hernández and Gaskin (2007) a study of spatio-temporal analysis using minimum temperature and precipitation data was carried out for Mexico. Several regression algorithms based on Kriging were applied in this case to more than 200 stations located in the Basin of Mexico for the period 1978–1985. In Kousari et al. (2013) maximum temperature data in Iran over the 1960–2005 period were analyzed with implications for

* Corresponding author.
E-mail address: antonio.caamano@urjc.es (A.J. Caamaño).

http://dx.doi.org/10.1016/j.gloplacha.2015.03.006
0921-8181/

climate change detection in the area. The authors applied different techniques including statistical test, filtering, and hierarchical clustering to data from 32 synoptic stations that cover Iran, and showed the trends and variability of temperature at different scales (annual, seasonal, and monthly). Another recent work on temperature variability is Kloog et al. (2012), where satellite surface measurements are used to analyze the spatio-temporal trends of minimum temperature for Massachusetts (USA). The study applies regression techniques and spatial smoothing and incorporates alternative meteorological data. Somehow related to that work is Van De Kerchove et al. (2013), in which the authors present a study of spatio-temporal variability of temperature in a remote region of Russia. Signal processing methods such as Fast Fourier Transform and others such as multi-linear regression are applied in this case to carry out the study.

In this paper we focus on the study of spatio-temporal analysis of air temperature. We propose the application of a novel clustering algorithm to this end, with focus on European data. Specifically, we use the Second Order Data Coupled Clustering (SODCC) algorithm, that was initially developed for joint data codification and self-organization of wireless sensor networks (Chidean et al., 2013). The SODCC is a self-organized clustering approach, that uses characteristics of the measured data (air temperature in this case) and geographically groups the measuring stations. To achieve this end, the proposed algorithm uses second-order statistics to compute the minimum amount of linearly independent components in the cluster, i.e., the number of principal components that explain most of the variance in the data. Note that previous spatio-temporal clustering approaches such as Carrera-Hernández and Gaskin (2007), Horenko(2010), and Kousari et al. (2013) exploit the similarities among air temperature temporal series of different measuring stations to ascribe them to different clusters, which may or may not be spatially compact. In contrast, SODCC performs joint spatio-temporal clustering by: 1) minimizing the number of principal components needed to explain the variance of the data and 2) not allowing for spatially disjoint clusters to occur.

We apply the SODCC algorithm to temperature fields obtained from measuring stations throughout Europe (Klein Tank et al., 2002; Chimani et al., 2013) considering different initial points (at decadal time scale). From the obtained cluster distributions we are able to evaluate the relations between neighboring measuring stations, the probability of cluster formation, and their time evolution. This fact helps us identify a possible climate change (or climate variability) pattern in the Iberian Peninsula from the 1970s onwards.

The rest of this paper is organized as follows: Section 2 describes the SODCC algorithm and its implementation; in Section 3 we describe the experiments made using data from weather stations and we show a possible climate change pattern that can be identified from our results. Finally, Section 4 concludes the article.

2. Self-organized clustering based on second order statistics

In this section we describe the Second Order Data Coupled Clustering (SODCC) algorithm, which we use in this work to analyze air temperature data. SODCC is a self-organized clustering algorithm and was initially proposed for wireless sensor networks. By using characteristics of the measured data (e.g. temperature), SODCC groups the measuring stations by coupling the clusters to the measured data field. For it, the algorithm uses second-order statistics to compute the minimum amount of linearly independent components in the cluster (the dimension of the signal subspace or, similarly, the number of principal components that explain most of the variance in the data). In the subsequent stages, the clusters merge until a stopping criterion based on the dimension of signal subspace is reached. This approach ensures the non-singularity of the signal subspace of the data measured by each cluster (the covariance matrix of the data is well-posed).

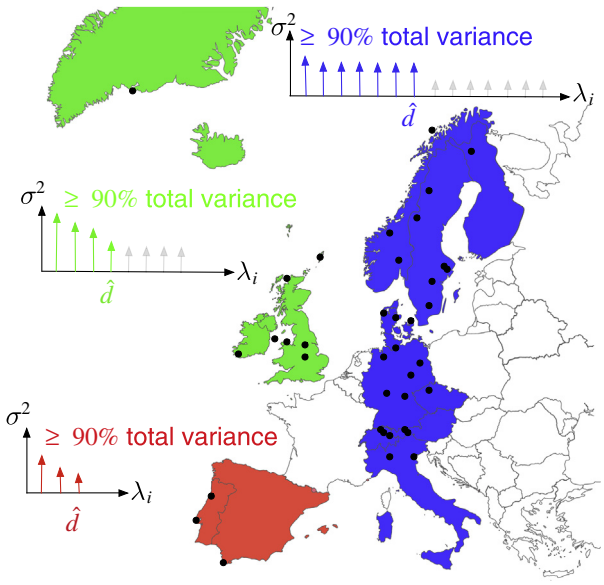

Fig. 1. The SODCC clustering of the measuring stations can be understood in terms of explained variance. The variance of the measurements from measuring stations in small clusters (red) can be explained with almost all of the (few) present principal components. As clusters enlarge, more principal components are needed to explain the variance of the measurements (green and blue).

By using the signal subspace dimension as a clustering criteria for the different measuring stations, we ensure that most of the variance in the data is captured within that cluster (see Fig. 1). It means that, if the cluster is small, the data from a small number of stations are needed to extract the eigenvalues of the largest principal components that explain their variance. Thus, the cross-correlation of the data among those stations is high. However, if the cluster is large, the data from many more stations are needed to explain the same percentage of the variance. Therefore, by using the dimension of the signal subspace we are ensuring that closely correlated temperature series are clustered together.

In the following subsections we describe the system model used, in which each measuring station is one of the nodes of the network. Next, we introduce the Fast Subspace Decomposition (FSD) algorithm (Xu & Kailath, 1994), the one used by SODCC to estimate the dimension of the signal subspace in each cluster. The third issue of this section is to carry out a detailed description of the SODCC algorithm. The main goal of SODCC is to group in the same cluster stations with high spatio-temporal correlation, linking the cluster configuration with the measured field. For this purpose, the output of FSD forms part of the criteria for deciding if a cluster has the minimum size required: when the signal subspace dimension is lower than the cluster size and the covariance matrix of the measured dataset is separable.

2.1. System model

Each measuring station considered is modeled as a node forming part of a network. Therefore, let $G = (V, E)$ be the graph modeling the network, with $|V| = N$ measuring stations and $|E| = C$ links or connections between the stations.

Each of the N stations takes a measurement each T_s time instants starting in T_I, obtaining a total of M measurements. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]$, with $\mathbf{x}_m \in \mathbb{R}^N$, be the dataset measured by the entire network. The covariance matrix of $\mathbf{X}$ can be estimated as $\boldsymbol{\Sigma} = \mathbf{X}\mathbf{X}^\top/M$, where $(\cdot)^\top$ indicates the transpose operation.

2.2. Dimension of the signal subspace of a cluster

In order to compute the dimension of the signal subspace robustly and iteratively as the cluster evolves, we employ the FSD algorithm (Xu & Kailath, 1994). This algorithm uses the estimation of the first $\hat{d}$ Rayleigh–Ritz eigenvalues and eigenvectors (spanning the signal subspace) using at most $\hat{d} = d$ iterations, being d the actual dimension. FSD is based on the Lanczos method and it has $O(N^2d)$ computational complexity, significantly less than the traditional eigendecomposition that has a $O(N^3)$ complexity.

For a $N \times M$ data matrix, the statistic $\varphi_{\hat{d}}$ is defined as (Xu & Kailath, 1994):

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

where $||\cdot||$ is the Fröbenius norm and θ_n is the Rayleigh–Ritz eigenvalues. In each iteration, for $\hat{d} \geq d+1$, the statistic $\varphi_{\hat{d}}$ tends to a χ^2 distribution with $(1/2)\left(N-\hat{d}\right)\left(N-\hat{d}+1\right)-1$ degrees of freedom.

It is shown in Xu and Kailath (1994) that

$$\varphi_{\hat{d}} \leq \gamma_{\hat{d}}\ c(M) \quad (2)$$

holds for M measurements, where $\gamma_{\hat{d}}$ is a threshold for the χ^2 distribution that is computed a priori. Moreover, the function $c(M)$ must fulfill the following conditions:

$$\lim_{M\to\infty}\frac{c(M)}{M} = 0 \qquad \text{and} \qquad \lim_{M\to\infty}\frac{c(M)}{\log\ \log M} = \infty. \quad (3)$$

In practice, the asymptotic behavior of $c(M)$ has to be "slower" than linear and "faster" than log log. Functions such as $c(M) = log(M)$ or $c(M) = \sqrt{log(M)}$ can be used (Xu & Kailath, 1994).

From the matrix perturbation theory (Nadler, 2008), we establish a lower bound for the amount of measurements M needed to recover the principal eigenvalue. Consider σ^2 as the noise power, $||\mathbf{v}||^2$ as the modulus of the first principal component of matrix $\boldsymbol{\Sigma}$ and, $\mathrm{SNR}_v = ||\mathbf{v}||^2/\sigma^2$ as the minimum signal to noise ratio needed to recover the first eigenvalue. The stochastic and self-adjoint matrix $\boldsymbol{\Sigma}$ experiences a phase-transition for M measurements per each of the N stations, such that $M/N \geq \mathrm{SNR}_v^{-2}$ (Eq. (2.19) from Nadler (2008)). The phase transition is characterized by the "collapse" of the first $\hat{d}$ eigenvalues from the noise to the signal subspace. This fact indicates that the minimum amount of measurements M needed to detect the principal eigenvalues with a signal to noise ratio SNR_v is

$$M \geq \frac{N}{\mathrm{SNR}_v^2}. \quad (4)$$

The concept of SNR in the context of the present work merits a more detailed explanation. The threshold $\gamma_{\hat{d}}\ c(M)$ in Eq. (2) is established in terms of the tail of the χ^2 distribution with $(1/2)\times\left(N-\hat{d}\right)\left(N-\hat{d}+1\right)-1$ degrees of freedom. Fixing this threshold to account for a percentage of such distribution (e.g. 90%) means that the eigenvalues in the signal subspace account for 90% of the variance of the data. Thus, the SNR is a ratio of the power of the signal subspace principal components to those of the noise subspace or, in other words, SNR establishes the ratio of explained vs. unexplained variances of the signal in the cluster. In the present work, the minimum SNR is established such that it accounts for 90% of the variance.

Algorithm 1. First stage: cluster initialization

```
1: Random decision to turn into MSH
2: if MSH then
3:   REQUEST to join the cluster to first neighbours (only free stations are able to join)
4:   UNION of selected stations from the first neighbours
5: end if
```

2.3. Second-Order Data-Coupled Clustering algorithm (SODCC)

In this work, we propose a clustering algorithm that has two stages: 1) random initialization of the first Measuring Station Heads (MSHs) (that will act as the first seed of the cluster) and formation of the first clusters and 2) fusion between existing clusters in order to fulfill the convergence criterion in each of the final clusters. The decision criterion in the second stage uses second-order statistics of the measured data, specifically, the dimension of the signal subspace of the dataset measured by the stations in each cluster. In order to ensure a separable correlation matrix for each cluster, the necessary and sufficient condition is that the dimension of the signal subspace has to be lower than the amount of stations that form the cluster (the cluster size).

2.3.1. First stage

The first stage of the SODCC algorithm (see Algorithm 1) starts with all N stations as "role-free measuring station", namely these are not neither MSH nor belong to any cluster. The role-free stations decide to switch their role to MSH, with an a priori probability P, in a completely independent way. The new MSHs request their first neighbors to be part of the newly formed cluster. Only the role-free stations are able to response. Each MSH picks up a subset of the stations and adds them to the cluster and terminates its particular first stage.

The remaining role-free stations repeat Algorithm 1 after a T_1 timer is reached. Moreover, after a $T_2 >> T_1$ timer is reached, all remaining role-free stations automatically switch their role to MSH, finishing the first stage for all N stations.

In this stage, a reasonable value for the maximum size allowed for a cluster (N^{1st}) is needed as in this initialization stage the data and its structure is irrelevant. However, if any MSH is not able to acquire N^{1st}-1 stations, the cluster is initialized with a lower size. Therefore, the resulting clusters of the first stage may have sizes of 1, 2, 3..., N^{1st} stations. As the N^{1st} parameter is data-independent, its value does not interfere in the final cluster configuration.

2.3.2. Second stage

The operation of the second stage of SODCC (see Algorithm 2) is the following: as the stations that form a cluster gather data (at least M_i measurements per station), the dimension of the signal subspace $\hat{d}$ of the dataset is estimated and a decision is taken depending on its value. If FSD is able to compute a dimension lower than the cluster size, that cluster fulfills the convergence criterion. If not, a cluster-fusion process is needed for this particular cluster.

Algorithm 2. Second stage: cluster growing

```
1: M_i = N_i × SNR_v^{-2}
2: if MSH then
3:   WAIT for M_i measurements per station
4:   d̂ ← dimension of the signal subspace, estimated using FSD
5:   if d̂ ≥ N_i then
6:     FUSION with selected cluster
7:     if Newly fused cluster then
8:       Gather all data from the joining cluster; update N_i
9:     end if
10:  end if
11: else
12:   Construct the new correlation matrix
13: end if
```

The clusters that are formed by stations with high spatio-temporal correlation among them estimate a signal subspace dimension lower than the cluster size ($\dot{d}<N_i$), meaning that the noise subspace dimension is at least 1, and that the signal and noise subspaces are separable. This type of clusters fulfills the convergence criterion of SODCC. On the other hand, clusters formed by stations with low spatio-temporal correlation between them do not satisfy Eq. (2), i.e. the signal and noise subspaces are not separable. In this second case, the cluster must grow (through fusion with another cluster) in order to fulfill the convergence criterion of SODCC in the future.

The criterion used to select a cluster to merge is crucial, as it directly impacts in the final cluster configuration. The SODCC algorithm aims to maximize the spatio-temporal correlations between the stations that form each cluster. This goal is obtained if the fusion criterion: 1) minimizes the distance between cluster centers-of-mass or 2) minimizes the cluster physical area. The cluster-fusion process is mandatory for the selected cluster, even if it has already fulfilled the convergence criterion.

3. Experiments and results

3.1. Datasets and data interpolation

In our experiments we use data gathered by measuring stations from the European Climate Assessment & Dataset (Klein Tank et al., 2002) and from several measuring stations from the European HISTALP project (Chimani et al., 2013), from Central Europe and Alps, measured between January 1940 and December 2010. The climatological variable used is the average temperature, calculated over monthly periods, i.e. $T_s = 1$ month. The time series used in the present work are from measured data and exhibit gaps. Specifically, we use 123 measuring stations (see Fig. 2), free from error bursts larger than 24 consecutive values (2 yrs).

Note that, to fill in data gaps, typical interpolation methods in climatology research involve spatial interpolation techniques (among measuring stations) such as global, local, and geostatistical methods (Vicente Serrano et al., 2003). These can be used to generate very high resolution interpolated climate surfaces (Hijmans et al., 2005). However, these methods may introduce artificial correlations among measuring stations which can mask actual spatial interactions detected by the proposed SODCC algorithm.

In order to avoid the introduction of spurious spatial correlations, we propose a simple, low-complexity interpolation method that deals with independent time-series from the measuring stations, i.e., data from each station are interpolated individually. The introduction of artificial time-correlations is a possibility. However the percentage of interpolated data is marginal compared to the total available data, as can be seen in Fig. 3(a). In this figure, we represent the percentage of missing values in each decade for the 123 stations, using the histogram method. For the worst case (1990s), the proportion of missing values is lower than 6%, which ensures an acceptable data quality.

For each data series we interpolated the remaining missing values, in an independent way, using the following model:

$$y = A\sin(2\pi f_0 x) + B\cos(2\pi f_0 x) \tag{5}$$

where A and B are weighting constants, f_0 is the fundamental frequency, and x is the vector of timestamps. In this paper we used a fundamental period of 1 year for the interpolation. The fitting between the model and the data minimizes the mean square error. In Fig. 3(b) we show, as an example, a segment of one of the data series where 11 values were interpolated using the aforementioned model. In this figure, the black circles represent the temperature data that is correctly measured, the red squares represent the values that were interpolated and the dotted gray line is used as a guide for the figure.

3.2. Experimental methodology

We apply the SODCC algorithm to the dataset of the 123 measuring stations for time intervals beginning in, respectively, $T_I \in \{1940, 1950, 1960, 1970, 1980, 1990\}$. As stated in the previous section, the output of each independent SODCC realization is a set of clusters of neighboring stations and thus, is spatially proximal. Furthermore, those clusters are ensured to encompass the measuring stations that allow us to obtain the minimum number of principal components that explain 90% of the variance of their data. The objective is then to look for changes in clustering distribution probabilities through the different considered time intervals. Thus, the methodology applied is the following: for each decade, we perform 10,000 independent random cluster seeds, i.e., the initial measuring stations in stage 1 of SODCC are randomly selected. This is done in order to obtain statistical significant results and thus, independent of the initialization of the clustering algorithm.

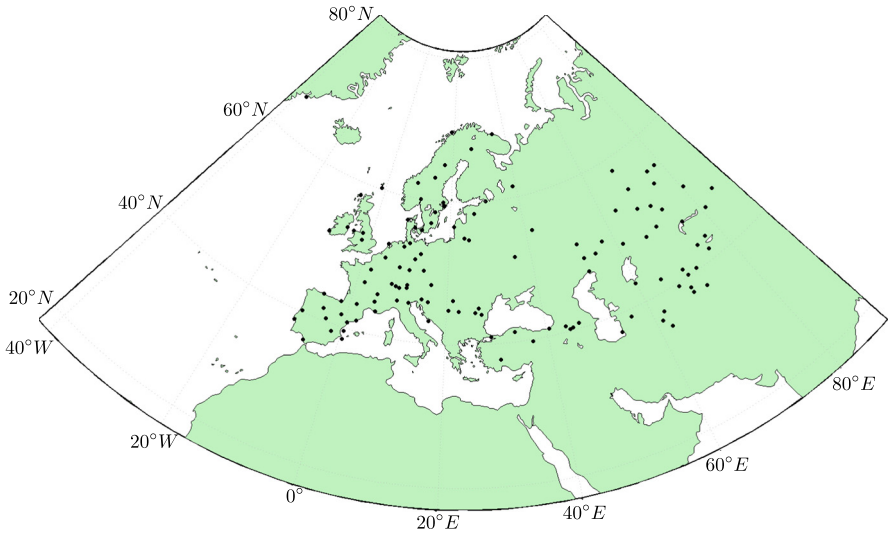

Fig. 2. Location of the 123 measuring stations (black dots) used in the present work.

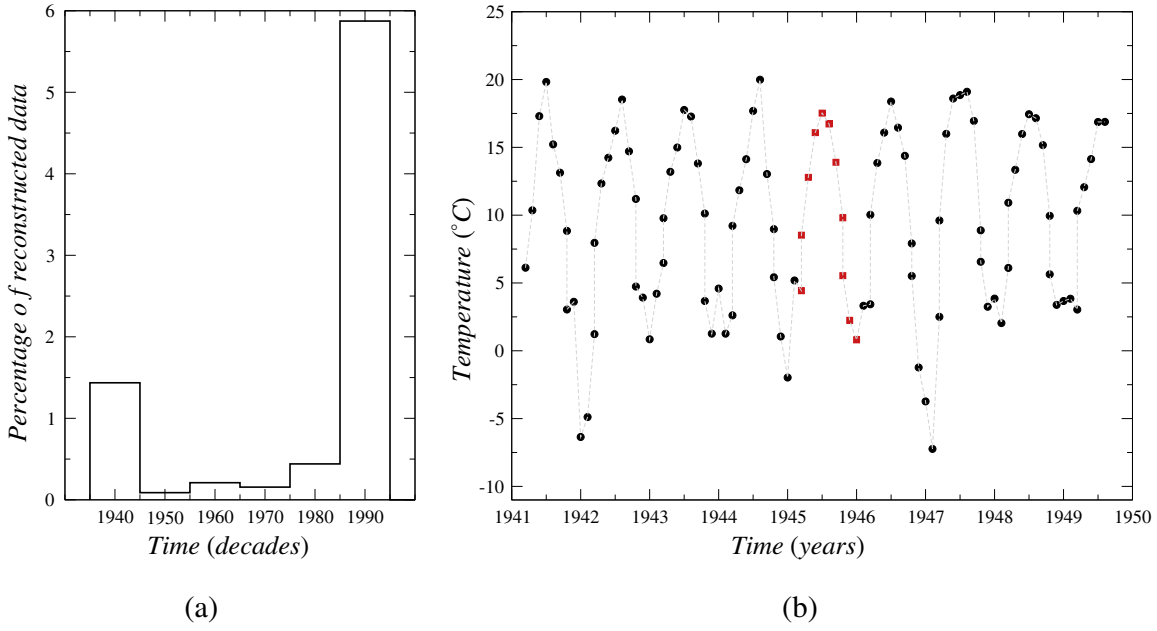

Fig. 3. (a) Percentage of missing data per decade. All this missing data was interpolated using the model in Eq. (5). (b) Example of reconstructed data (red squares) using original data (black circles) and the model in Eq. (5). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

As results, we obtain information about the geographical location of the most probable cluster borders and their time evolution. The appearance of highly probable domain frontiers (i.e. cluster border) and its change through time, will also reflect the spatial extent of the correlation and its change through the different time series. Moreover, we also obtain data that allows us to relate the stations pertaining both to the same cluster and to different clusters. We focus on the probability of a given station to be in a cluster of a given size. Any change through time of this statistical distribution will reflect a change in the spatial domain of the time correlations among neighboring stations.

3.3. Results

In the following subsections we define three metrics that allow us statistically evaluating the spatio-temporal correlations. The first one is a global measure of the spatial-extent statistical distribution of the correlations among measuring stations present in the decade, the so called cluster size probability. The second and third metrics are related to the individual affinity of the stations to be associated with their neighbors. The Pairwise Frontier Probability measures the affinity (or lack thereof) of two neighboring stations based on their cross-

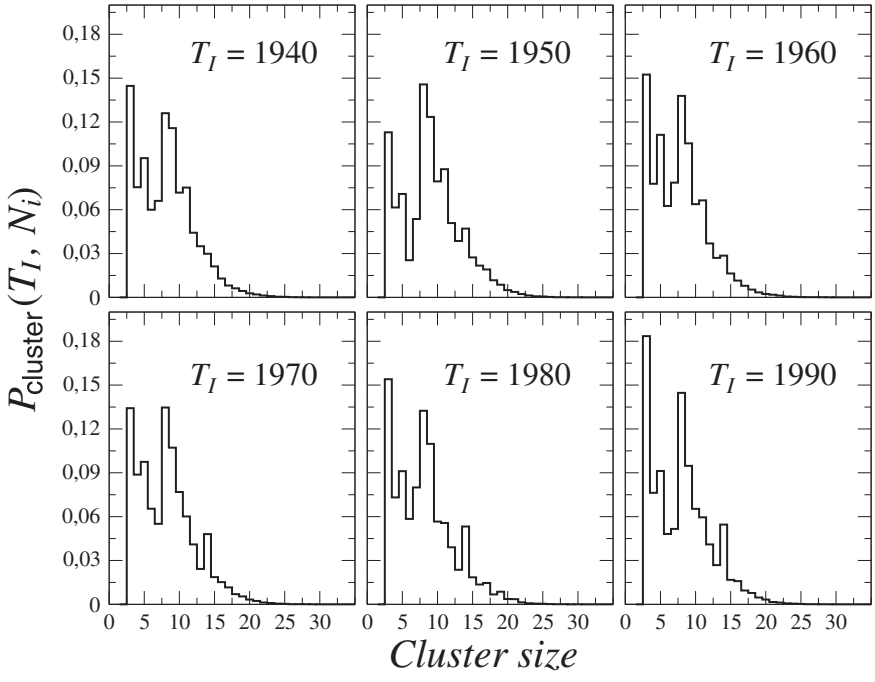

Fig. 4. Cluster size distribution resulting from the SODCC clustering of the data from the 123 measuring stations in the six considered decades.

correlations. The Station-to-Cluster-Size Probability measures the probability of a given station to be associated with a cluster of a certain size. With these three measures, we analyze the statistical distributions and time evolution of the clustering that SODCC algorithm achieves on air temperature datasets of stations throughout Europe and Western Asia considered.

3.3.1. Cluster size probability

In order to follow the time evolution of the spatio-temporal correlations, we use the cluster size probability distribution for the decade beginning in T_I and different cluster sizes, namely $P_{\text{cluster}}(T_I, N_i)$. In Fig. 4 we show the normalized histogram of $P_{\text{cluster}}(T_I, N_i)$ obtained. In the first three decades (1940, 1950, 1960), we can clearly see a bimodal asymmetric distribution with maxima in cluster sizes of $N_i = 3$ and $N_i = 8$ stations. The cluster sizes corresponding to these peaks are related to the number of principal components that explain 90% of the variance of the data in the cluster. It has been shown elsewhere (Chidean et al., 2014) that the first peak is related to the minimum signal subspace dimension present throughout the stations, while the second peak is accordingly related to the maximum signal subspace dimension. The exponential decay after the two peaks can be directly related with the criteria for the cluster fusion process. Thus, the cluster sizes in those peaks can be directly related to the extent of the spatio-temporal correlations of the air temperature measurements.

The most striking feature appears in the following three decades (1970, 1980, 1990) where a third peak appears in the cluster size value of $N_i = 14$ stations. Thus, we can see that larger spatio-temporal correlations appear (affecting to a larger number of stations) from 1970 onwards.

From the previous results we identify three clearly differentiated cluster size regimes which can be associated with the according extent of spatio-temporal correlations among measuring stations. We define

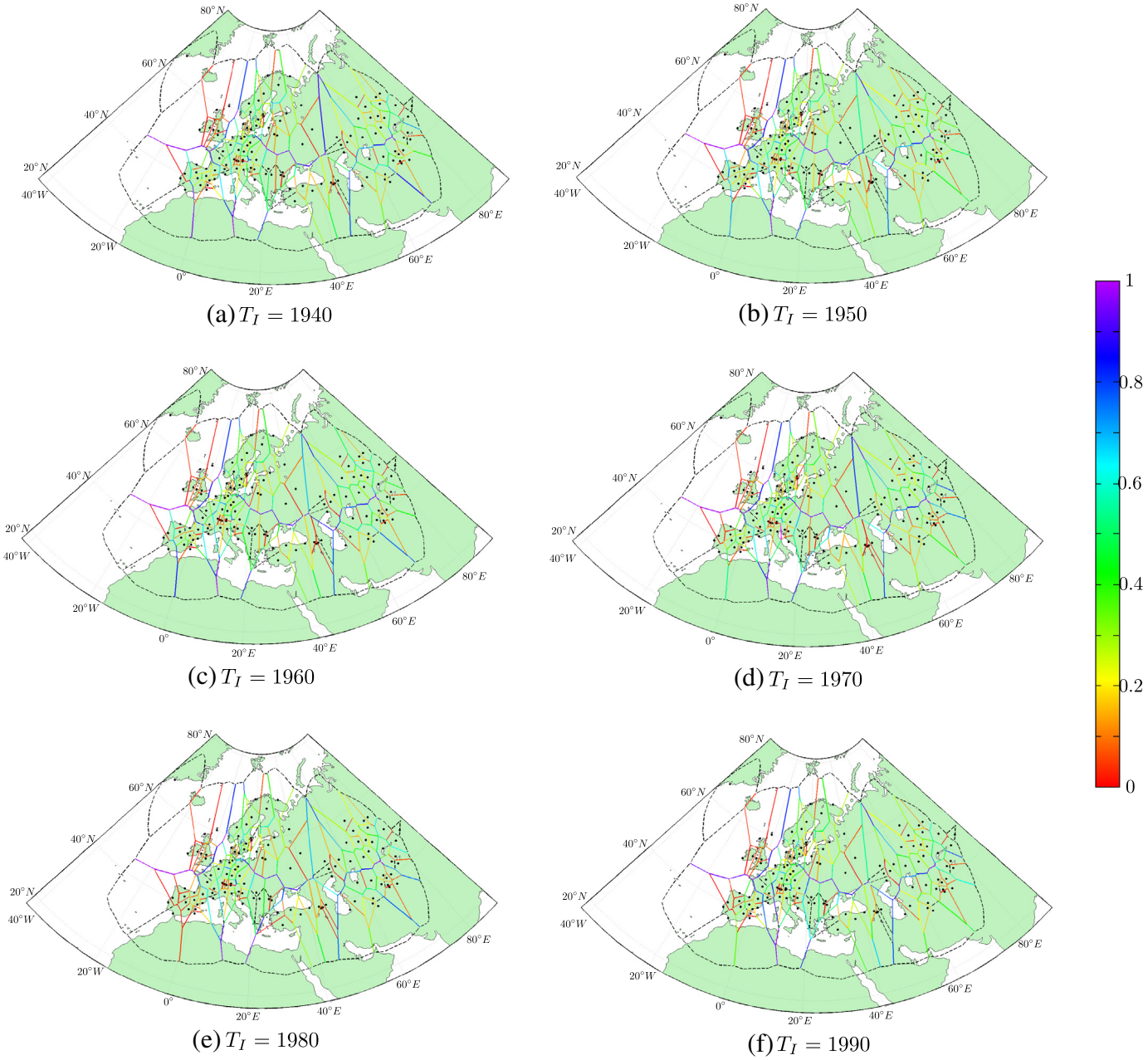

Fig. 5. Pairwise Frontier Probability (with reference to the Voronoi cell borders) in the decades $T_I \in \{1940, 1950, 1960, 1970, 1980, 1990\}$. 10^4 independent realizations have been performed for each decade. Probabilities lower than $2 \cdot 10^{-4}$ are indicated with discontinuous lines.

those regimes as 1) small-scale correlations that we identify with clusters of sizes 3, 4, and 5; 2) medium-scale correlations that we associate with clusters of sizes 7, 8, and 9; and 3) large-scale correlations that are bound to appear in clusters of sizes 14, 15, and 16.

To further facilitate the spatial interpretation of the results that will be shown next (location and extent of clusters), we depict the measuring stations as centroids of a Voronoi region. This type of representation not only helps visualize the relation among each station and its neighbors, but it also offers information about the spatial density distributions of the stations: small Voronoi regions appear in areas with high station density. In order to limit the perimeter of the Voronoi regions located on the periphery, we added auxiliary centroids (not shown) at the 17°*N* and 81°*N* parallels and 42°*W* and 89°*E* meridians.

3.3.2. Pairwise frontier probability

To study the behavior of the cluster borders, we compute the Pairwise Frontier Probability (PFP), namely the probability that a cluster frontier occurs between two neighboring stations. The PFP represents the likelihood of a cluster frontier to appear with the SODCC algorithm and we depict the results obtained in Fig. 5. We see that most frontiers are stable in their PFP values. We also observe that composite frontiers from two or more borders are highly stable through time. The variation in PFP for the unstable (with respect to time) frontiers is, in general, small and without an evident temporal structure. However, a composite frontier that crosses diagonally the Iberian Peninsula purports the largest observed instability of the whole dataset. Its individual frontiers exhibit $0.6 \leq \text{PFP} \leq 0.8$ in the decades previous to 1970 but, from 1970 onwards, their probabilities lower to $\text{PFP} \leq 0.4$.

3.3.3. Station-to-Cluster-Size Probability

To further clarify the location, extent, and evolution of the SODCC clusters, in the following viewgraphs we represent the Station-to-Cluster-Size Probability (SCSP) $P_{\text{station}}(T_I, N_i)$ where we compute the

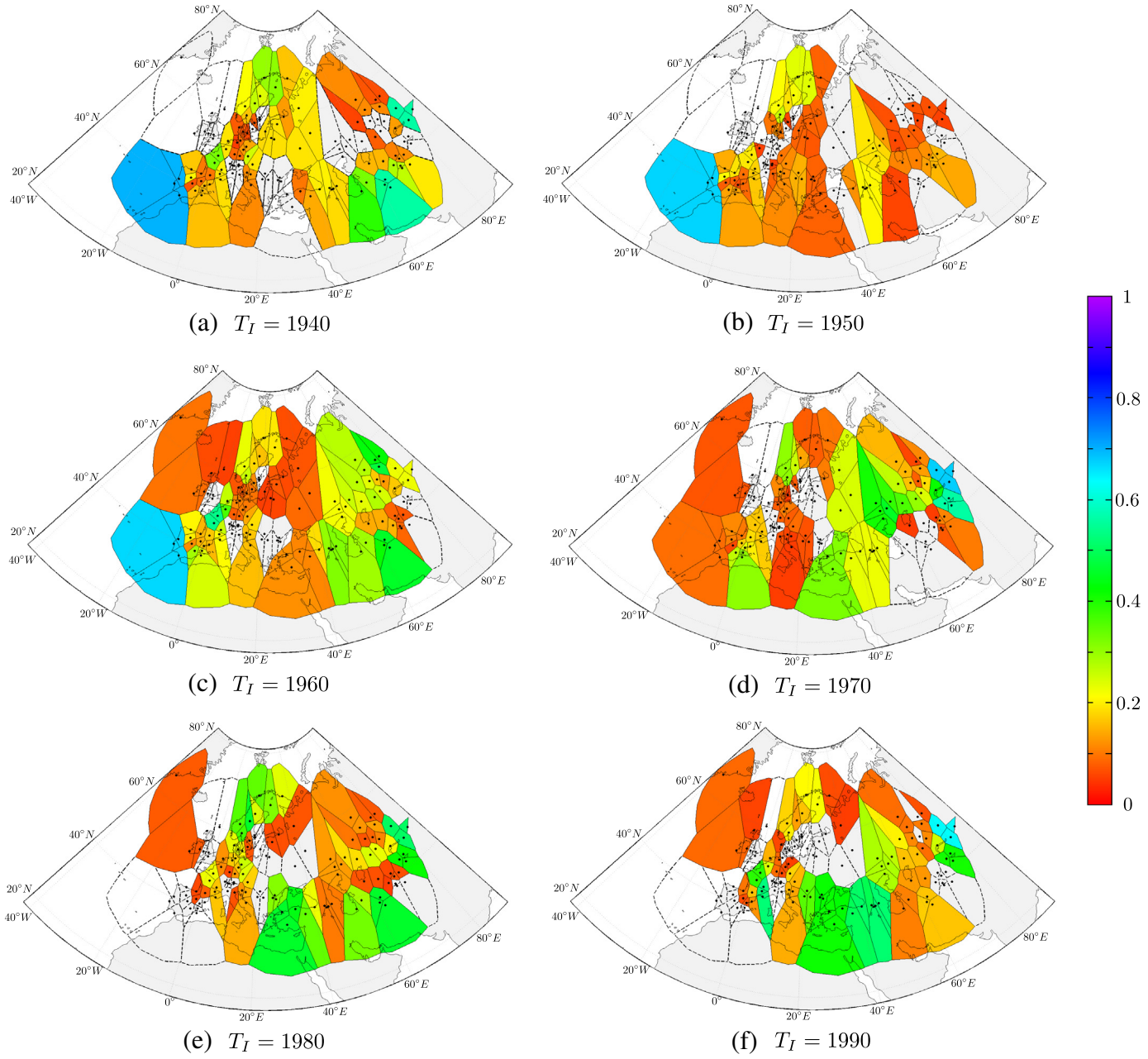

Fig. 6. Station-to-Cluster-Size Probability for clusters with size $N_i = 3$, 4, and 5, $\sum_{N_i=3}^{5} P_{\text{station}}(T_I, N_i)$ in decades $T_I \in \{1940, 1950, 1960, 1970, 1980, 1990\}$.

probability that any given node is part of a cluster of size N_i in the T_I decade.

In Fig. 6 we show $\sum_{N_i=3}^{5} P_{\text{station}}(T_I, N_i)$ to represent the location, spatial distribution, and time evolution of clusters associated with small-scale correlations and with the first peak of $P_{\text{cluster}}(T_I, N_i)$ in Fig. 4. We observe high values of $P_{\text{station}}(T_I, N_i)$ in three stations of the southwest Iberian Peninsula in the first three decades (1940, 1950, 1960), namely Coimbra, Lisbon, and Cadiz. However, the SCSP for these stations abruptly drops to negligible values from 1970 onwards. This fact is consistent with the disappearance of the composite frontier mentioned earlier.

In Fig. 7 we plot the SCSPs associated with the medium-scale correlations, namely $\sum_{N_i=7}^{9} P_{\text{station}}(T_I, N_i)$, and with the second peak of the PFP in Fig. 4. We see that stations in the Iberian Peninsula show low values of SCSP for the medium-scale correlations. However, in the 1970 decade there is a significant increase, consistent with a small-to-large cluster transition through of a medium-size cluster. In this decade, the spatial correlations among the stations in the peninsula would increase to reach southern France.

And in Fig. 8 we represent $\sum_{N_i=14}^{16} P_{\text{station}}(T_I, N_i)$, namely the SCSPs associated with the large-scale correlations and with the third peak of the PFPs that only appears in the (1970, 1980, 1990) decades in Fig. 4. We clearly see an increase in the probability of being associated with large clusters for all the measuring stations in the Iberian Peninsula and Southern France in the 1970s. We also observe that $\sum_{N_i=14}^{16} P_{\text{station}}(T_I, N_i)$ arises from baseline values ($\approx 10^{-4}$) in the 1940–1969 period to significant values in the 1970 decade and stabilizes in values ≈ 0.45 in the 1980–1999 period.

With these results, we are able to detect a stable trend which shows an increase in the scale of the spatial correlations among stations in

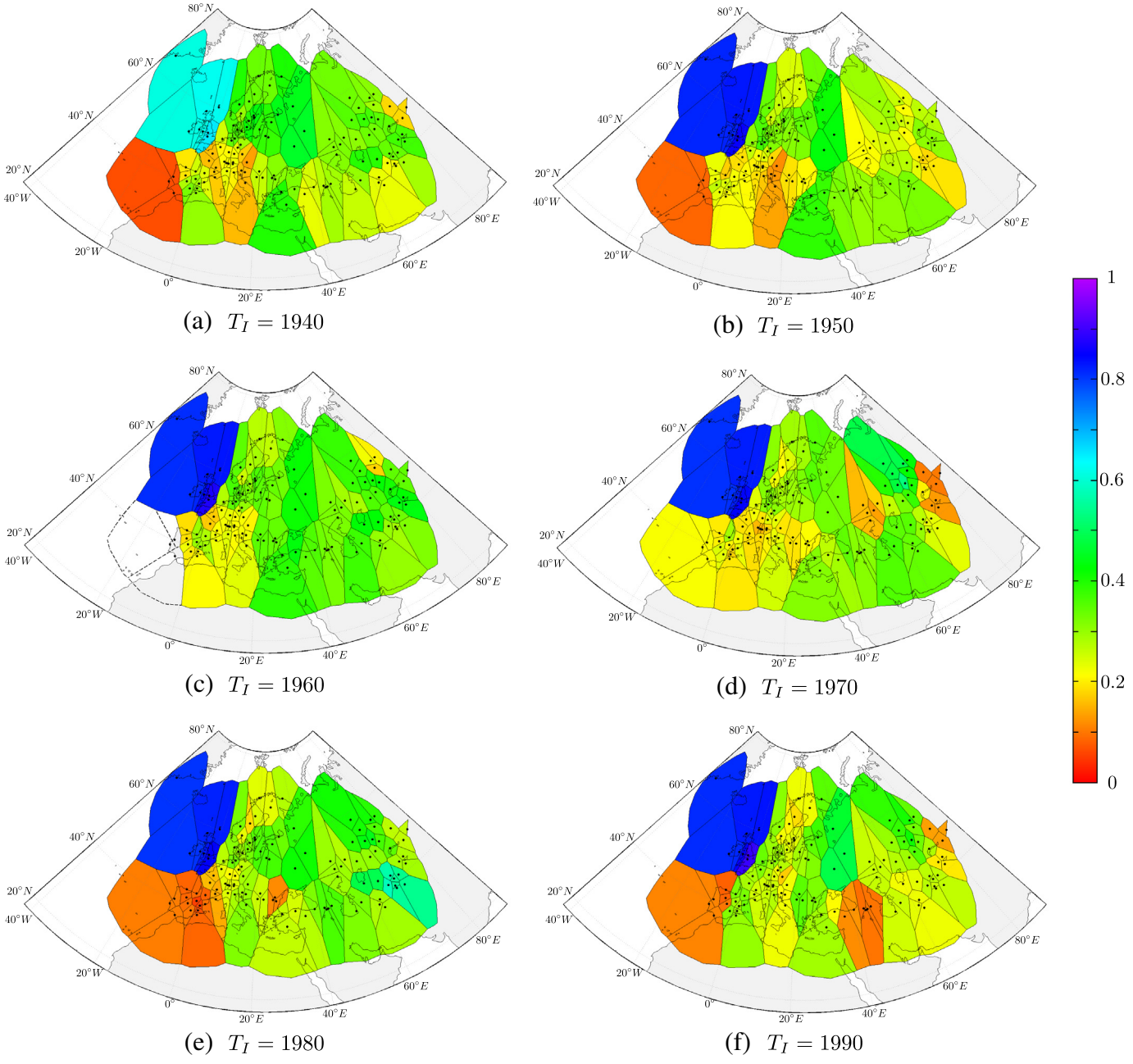

Fig. 7. Station-to-Cluster-Size Probability for clusters with size $N_i = 7$, 8, and 9, $\sum_{N_i=7}^{9} P_{\text{station}}(T_I, N_i)$ in decades $T_I \in \{1940, 1950, 1960, 1970, 1980, 1990\}$.

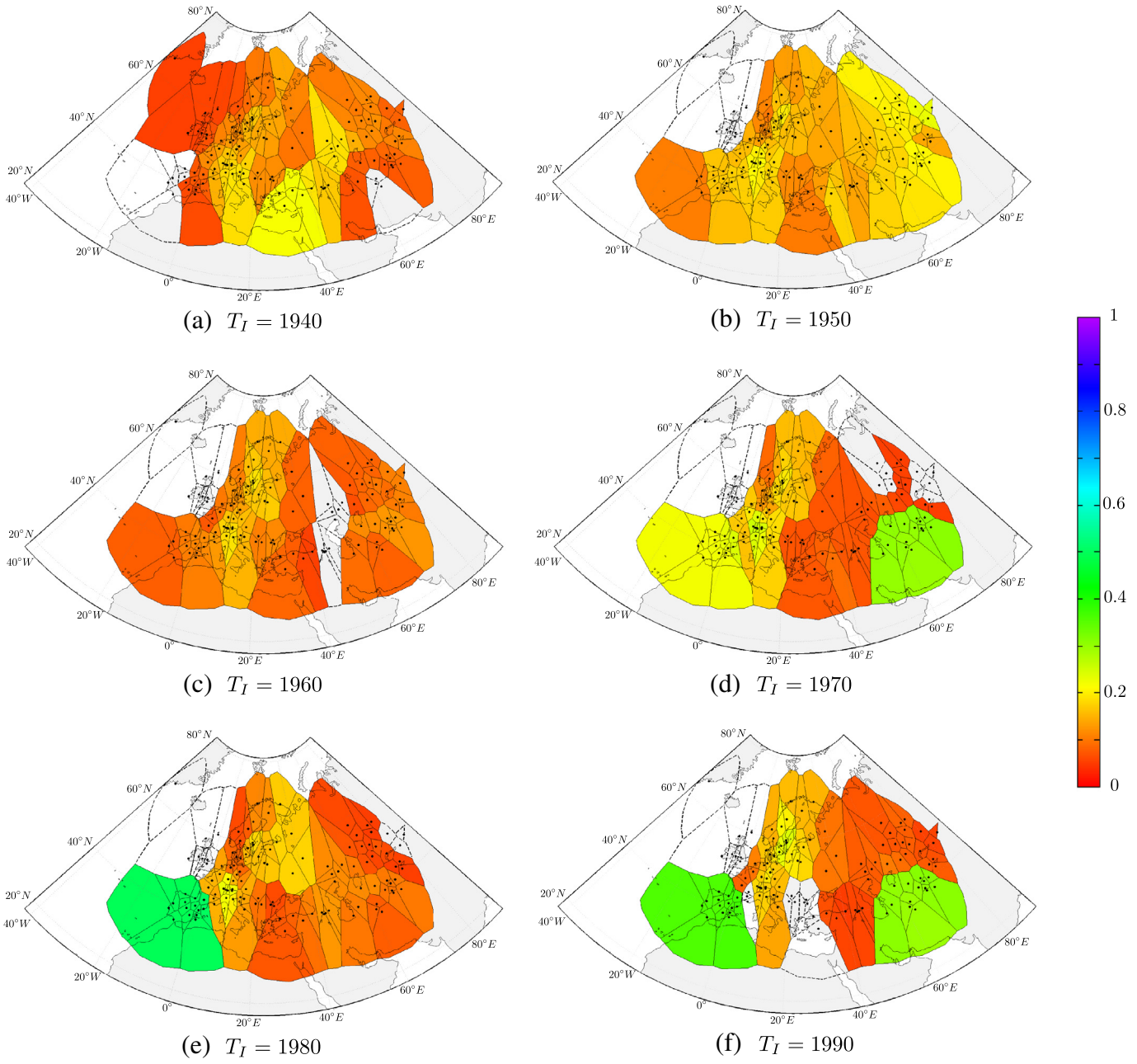

Fig. 8. Station-to-Cluster-Size Probability for clusters with size $N_i = 14$, 15, and 16, $\sum_{N_i=14}^{16} P_{\text{station}}(T_I, N_i)$ in decades $T_I \in \{1940, 1950, 1960, 1970, 1980, 1990\}$.

Southern Europe by means of the spatio-temporal clustering of air-temperature data. In the 1940 decade the 90% of the variance of the data from three measuring stations can be explained within the subspace of principal components spanned by these datasets. In the 1950–1969 period, we observe an increasing probability of these measuring stations to be associated with neighboring measuring station datasets. The spatio-temporal correlations in these first three decades are in the small-scale regime. However, the growth in the average cluster size associated with the Iberian Peninsula region in the 1970 decade corroborates a shift of the spatio-temporal correlations to the medium-scale regime. The further increase in the average cluster size in that region in the 1980–1999 period evidences a shift of the spatio-temporal correlations regime, extending the region affected to stations in Southern France. This fact points to a possible evidence of a climate change (or climate variability) pattern affecting Southern Europe.

Although the previous trend can be identified with low effort, other trends can be seen to be present. These trends show, for example, extremely stable medium-scale regimes associated with the British Isles (see Fig. 7) or seemingly oscillating (with a two-decades period) large-scale regimes in the Turkmenistan, Tajikistan, and Uzbekistan stations (see Fig. 8) which is a strong indicator of a climate variability pattern.

3.4. Analysis of the results

In the present section we analyze the significance in the change of probability distributions for individual "tell-tale" (representative) measuring stations. These are stations that can be identified as having the largest change in their SCSP $P_{\text{station}}(T_I, N_i)$ with respect to the reference SCSP which we fixed to be their respective initial $P_{\text{station}}(1940, N_i)$.

In order to quantify the change in the SCSP for the 123 stations we use the Kullback–Leibler distance (or divergence) $D_{\text{KL}}(P\|Q)$. This metric is commonly used in information theory to quantify the differences between two probability distributions (either discrete or continuous).

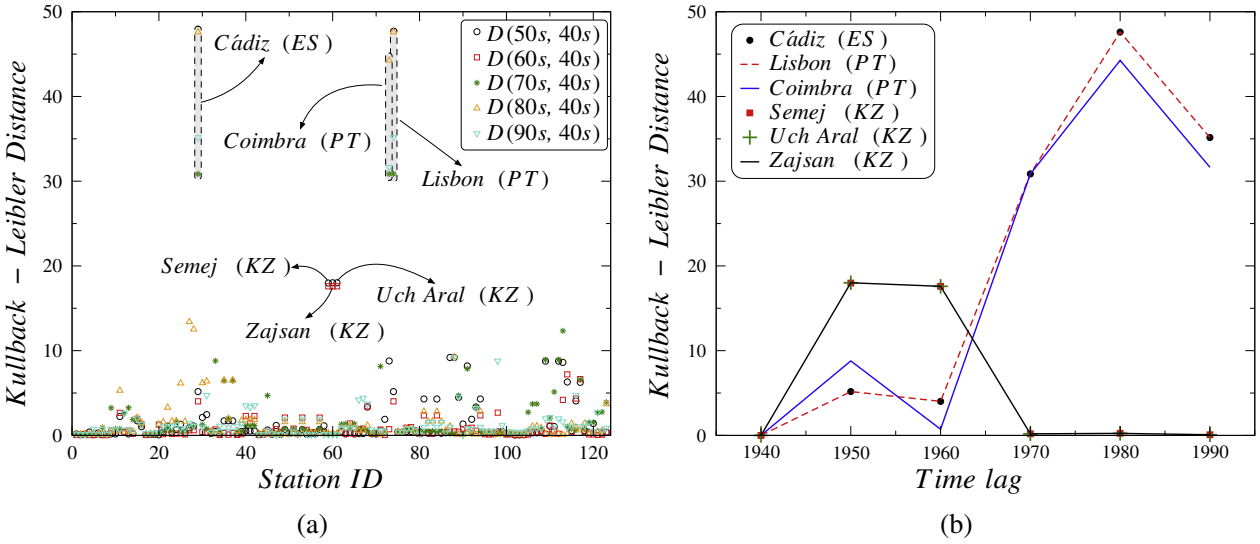

Fig. 9. Kullback–Leibler distance $D_{\text{station}}(T_I, 1940)$ of the Station-to-Cluster-Size Probability Mass Function $P_{\text{station}}(T_I, N_i)$ to $P_{\text{station}}(1940, N_i)$ for (a) the 123 stations and (b) for the 6 stations with the largest $D_{\text{station}}(T_I, 1940)$ vs. the time lag T_I.

Specifically, it accounts for the information lost if a given probability distribution P is used to approximate a different probability distribution Q. Although termed a "distance", it is not a true distance metric as it is not symmetric with the interchange of distributions, i.e. $D_{\text{KL}}(P||Q) \neq D_{\text{KL}}(Q||P)$. However, this last property has no consequences in the present work.

We define the Kullback–Leibler distance between two SCSP belonging to the same stations, $P_{\text{station}}(T_I, N_i)$ and $P_{\text{station}}(1940, N_i)$ as

$$D_{\text{station}}(T_I, 1940) = \sum_{N_i=1}^{N_{max}} P_{\text{station}}(T_I, N_i) log \frac{P_{\text{station}}(T_I, N_i)}{P_{\text{station}}(1940, N_i)}. \quad (6)$$

We then compute $D_{\text{station}}(T_I, 1940)$ for all the decades of interest and for all 123 stations. In Fig. 9(a) we represent the results of such computation. We can then identify 6 stations that rise above the fray of values $D_{\text{station}}(T_I, 1940) \leq 15$. Within those 6, we can identify two classes: 1) Cadiz, Lisbon, and Coimbra in the Iberian Peninsula with $D_{\text{station}}(T_I, 1940) \geq 30$ in the 1970–1990 period and 2) Semej, Uch Aral, and Zajsan in the Kazakhstan region with $17 \leq D_{\text{station}}(T_I, 1940) \leq 18$ in the 1950–1969 period.

We show the time evolution of $D_{\text{station}}(T_I, 1940)$ for the six selected stations in Fig. 9(b). We see that, for the Iberian stations, an increasing trend is clearly present from 1970 onwards that drastically differentiates their initial SCSPs in the 1940 decade from those in the three decades from 1970 to 1999. The change in the three Kazakhstan stations is not as abrupt as the Iberians and it is not stable in time.

In Fig. 10 we plot the SCSP for the Cadiz measuring station (a) and for the Semej station (b). We observe the change in probability distribution of the first is clearly more drastic than in the second case.

Therefore, SCSP from individual stations can help us identify "tell-tale" or representative stations within a spatio-temporal trend. However, despite this fact, these stations cannot help us envision the extent of the spatio-temporal trend as data variance from the different stations is not homogeneously distributed. For example, stations that are bound to the large-scale regime in the Iberian Peninsula but do not exhibit a large $D_{\text{station}}(T_I, 1940)$ are difficult to be identified as part large spatio-temporal correlations. Another example of such apparent mismatch

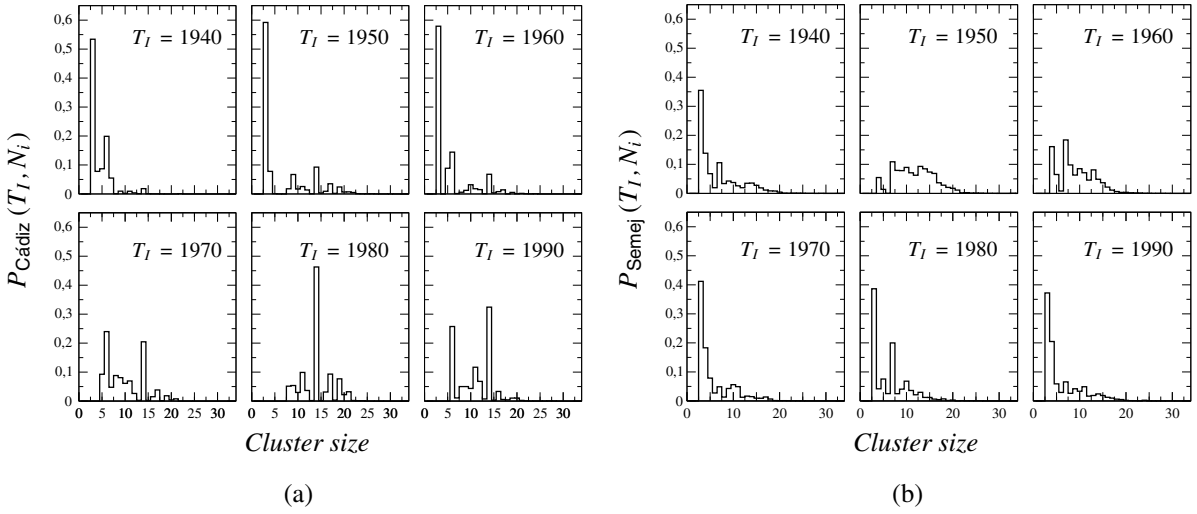

Fig. 10. Station-to-Cluster-Size Probability Mass Function $P_{\text{station}}(T_I, N_i)$ for two "tell-tale" measuring stations: (a) Cadiz (ES) and (b) Semej (KZ).

between individual and collective representations occurs in the case of the Kazakhstan stations. Despite we are able to identify a trend change in this region in the 1950 and 1960 decades, we can only identify a small-scale regime (see Fig. 6(a–c)) in the region, in contrast to the previous large-scale regime in the Iberian region. That being the case, collective measurements are to be preferred to individual measurements in probability of cluster size formation if spatio-temporal trends are to be identified.

The results obtained in this work, pointing out possible climate change related to air temperature in Europe and Western Asia in the last decade of 20th Century, are in accord with previous studies of climate change impact such as Thuiller et al. (2005). In that paper, a study of climate change threats to plant diversity was presented for Europe, showing zones of special sensitivity. In fact, an important result obtained in Thuiller et al. (2005) points out that "the greatest changes are expected in the transition between the Mediterranean and Euro-Siberian regions", and also the Iberian Peninsula is pointed out to be a zone of important species loss. The results obtained in this paper support those previous findings. In conclusion, SODCC can help identify, not only trend changes in the climate in time, but also their geographical extent and intensity.

4. Conclusions

In this work we used a self-organized clustering algorithm, SODCC, that extracts the spatio-temporal correlations of air temperature measurements between different measuring stations. This algorithm works by clustering stations, in a decadal time basis, in which 90% of the data variance can be explained with a minimum number of principal components. We apply this algorithm to the data of 123 stations throughout Europe and Western Asia.

We were able to distinctly identify three different regimes of spatio-temporal correlations based on their geographical extent which are, respectively, small, medium, and large-scale regimes. Based on these regimes, we were able to identify a change in the spatio-temporal trend of air temperature, that reflects in a transition from the small-scale regime to the large-scale regime (passing through the medium-scale regime) of the stations in the Iberian Peninsula and Southern France. We were also able to identify an oscillating spatio-temporal trend in the Turkmenistan, Tajikistan, and Uzbekistan regions (with a two-decade period) that manifest in a seesaw transition between the small and large-scale regimes and a stable medium-scale regime affecting the British Isles. These results are in accord with previous findings in the scientific literature.

We conclude that SODCC has helped us detect an air temperature spatio-temporal trend change in Southern Europe that can be associated with climate change threats to plant diversity and points to an evidence of a climate-change (or climate variability) pattern.

Acknowledgments

We acknowledge the data providers in the ECA & D project. Data and metadata are available at http://www.ecad.eu. We also acknowledge the data providers in the HISTALP project. Data and metadata are available at http://www.zamg.ac.at/histalp.

This work has been partially supported by the Research Projects S2013/MAE-2835 and S2013/ICE-2933 from the Autonomous Community of Madrid.

Mihaela I. Chidean is supported by the FPU Research Grant AP2012-2981 from the Spanish Ministry of Education, Culture and Sports.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gloplacha.2015.03.006.

References

Alva-Basurto, J.C., Arias-González, J.E., 2014. Modelling the effects of climate change on a Caribbean coral reef food web. Ecol. Model. 289, 1–14.
Carrera-Hernández, J., Gaskin, S., 2007. Spatio temporal analysis of daily precipitation and temperature in the Basin of Mexico. J. Hydrol. 336 (3), 231–249.
Chidean, M.I., Morgado, E., Ramiro-Bargueno, J., Caamano, A.J., 2013. Self-organized distributed compressive projection in large scale wireless sensor networks. Proc. IEEE 24th Annual Int. Symp. Personal, Indoor, and Radio Commun. (PIMRC'13).
Chidean, M., Morgado, E., del Arco, E., Ramiro-Bargueno, J., Caamano, A.J., 2014. Scalable Data-Coupled Clustering for Large Scale WSN (submitted).
Chimani, B., Matulla, C., Bohm, R., Hofstatter, M., 2013. A new high resolution absolute temperature grid for the Greater Alpine Region back to 1780. Int. J. Climatol. 33 (9), 2129–2141.
Cobaner, M., Citakoglu, H., Kisi, O., Haktanir, T., 2014. Estimation of mean monthly air temperatures in Turkey. Comput. Electron. Agric. 109, 71–79.
Cramer, W., Yohe, G., et al., 2013. Detection and attribution of observed impacts. IPCC5 Work Group 2, 5th Assessment Report Chapter 18, pp. 1–94.
Douglass, D.H., Blackman, E.G., Knox, R.S., 2004. Temperature response of Earth to the annual solar irradiance cycle. Phys. Lett. A 323 (3), 315–322.
Garske, T., Ferguson, N.M., Ghani, A.C., 2013. Estimating air temperature and its influence on malaria transmission across Africa. PLoS One 8 (2), e56487.
Gomiero, A., Viarengo, A., 2014. Effects of elevated temperature on the toxicity of copper and oxytetracycline in the marine model, *Euplotes crassus*: a climate change perspective. Environ. Pollut. 194, 262–271.
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25 (15), 1965–1978.
Horenko, I., 2010. On clustering of non-stationary meteorological time series. Dyn. Atmos. Oceans 49 (2), 164–187.
Jaglom, W.S., McFarland, J.R., Colley, M.F., Mack, C.B., Venkatesh, B., Miller, R.L., Haydel, J., Schultz, P.A., Perkins, B., Casola, J.H., Martinich, J.A., Cross, P., Kolian, M.J., Kayin, S., 2014. Assessment of projected temperature impacts from climate change on the U.S. electric power sector using the integrated planning model. Energy Policy 73, 524–539.
Kaufmann, R.K., Stern, D.I., 1997. Evidence for human influence on climate from hemispheric temperature relations. Nature 388 (6637), 39–44.
Kaufmann, R.K., Stern, D.I., 2002. Cointegration analysis of hemispheric temperature relations. J. Geophys. Res.-Atmos. (1984–2012) 107 (D2) (ACL–8).
Kaufmann, R.K., Kauppi, H., Mann, M.L., Stock, J.H., 2011. Reconciling anthropogenic climate change with observed temperature 1998–2008. Proc. Natl. Acad. Sci. 108 (29), 11790–11793.
Klein Tank, A.M.G., et al., 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European climate assessment. Int. J. Climatol. 22 (12), 1441–1453.
Kloog, I., Chudnovsky, A., Koutrakis, P., Schwartz, J., 2012. Temporal and spatial assessments of minimum air temperature using satellite surface temperature measurements in Massachusetts, USA. Sci. Total Environ. 432, 85–92.
Kousari, M.R., Ahani, H., Hendi-zadeh, R., 2013. Temporal and spatial trend detection of maximum air temperature in Iran during 1960–2005. Glob. Planet. Chang. 111, 97–110.
Nadler, B., 2008. Finite sample approximation results for principal component analysis: a matrix perturbation approach. Ann. Stat. 36 (6), 2791–2817.
Paniagua-Tineo, A., Salcedo-Sanz, S., Casanova-Mateo, C., Ortiz-Garca, E., Cony, M., Hernández-Martn, E., 2011. Prediction of daily maximum temperature using a support vector regression algorithm. Renew. Energy 36 (11), 3054–3060.
Smith, B.A., Hoogenboom, G., McClendon, R.W., 2009. Artificial neural networks for automated year-round temperature prediction. Comput. Electron. Agric. 68 (1), 52–61.
Solomon, S., Plattner, G.-K., Knutti, R., Friedlingstein, P., 2009. Irreversible climate change due to carbon dioxide emissions. Proc. Natl. Acad. Sci. 106 (6), 1704–1709.
Stone, D.A., Allen, M., 2005. Attribution of global surface warming without dynamical models. Geophys. Res. Lett. 32 (18).
Thuiller, W., Lavorel, S., Araújo, M.B., Sykes, M.T., Prentice, I.C., 2005. Climate change threats to plant diversity in Europe. Proc. Natl. Acad. Sci. U. S. A. 102 (23), 8245–8250.
Van De Kerchove, R., Lhermitte, S., Veraverbeke, S., Goossens, R., 2013. Spatio-temporal variability in remotely sensed land surface temperature, and its relationship with physiographic variables in the Russian Altay Mountains. Int. J. Appl. Earth Obs. Geoinf. 20, 4–19.
Vicente Serrano, S.M., Sánchez, S., Cuadrat, J.M., et al., 2003. Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): application to annual precipitation and temperature. Clim. Res. 24 (2), 161–180.
WGCM, 2014. Coupled Model Inter-comparison Project Phase 5. http://cmip-pcmdi.llnl.gov/cmip5/ (online; accessed 27-April-2014).
Xu, G., Kailath, T., 1994. Fast subspace decomposition. IEEE Trans. Signal Process. 42 (3), 539–551.
Xu, Z., Liu, Y., Ma, Z., Li, S., Hu, W., Tong, S., 2014. Impact of temperature on childhood pneumonia estimated from satellite remote sensing. Environ. Res. 132, 334–341.